Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Οι οριακές επιδράσεις των επεξηγηματικών μεταβλητών ενός γενικευμένου γραμμικού μοντέλου: Ερμηνεία και εφαρμογές

Κωνσταντίνος Γούτος

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς

Ιούλιος 2021

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Καθηγητής Ηλιόπουλος Γεώργιος (Επιβλέπων)
- Αναπλ. Καθηγητής Πολίτης Κωνσταντίνος
- Αναπλ. Καθηγητής Τζαβέλας Γεώργιος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN

APPLIED STATISTICS

Marginal effects of explanatory variables in a generalized linear model: Interpretation and applications.

By

Konstantinos Goutos

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece

July 2021

Στην οικογένεια μου

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κ. Ηλιόπουλο Γεώργιο για την αμέριστη βοήθειά του, την άψογη καθοδήγησή του και τη στήριξή του κατά τη διάρκεια συγγραφής της παρούσας διπλωματικής εργασίας. Επιπλέον θα ήθελα να ευχαριστήσω τα μέλη της Τριμελούς Συμβουλευτικής Επιτροπής κ. Πολίτη Κωνσταντίνο και τον κ. Τζαβελά Γεώργιο, Αναπληρωτές Καθηγητές του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς.

Θα ήθελα να ευχαριστήσω τον πολύ καλό φίλο και συνάδελφο Παπαδάκη Κωνσταντίνο, ο οποίος καθ' όλη τη διάρκεια της διπλωματικής εργασίας συνέδραμε τα μέγιστα με την πολύτιμη βοήθεια και τις συμβουλές του.

Εν συνεχεία, θα ήθελα να δώσω ένα πολύ μεγάλο ευχαριστώ στη σύντροφό μου Μπρίκου Χρυσούλα, η οποία με την καθημερινή υποστήριξή της με βοήθησε στην ολοκλήρωση της διπλωματικής εργασίας.

Κλείνοντας, θα ήθελα να ευχαριστήσω βαθύτατα τον πατέρα μου Γούτο Θεόδωρο, την μητέρα μου Λιάκου Ευαγγελία και τον αδελφό μου Γούτο Δημήτρη, οι οποίοι όλα αυτά τα χρόνια είναι δίπλα μου και με στηρίζουν σε κάθε μου προσπάθεια.

Περίληψη

Η ανάλυση παλινδρόμηση αποτελεί ένα από τα βασικότερα εργαλεία της σύγχρονης στατιστικής. Οι εκτιμήσεις των συντελεστών ενός μοντέλου παλινδρόμησης υπολογίζονται, προκειμένου να περιγράψουν σχέσεις σε πολυμεταβλητά δεδομένα και να διατυπωθούν προβλέψεις. Στα κλασικά γραμμικά μοντέλα παλινδρόμησης οι ερμηνείες τους είναι άμεσες. Ωστόσο, η γενικότητα του πλαισίου παλινδρόμησης σημαίνει ό,τι είναι ευρέως γενικευμένη για να εξεταστούν πιο πολύπλοκες σχέσεις, συμπεριλαμβανομένης της εξειδίκευσης των μη γραμμικών σχέσεων μεταξύ των επεξηγηματικών μεταβλητών και της μεταβλητής απόκρισης. Με αυτή την ευελιξία να προσδιοριστούν δυνητικά πολύπλοκες σχέσεις πολλών μεταβλητών, υπάρχει κίνδυνος παρερμηνείας. Σκοπός της παρούσας εργασίας είναι η παρουσίαση, η ερμηνεία και η εφαρμογή των Οριακών Επιδράσεων, κυρίως για τα γενικευμένα γραμμικά μοντέλα. Οι Οριακές Επιδράσεις, είναι ένα σημαντικό εργαλείο της συμπερασματικής στατιστικής, αποτελώντας ένα μέτρο υπολογισμού της επίδρασης μίας αλλαγής μιας επεξηγηματικής μεταβλητής στην μεταβλητή απόκρισης, όπου η αλλαγή αυτή μετριέται στις φυσικές της μονάδες.

Η δομή της διπλωματικής εργασίας αποτελείται από 6 Κεφάλαια. Στο 1^{ο} Κεφάλαιο, παρουσιάζεται το θεωρητικό υπόβαθρο των γενικευμένων γραμμικών μοντέλων. Ειδικότερα για τα μοντέλα με δίτιμες μεταβλητές απόκρισης (logit, probit & complementary log log) και για τα μοντέλα με μεταβλητές απόκρισης απαριθμήσεων (Poisson & zero inflated Poisson). Ακολουθεί το 2^{ο} Κεφάλαιο στο οποίο γίνεται η εισαγωγή των Οριακών Επιδράσεων και ο τρόπος υπολογισμού τους. Έπειτα στο 3^{ο} Κεφάλαιο γίνεται αναλυτική ανασκόπηση των προσεγγίσεων, συμφωνά με τις οποίες μπορούν να υπολογιστούν οι Οριακές Επιδράσεις, παρέχοντας επιστημονικά επιχειρήματα ως προς την καταλληλόλητά τους. Στην συνέχεια, στο 4^{ο} Κεφάλαιο περιγράφονται αναλυτικά οι μέθοδοι που χρησιμοποιούνται προκειμένου να εκτιμηθούν τα ασυμπτωτικά τυπικά σφάλματα των Οριακών Επιδράσεων, δίνοντας περισσότερη έμφαση στην μέθοδο Δέλτα. Εν συνεχεία στο 5^{ο} Κεφάλαιο, διατυπώνεται πώς από μια απλή τυποποίηση των επεξηγηματικών μεταβλητών, οι Οριακές Επιδράσεις στα γενικευμένα γραμμικά μοντέλα logit και probit, μπορούν να απλοποιηθούν. Η εργασία ολοκληρώνεται με το 6^{ο} Κεφάλαιο, στο οποίο εφαρμόζεται η θεωρία των Οριακών Επιδράσεων που αναπτύχθηκε στα προηγούμενα Κεφάλαια, στην βάση δεδομένων που σχετίζεται με το σπάνιο είδος καβουριού που ονομάζεται πεταλοειδές καβούρι (Limulus polyphemus), καταλήγοντας σε αξιοσημείωτα συμπεράσματα. Για την στατιστική ανάλυση των δεδομένων χρησιμοποιείται η γλώσσα προγραμματισμού R.

Abstract

Regression analysis is one of the most essential tools in modern Statistics. The estimates of the coefficients of explanatory variables are calculated to describe potential relationships in multivariate data and make predictions. For the classic Linear Regression Models, the interpretation of the estimates is straight forward. However, the regression framework is so wide that it allows to consider more complex relationships, including nonlinear relationships between the explanatory and the response variables. This provides the flexibility to identify potential complex relationship between many variables. On the other hand, there is a risk of misinterpretation. The purpose of this M.Sc. Thesis is the presentation, interpretation and application of the Marginal Effects, with a focus on the generalized linear models. Marginal Effects is an important tool for Statistical inference as they provide a measure of calculating the effect of a change of an explanatory variable on the response variable, where this change is measured in its physical units.

This M.Sc. Thesis consists of 6 Chapters. In Chapter 1, we present the generalized linear models, emphasizing on the models for binary outcomes (logit, probit and complementary log-log) and for count outcomes (Poisson and zero inflated Poisson). In chapter 2, we introduce the Marginal effects and we describe the ways of calculating them. Chapter 3 consists of a detailed overview of the various approaches in which the Marginal Effects can be calculated with. Also, it includes the comparison for two of the presented methods. In Chapter 4, we describe the methods used to estimate the asymptotic standard errors of the Marginal Effects, emphasizing on the Delta method. In Chapter 5, we present the simplified models of Marginal Effects for the generalized linear models logit and probit after normalization of the explanatory variables. In Chapter 6, we apply the theoretical foundation of Marginal Effects. For the Case Study, a dataset related to a rare species of a crab, called horseshoe crab (Limulus polyphemus), is used ending up to remarkable conclusions. For the computational part of this thesis, we used the R Programming Language.

Περιεχόμενα

Κεφάλαιο 1

1.1 Εισαγωγή στα Γενικευμένα Γραμμικά Μοντέλα

Τα γενικευμένα γραμμικά μοντέλα (Generalized Linear Models ή GLMs), αρχικά αναπτύχθηκαν από τους John Nelder και Robert Wedderburn (1972), ωστόσο σημαντική ήταν και η συμβολή του Peter McCullagh, ο οποίος μαζί με τον John Nelder είναι οι συγγραφείς του βιβλίου Generalized Linear Models του 1983. Το μεγαλύτερο ποσοστό της θεωρίας των γενικευμένων γραμμικών μοντέλων, δεν αποτελεί άγνωστο πεδίο στο χώρο της στατιστικής. Στην ουσία τα γενικευμένα γραμμικά μοντέλα συνθέτουν την γενίκευση των κλασσικών γραμμικών μοντέλων. Αποτελούν μία σύνδεση και επέκταση γνωστών μοντέλων παλινδρόμησης τα οποία εμφανίζουν κοινές ιδιότητες και έχουν κοινή μέθοδο εκτίμησης παραμέτρων.

1.2 Δομή των Γενικευμένων Γραμμικών Μοντέλων

Ένα Γενικευμένο Γραμμικό Μοντέλο (*GLM*), αποτελείται από τρεις συνιστώσες:

i. Τυχαία συνιστώσα (*random component*)

Η τυχαία συνιστώσα, αποτελείται από την μεταβλητή απόκρισης Y_i με $i = 1, .., n$ ανεξάρτητες παρατηρήσεις και την κατανομή της, η οποία σύμφωνα με την αρχική θεμελίωση των John Nelder και Robert Wedderburn το 1972, είναι ένα μέλος μίας εκθετικής οικογένειας κατανομών. Ορισμένες κατανομές που ανήκουν στις εκθετικές οικογένειες κατανομών, είναι η κανονική, η διωνυμική, η Poisson, η εκθετική κ.α.

ii. Γραμμική συνάρτηση πρόβλεψης (*linear predictor*)

Η γραμμική συνάρτηση πρόβλεψης, ενσωματώνει τις πληροφορίες των επεξηγηματικών μεταβλητών μέσα στο μοντέλο. Έστω $\boldsymbol{x}_i = \left(1, x_{i1}, x_{i2}, \dots, x_{ip}\right)'$, είναι ένα διάνυσμα των τιμών των επεξηγηματικών μεταβλητών για την παρατήρηση Y_i. Οι x_{ij}, είναι προκαθορισμένες συναρτήσεις των επεξηγηματικών μεταβλητών, ως εκ τούτου μπορεί να περιλαμβάνει είτε συνεχείς είτε κατηγορικές επεξηγηματικές μεταβλητές, όπως συμβαίνει και σε ένα κλασσικό γραμμικό μοντέλο. Επί πλέον μπορεί να περιλαμβάνει δυνάμεις επεξηγηματικών μεταβλητών, αλληλεπιδράσεις και ούτω καθεξής.

Η γραμμική συνάρτηση πρόβλεψης συμβολίζεται με το ελληνικό γράμμα η ("eta"). Εκφράζει παραμέτρους η_i που σχετίζονται με την μέση τιμή $E(Y_i)$ για κάποιους παραμέτρους $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ και επεξηγηματικές μεταβλητές $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ip})'$ και ορίζεται ως εξής:

$$\eta_i = \eta_i(\beta) = \boldsymbol{x}_i{}'\boldsymbol{\beta}$$

$$\Leftrightarrow$$

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

iii. Συνάρτηση σύνδεσης (*link function*)

Η συνάρτηση σύνδεσης $g(\cdot)$, η οποία είναι μονότονη και παραγωγίσιμη, μετασχηματίζει την αναμενόμενη τιμή της μεταβλητής απόκρισης $\mu_i = E(Y_i)$ στη γραμμική πρόβλεψη, δηλαδή:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

Επειδή η συνάρτηση σύνδεσης είναι αντιστρέψιμη μπορεί να γραφτεί και ως εξής:

$$\mu_i = g^{-1}(\eta_i) = g^{-1}\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right)$$

Από το προηγούμενο τύπο, συμπεραίνει κανείς ό,τι τα GLM, μπορούν να θεωρηθούν ως γραμμικά μοντέλα τα οποία έχουν μετασχηματισμένη μεταβλητή απόκρισης ή ως μη γραμμικά μοντέλα παλινδρόμησης για την μεταβλητή απόκρισης. Η συνάρτηση σύνδεσης είναι και ο λόγος όπου οι εκτιμώμενη συντελεστές ενός GLM, λόγο της έλλειψης γραμμικότητας, χάνουν την άμεση ερμηνεία τους.

Στον επόμενο Πίνακα 1.1, παρουσιάζονται οι συναρτήσεις σύνδεσης και οι αντίστροφές τους, που χρησιμοποιούνται ως επί το πλείστον.

Πίνακας 1.1: Ορισμένες σύνηθες συναρτήσεις και οι αντίστροφές τους

Σύνδεση	$\boldsymbol{\eta_i = g(\mu_i)}$	$\boldsymbol{\mu_i = g^{-1}(\eta_i)}$
Ταυτοτική	μ_i	η_i
Log	$\log \mu_i$	e^{η_i}
Αντίστροφη	${\mu_i}^{-1}$	${\eta_i}^{-1}$
Τετραγωνική ρίζα	$\sqrt{\mu_i}$	${\eta_i}^2$
Logit	$\log \frac{\mu_i}{1-\mu_i}$	$\frac{e^{\eta_i}}{1+e^{\eta_i}}$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Complementary log-log	$\log[-\log \mu_i(1-\mu_i)]$	$1 - \exp[-\exp(\eta_i)]$

Σημείωση: Το $\Phi(\cdot)$, δηλώνει την αθροιστική συνάρτηση της κανονικής κατανομής.

1.3 Γενικευμένα Γραμμικά Μοντέλα με Δίτιμες Μεταβλητές Απόκρισης

Τα μοντέλα παλινδρόμησης των οποίων η μεταβλητή απόκρισης λαμβάνει διακριτές τιμές, ονομάζονται μοντέλα διακριτής απόκρισης (discrete outcome) ή μοντέλα ποιοτικής απόκρισης (qualitative response models) (Cameron & Trivedi, 2005). Στην παρούσα εργασία, δίνεται ιδιαίτερη προσοχή κυρίως στα γενικευμένα γραμμικά μοντέλα με δίτιμη μεταβλητή απόκρισης (binary outcomes) και στα μοντέλα με μεταβλητή απόκρισης για απαριθμήσεις (count models). Τα μοντέλα δίτιμης μεταβλητής απόκρισης, είναι το θεμέλιο από τα οποία μπορούν να εξαχθούν πιο περίπλοκα μοντέλα, όπως τα διατακτικά (ordinal), τα κατηγορικά (nominal) και τα μοντέλα απαριθμήσεων (count models). Τα διατακτικά και κατηγορικά μοντέλα παλινδρόμησης ισοδυναμούν με την ταυτόχρονη εκτίμηση μιας σειράς δυαδικών αποτελεσμάτων.

Η κατανομή Bernoulli

Τα μοντέλα δίτιμης μεταβλητής απόκρισης, είναι εκείνα των οποίων η μεταβλητή απόκρισης μπορεί να λάβει δύο ακριβώς τιμές. Τέτοιου είδους μοντέλα, είναι απλά και η εκτίμηση τους γίνεται συνήθως με τη μέθοδο μέγιστης πιθανοφάνειας, επειδή η κατανομή των δεδομένων τους, καθορίζεται από την κατανομή Bernoulli («επιτυχία» ή «αποτυχία»). Επί πλέον επιτρέπουν στον ερευνητή να διερευνήσει πώς κάθε επεξηγηματική μεταβλητή επηρεάζει την πιθανότητα εμφάνισης ενός συμβάν. Σημειώνεται, ό,τι η συγκεκριμένη μεταβλητή απόκρισης, ενδέχεται να είναι και ποσοτική, όπου με κατάλληλες ενέργειες, ο εκάστοτε στατιστικός αναλυτής μπορεί να την μετατρέψει σε δίτιμη. Αυτό πραγματοποιείται ορίζοντας ένα σημείο αναφοράς έστω *τ*, για το οποίο αν οι τιμές της μεταβλητής απόκρισης ξεπερνάνε την τιμή *τ*, να λαμβάνουν μια συγκεκριμένη τιμή (για παράδειγμα την τιμή 1) και αν είναι μικρότερες από την τιμή τ να λαμβάνει μία άλλη τιμή (έστω το 0). Αυτό το μοντέλο ονομάζεται, μοντέλο "κρυμμένης" απόκρισης (latent variable model), για περισσότερες πληροφορίες, μπορεί να ανατρέξει κάποιος στο βιβλίο του Long (1997). Η κλασσική επιλογή των τιμών που κωδικοποιούνται οι κατηγορίες της μεταβλητής απόκρισης είναι, το 0 (αναφέρεται κυρίως σε αρνητική έκβαση) και το 1 (αναφέρεται κυρίως σε θετική έκβαση), ώστε η κατανομή της απόκρισης Y_i, να είναι κατανομή *Bernoulli* με συνάρτηση πυκνότητας πιθανότητας:

$$f(Y_i) = P(Y_i = y_i) = \begin{cases} {\pi_i}^{y_i}(1-\pi_i)^{1-y_i}, & \gamma\iota\alpha\ y_i = 0\ \eta\ 1 \\ 0, & \alpha\lambda\lambda o\acute{\upsilon} \end{cases} \qquad (1.3)$$

Συνεπώς η πιθανότητα ενός αποτελέσματος, όταν $y_i = 1$, ισούται με $P(Y_i = 1) = \pi_i$, ενώ για $y_i = 0$, είναι ίση με $P(Y_i = 0) = 1 - \pi_i$.

Επί πλέον επαληθεύεται άμεσα ύστερα από υπολογισμό, ό,τι η αναμενόμενη τιμή και η διασπορά ισούνται με:

$$\begin{aligned} E(Y_i) &= \mu_i = \pi_i \\ \mathrm{Var}(Y_i) &= \sigma_i^2 = \pi_i\,(1-\pi_i) \end{aligned} \qquad (1.4)$$

Από τους παραπάνω τύπους, φαίνεται ό,τι τόσο η αναμενόμενη τιμή, όσο και η διασπορά εξαρτάται από την πιθανότητα π_i. Ως εκ τούτου, οποιοσδήποτε παράγοντας επηρεάσει την πιθανότητα π_i, θα επηρεάσει και την μέση τιμή και την διασπορά.

Οι παραπάνω τύποι ισχύουν όταν τα δεδομένα εκφράζονται ως ένα διάνυσμα $(Y_1, Y_2, \ldots\ Y_n)$, όπου η i-οστή μεταβλητή απόκρισης $Y_i = 0$ ή 1, παριστάνει το αποτέλεσμα του i-οστού πειράματος.

Όταν όμως η μεταβλητή απόκρισης Y_i λαμβάνει τιμές $1, 2, \ldots..,\ n_i$, τότε η Y_i παριστάνει την συχνότητα εμφάνιση της σε n_i ανεξάρτητα πειράματα *Bernoulli*, τα οποία έχουν όλα πιθανότητα εμφάνισης π_i. Συνεπώς η Y_i ακολουθεί την διωνυμική κατανομή, με παραμέτρους n_i και π_i.

$$Y_i \sim B(n_i, \pi_i)$$

Με συνάρτηση πυκνότητας πιθανότητας:

$$f(Y_i) = P(Y_i = y_i) = \begin{cases} \binom{n_i}{y_i} \pi_i{}^{y_i}(1-\pi_i)^{n_i - y_i}, & \gamma\iota\alpha\ y_i = 0 \text{ ή } 1 \\ 0, & \alpha\lambda\lambda\circ\acute{\upsilon} \end{cases} \qquad (1.5)$$

Παρομοίως, η αναμενόμενη τιμή και η διασπορά προκύπτουν αντιστοίχως:

$$\begin{aligned} E(Y_i) &= \mu_i = n_i\pi_i \\ \mathrm{Var}(Y_i) &= \sigma_i^2 = n_i\pi_i\,(1-\pi_i) \end{aligned} \qquad (1.6)$$

Το επόμενο βήμα είναι να καθοριστεί ένα μοντέλο παλινδρόμησης το οποία θα εκφράζει την πιθανότητα εμφάνισης ενός χαρακτηριστικού π_i. Ο πιο απλός τρόπος είναι να εκφραστεί ως ένα κλασσικό γραμμικό μοντέλο παλινδρόμησης, δηλαδή:

$$\pi_i = \boldsymbol{x}_i{}'\boldsymbol{\beta} \qquad (1.7)$$

Το μοντέλο (1.7), είναι γνωστό ως γραμμικό μοντέλο πιθανότητας (linear probability model- LBM), οι συντελεστές του οποίου εκτιμώνται με την μέθοδο ελαχίστων τετραγώνων. Το πρόβλημα που προκύπτει όμως, είναι ό,τι η δεξιά πλευρά της εξίσωσης, αποτελεί μία πιθανότητα π_i, συνεπώς έχει εύρος τιμών από το 0 έως το 1, αντίθετα η γραμμική συνάρτηση πρόβλεψης $\boldsymbol{x}_i{}'\boldsymbol{\beta}$ στην δεξιά πλευρά της εξίσωσης μπορεί να πάρει οποιαδήποτε πραγματική τιμή. Ένα τέτοιο μοντέλο αντιστοιχεί λανθασμένες τιμές για της προβλεπόμενες πιθανότητες και παράγει αρνητικές διακυμάνσεις, για αυτούς τους λόγους δεν χρησιμοποιείται συχνά.

Για αυτό τον λόγο απαιτείται μια προσέγγιση που χρησιμοποιεί μια μη γραμμική συνάρτηση για τη μοντελοποίηση της συνάρτησης πιθανότητας υπό όρους μιας δυαδικής

εξαρτώμενης μεταβλητής (Greene, 1997). Αυτή η προσέγγιση, είναι να χρησιμοποιηθούν στον τύπο (1.7) συναρτήσεις σύνδεσης $F(\cdot)$ (όπου $F(\cdot)$ δηλώνει μία αθροιστική συνάρτηση κατανομής), οι οποίες να παράγουν στο εύρος (0, 1) τις τιμές της γραμμικής συνάρτησης πρόβλεψης $\boldsymbol{x}'\boldsymbol{\beta}$, δηλαδή:

$$\pi = F(\boldsymbol{x}'\boldsymbol{\beta}) \qquad (1.8)$$

Έτσι ώστε να ισχύει:

$$P(Y = 1|\boldsymbol{x}) = 1$$

$$P(Y_i = 0|\boldsymbol{x}) = 0 \qquad (1.9)$$

1.3.1 Λογιστική παλινδρόμηση (Logistic regression)

Το μοντέλο logit, είναι ευρέως γνωστό στον τομέα της Στατιστικής και ως μοντέλο λογιστικής παλινδρόμησης (Logistic Regression Model). Το μοντέλο logit, είναι μία ειδική περίπτωση ενός γενικευμένου γραμμικού μοντέλου. Χρησιμοποιείται για την περιγραφή δεδομένων και για την διερεύνηση της σχέσης μεταξύ μίας κατηγορικής μεταβλητής απόκρισης και μιας ή περισσότερων επεξηγηματικών μεταβλητών, η σύνδεση των οποίων γίνεται μέσω της συνάρτησης σύνδεσης logit. Ανάλογα την φύση των κατηγοριών της μεταβλητής απόκρισης, η λογιστική παλινδρόμηση, ανήκει σε τρεις κατηγορίες μοντέλων, τη διωνυμική λογιστική παλινδρόμηση (μόνο δύο κατηγορίες), την πολυωνυμική λογιστική παλινδρόμηση (με περισσότερες από δύο κατηγορίες) και την διατακτική λογιστική παλινδρόμηση (όταν οι κατηγορίες διατάσσονται με αυξητική τάση). Παρ' όλα αυτά, δίνεται μεγαλύτερη προσοχή στη διωνυμική λογιστική παλινδρόμηση, διό,τι είναι και αυτή που χρησιμοποιείται ως επί το πλείστον σε περισσότερες εφαρμογές.

Στην ουσία η λογιστική παλινδρόμηση, είναι ένα μοντέλο το οποία ταξινομεί τις τιμές της μεταβλητής απόκρισης, με βάση τη θεωρία των πιθανοτήτων. Η διαφορά της με την γραμμική παλινδρόμηση βασίζεται στο ό,τι η μεταβλητή απόκρισης της είναι κατηγορική, αντιθέτως στην γραμμική παλινδρόμηση είναι ποσοτική. Επί πλέον, στην κλασσική γραμμική παλινδρόμηση, η εκτίμηση των παραμέτρων γίνεται με βάση τη μέθοδο των ελαχίστων τετραγώνων, ενώ στην λογιστική παλινδρόμηση γίνεται με τη μέθοδο μέγιστης πιθανοφάνειας (που όπως προαναφέρθηκε είναι η μέθοδος που εφαρμόζεται συνήθως στα GLMs).

Το μοντέλο logit χρησιμοποιείται σε διάφορους επιστημονικούς κλάδους, όπως στις πολιτικές και κοινωνικές επιστήμες, στα περισσότερα ιατρικά πεδία, στο machine learning, στην οικονομία και ούτω καθεξής. Για μία πιο λεπτομερή περιγραφή των μεθόδων της λογιστικής παλινδρόμησης, μπορεί να ανατρέξει κάποιος στα συγγράμματα των Cox & Snell (1989), των Hosmer & Lemeshow (2000) και των Long & Freese (2014).

Συνάρτηση σύνδεσης logit

Στη Στατιστική, η συνάρτηση σύνδεσης logit, είναι ο λογάριθμος της σχετικής πιθανότητας (odds) για ένα γεγονός. Η σχετική πιθανότητα, ενός ενδεχομένου για $\pi \epsilon (0,1)$, ισούται με τον λόγο της πιθανότητας εμφάνισης π, ως προς την πιθανότητα μη εμφάνισης 1- π . Δηλαδή:

$$\text{odds} = \frac{\pi}{1-\pi}, \qquad \text{odds } \epsilon\ (0, \infty)$$

$$\log \text{odds} = \text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right), \qquad \text{LO } \epsilon\ (-\infty, +\infty)$$

Η συνάρτηση logit, έχει και πεδίο ορισμού το διάστημα (0, 1) και πεδίο τιμών το $\mathbb{R}$. Οπότε για $\pi\epsilon\ (0,1)$, ορίζεται ως ο λογάριθμος της σχετική πιθανότητας (log odds - LO):

Με αντίστροφη συνάρτηση:

$$\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$$

Επί πλέον η συνάρτηση logit, είναι γνησίως αύξουσα και ισχύει ό,τι:

$$\pi_i < \frac{1}{2} \Leftrightarrow odds < 1 \Leftrightarrow \text{logit}(\pi_i) < 0$$

$$\pi_i = \frac{1}{2} \Leftrightarrow odds = 1 \Leftrightarrow \text{logit}(\pi_i) = 0$$

$$\pi_i > \frac{1}{2} \Leftrightarrow odds > 1 \Leftrightarrow \text{logit}(\pi_i) > 0$$

Το μοντέλο Logit

Όπως διατυπώθηκε προηγουμένως, το μοντέλο logit, αποτελεί ένα μοντέλο GLM, με δίτιμη απόκριση και συνάρτηση σύνδεσης την logit. Στην ουσία η αθροιστική συνάρτηση κατανομής αυτού του μοντέλου, είναι η τυπική λογιστική κατανομή.

$$\Lambda(x) = \frac{e^x}{1+e^x}, \qquad \gamma\iota\alpha\ x \in \mathbb{R}$$

Με συνάρτηση πυκνότητας πιθανότητας:

$$l(x) = \Lambda'(x) = \frac{e^x}{(1+e^x)^2}, \qquad \gamma\iota\alpha\ x \in \mathbb{R}$$

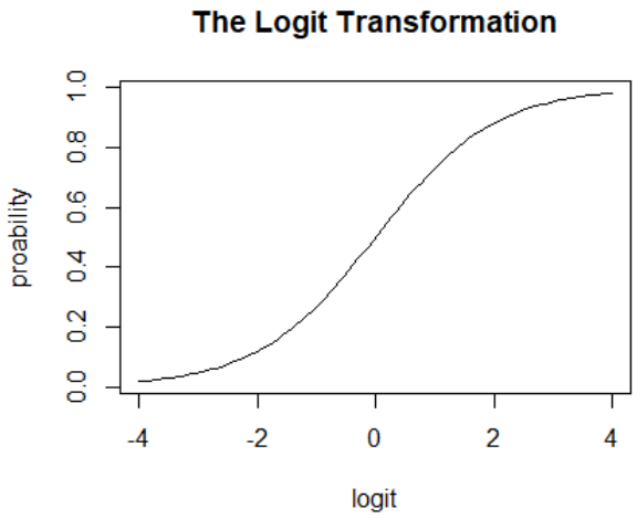
Η τυπική λογιστική κατανομή είναι συμμετρική ως προς το 0. Έχει μέση τιμή 0 και διασπορά $\pi^2/3$.

Έστω τώρα ό,τι δίνεται μια δίτιμη μεταβλητή απόκρισης Y_i, για την οποία ισχύει $E(Y) = P(Y = 1) = \pi$ και για κάποιες παραμέτρους $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ και επεξηγηματικές μεταβλητές $\boldsymbol{x} = (1, x_1, x_2, \dots, x_p)'$. Αντικαθιστώντας την αθροιστική συνάρτηση κατανομής $F(\cdot)$, της εξίσωσης (1.8), με την τυπική λογιστική κατανομή $\Lambda(\cdot)$, παράγεται το μοντέλο logit:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{x}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (1.10)$$

Αντιστρέφοντας τώρα την παραπάνω συνάρτηση, προκύπτει:

$$\pi = \Lambda(\boldsymbol{x}'\boldsymbol{\beta}) = \frac{e^{x'\boldsymbol{\beta}}}{1 + e^{x'\boldsymbol{\beta}}} \qquad (1.11)$$



Σχήμα 1.2: Ο μετασχηματισμός logit

Από το **Σχήμα 1.2**, παρατηρείται ό,τι καμπύλη της συνάρτηση σύνδεσης logit, είναι μία σιγμοειδής συνάρτηση και χαρακτηρίζεται από ένα στάδιο εκθετικής ανάπτυξης, στην συνέχεια ο ρυθμός αύξησης της μειώνεται σταδιακά με αποτέλεσμα το αριστερό και δεξιό πέρας της συνάρτησης να καταλήγουν σε ευθείες σχεδόν παράλληλες στον οριζόντιο άξονα X.

1.3.2 Η παλινδρόμηση Probit (Probit regression)

Όπως το μοντέλο logit, έτσι και το μοντέλο probit είναι ένα δημοφιλές μοντέλο δυαδικής απόκρισης και αποτελεί ένα είδος μοντέλου GLM. Επί πλέον η εκτίμηση των παραμέτρων και σε αυτό το μοντέλο, γίνεται με τη μέθοδο μέγιστης πιθανοφάνειας. Το μοντέλο προτάθηκε από τον Bliss (1935), ο οποίος είναι και αυτός που έδωσε το όνομα από τα μέρη των λέξεων ***prob****ability un****it***. Ο Bliss, χρησιμοποίησε το μοντέλο αυτό, σε διάφορα πειράματα τοξικολογίας (McCullagh, P. & Nelder, J. A., 1989). Πιο συγκεκριμένα, χώριζε τα πειραματόζωα ή έντομα σε ομάδες, έστω i το πλήθος, όχι απαραίτητα ίσου μεγέθους και στην συνέχεια σε κάθε ομάδα χορηγούσε μία δόση *x* ενός φαρμάκου, η οποία διέφερε από ομάδα σε ομάδα. Έπειτα κατέγραφε από την i ομάδα, τον αριθμό Y_i των πειραματόζωων που επιβιώνει από το αρχικό πλήθος m_i, καθώς και τη δόση *x* που είχε χορηγήσει. Για αυτόν τον λόγο συναρτήση της δόσης, χρησιμοποίησε το μοντέλο probit. Παρ' όλα αυτά, το μοντέλο probit, είναι ένα πολύ σημαντικό εργαλείο για τους οικονομολόγους, όταν θέλουν να μοντελοποιήσουν δίτιμες αποκρίσεις.

Το μοντέλο Probit

Η προσαρμογή του δεν διαφέρει πολύ από το μοντέλο logit. Στο μοντέλο probit, η συνάρτηση σύνδεσης είναι η probit. Σε αυτό το μοντέλο, η αθροιστική συνάρτηση κατανομής είναι η τυπική κανονική κατανομή $\Phi(\cdot)$, δηλαδή:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\, e^{-t^2/2}\, dt\,, \qquad \gamma\iota\alpha\ x \in \mathbb{R}$$

Με συνάρτηση πυκνότητας πιθανότητας:

$$\varphi(x) = \Phi'(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2} \quad , \qquad \gamma\iota\alpha\ x \in \mathbb{R}$$

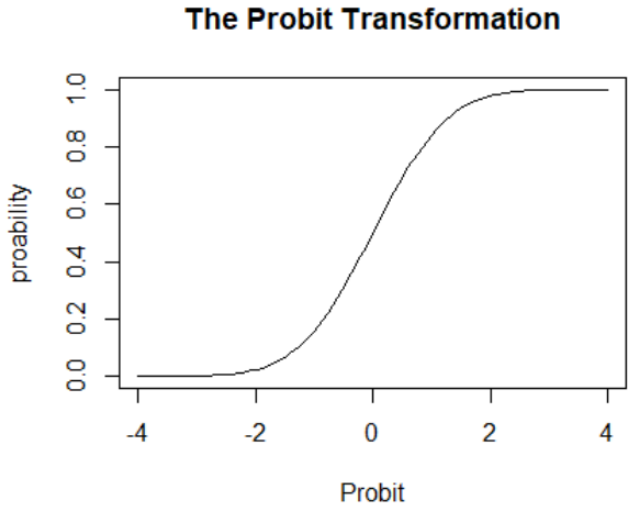
Από την θεωρία είναι γνωστό ό,τι η τυπική κανονική κατανομή είναι συμμετρική στο 0, συνεπώς έχει μέση τιμή 0 και διασπορά 1.

Παίρνοντας τώρα την εξίσωση (1.8) και αντικαθιστώντας την $F(\cdot)$ με τη τυπική κανονική κατανομή $\Phi(\cdot)$, προκύπτει το μοντέλο probit.

$$\pi = \Phi(\boldsymbol{x}'\boldsymbol{\beta}) \qquad (1.12)$$

Με αντίστροφη συνάρτηση:

$$\Phi^{-1}(\pi) = \boldsymbol{x}'\boldsymbol{\beta} \qquad (1.13)$$



Σχήμα 1.3: Ο μετασχηματισμός probit

Από την **Σχήμα 1.3**, παρατηρείται ό,τι καμπύλη της συνάρτηση σύνδεσης probit, έχει τις ίδιες ιδιότητες παρόμοια με τη συνάρτηση logit. Αυτό οφείλεται στο γεγονός ό,τι για κατάλληλους παραμέτρους η τυπική λογιστική κατανομή δεν διαφέρει με την τυπική κανονική κατανομή.

1.3.3 Η παλινδρόμηση Complementary log-log (Complementary log-log regression)

Το μοντέλο complementary log-log εμφανίσθηκε πρώτο στη βιβλιογραφία το 1922 από τον Άγγλο στατιστικό R.A. Ficher. Ο Fisher περιέγραψε ένα πείραμα στο οποίο πήρε ένα δείγμα εδάφους ή νερού και πραγματοποίησε μια σειρά αραιώσεων του δείγματος, με σκοπό να προσδιοριστεί η παρουσία ή η απουσία κάποιου μικροβιακού μολυσματικού παράγοντα. Για αυτό τον λόγο χρησιμοποίησε έναν μετασχηματισμό cloglog και εφάρμοσε την εκτίμηση της μέγιστης πιθανότητας (Piegorsch, 1992).

Το μοντέλο Complementary log-log

Η αθροιστική συνάρτηση κατανομής αυτού του μοντέλου, είναι η τυπική κατανομή Gumbel.

$$W(x) = 1 - e^{-e^{x}}, \qquad \gamma\iota\alpha\ x \in \mathbb{R}$$

Με συνάρτηση πυκνότητας πιθανότητας:

$$w(x) = W'(x) = e^{x}\big(1 - W(x)\big) = e^{x - e^{x}}, \qquad \gamma\iota\alpha\ x \in \mathbb{R}$$

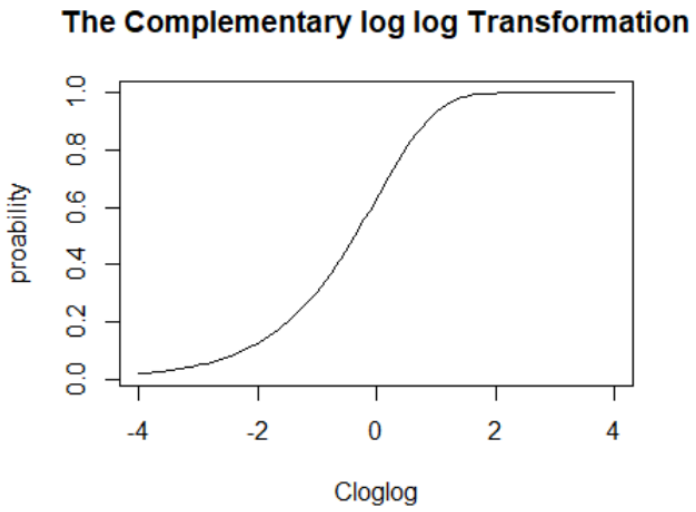
Η τυπική κατανομή Gumbel, σε αντίθεση με τις προηγούμενες τυπικές κατανομές δεν είναι συμμετρική στο 0. Έχει μέση τιμή περίπου ίση με -0.5772 και διασπορά ίση με $\pi^2/6$. Επί πλέον μία ιδιαιτερότητα αυτού του μοντέλου, είναι ό,τι έχει καλύτερη προσαρμογή σε περιπτώσεις που η π αργεί να πλησιάσει το 0 ενώ πλησιάζει γρήγορα το 1, ως συνάρτηση των επεξηγηματικών μεταβλητών.

Παίρνοντας την εξίσωση (1.8) και αντικαθιστώντας την $F(\cdot)$ με τη τυπική κατανομή $W(\cdot)$, προκύπτει το μοντέλο complementary log log.

$$\pi = W(\boldsymbol{x'\beta}) = 1 - e^{-e^{x'\beta}} \qquad (1.14)$$

Με αντίστροφη συνάρτηση:

$$\log(-\log(1 - \pi)) = \boldsymbol{x'\beta} \qquad (1.15).$$



Σχήμα 1.4: Ο μετασχηματισμός cloglog

1.4 Γενικευμένα Γραμμικά Μοντέλα με Μεταβλητές Απόκρισης Απαριθμήσεων

Τα μοντέλα με μεταβλητές απόκρισης απαριθμήσεων (count models), αποτελούν ένα υποσύνολο των μοντέλων παλινδρόμησης διακριτής απόκρισης. Στα συγκεκριμένα μοντέλα η μεταβλητή απόκρισης λαμβάνει ακέραιες μη αρνητικές τιμές. Πιο συγκεκριμένα, τα δεδομένα απαριθμήσεων, κατανέμονται ως μη αρνητικοί ακέραιοι αριθμοί, είναι εγγενώς ετεροσκεδαστικοί, με καμπύλη λοξή προς τα δεξιά και έχουν διασπορά που αυξάνεται με τον μέσο όρο (Cameron & Trivedi, 1998). Ορισμένα παραδείγματα δεδομένων απαριθμήσεων περιλαμβάνουν καταστάσεις όπως η διάρκεια της παραμονής στο νοσοκομείο ενός ασθενούς, ο αριθμός ορισμένων ειδών ψαριών ανά καθορισμένη περιοχή στον ωκεανό, ο αριθμός των θανάτων λόγω του covid-19, ο αριθμός πελατών στην τράπεζα κ.ο.κ

1.4.1 Η παλινδρόμηση Poisson (Poisson regression)

Η κατανομή Poisson

Η κατανομή Poisson χρησιμεύει πιο πολύ στην ανάλυση αυτών των δεδομένων και είναι γνωστό ό,τι δίνεται από τον τύπο,

$$f(y) = P(Y = y) = e^{-\mu}\frac{\mu^y}{y!}, \qquad για\ y = 0, 1, \ldots,$$

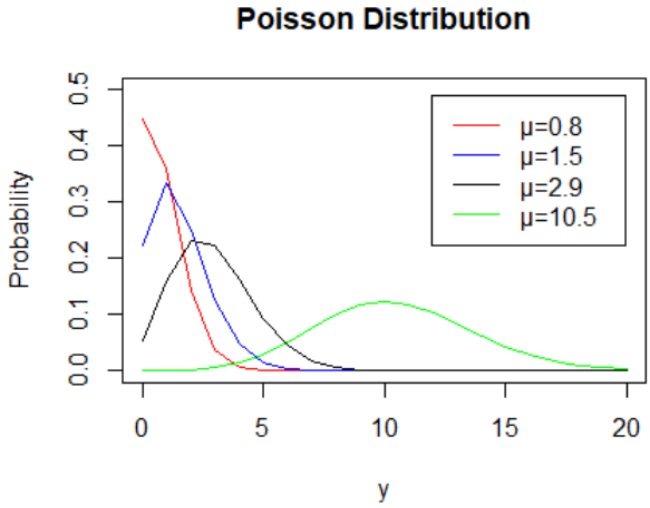
όπου

- μ: είναι μία θετική παράμετρος

Με μέση τιμή και διασπορά:

$$E(Y) = Var(Y) = \mu$$

Προκείμενου να αποκτηθεί μια καλύτερη αίσθηση όσον αφορά την κατανομή Poisson, παρουσιάζεται στο παρακάτω διάγραμμα, οι τιμές των προβλεπόμενων πιθανοτήτων, που προκύπτουν για τις διάφορες τιμές τις παραμέτρου μ.



Το διάγραμμα απεικονίζει τέσσερα χαρακτηριστικά της κατανομής Poisson που είναι σημαντικά για την κατανόηση των μοντέλων παλινδρόμησης για μετρήσεις, αυτά είναι:

- Όσο η μέση τιμή μ της κατανομής μεγαλώνει, τόσο η καμπύλη της κατανομής μετατοπίζεται προς τα δεξιά
- Σε πραγματικά δεδομένα, πολλές μεταβλητές απαριθμήσεων τυγχάνει να έχουν διασπορά μεγαλύτερη από τη μέση τιμή, φαινόμενο το οποίο ονομάζεται *υπερσκέδαση*.
- Καθώς η ποσότητα μ αυξάνεται, μειώνεται η πιθανότητα να υπάρχει μηδενικός αριθμός. Σημειώνεται ό,τι σε πολλές μεταβλητές απαριθμήσεων, ενδέχεται να υπάρχουν περισσότερα μηδενικά που παρατηρούνται από αυτά που προβλέπονται από την κατανομή Poisson.
- Για μεγάλες τιμές του μ, η κατανομή Poisson προσεγγίζει την κανονική κατανομή. Το οποίο παρατηρείται από το παραπάνω διάγραμμα για $\mu = 10.5$

Παλινδρόμηση Poisson

Η Poisson παλινδρόμηση είναι το βασικό μοντέλο στο οποίο βασίζονται πολλά μοντέλων απαριθμήσεων. Στην Poisson παλινδρόμησης, οι μεταβλητές απόκρισης είναι τιμές οι οποίες ακολουθούν την κατανομή Poisson. Πολλές φορές σε εφαρμογές παρατηρούνται δεδομένα συχνοτήτων, για τα οποία η κατανομή Poisson είναι πολύ χρήσιμη στην ανάλυση τους.

Το μοντέλο παλινδρόμησης Poisson προέρχεται από την κατανομή Poisson με παραμετροποίηση της σχέσης μεταξύ της παραμέτρου μ και των επεξηγηματικών μεταβλητών x του μοντέλου. Η πιο συνηθισμένη συνάρτηση σύνδεσης που χρησιμοποιείται είναι ο λογάριθμος. Δηλαδή:

$$\log \mu = \boldsymbol{x}'\boldsymbol{\beta} \qquad (1.16)$$

Το οποίο συνεπάγεται:

$$\mu = e^{x'\beta} \qquad (1.17)$$

Σημειώνεται, ό,τι παρ' όλο που το μοντέλο παλινδρόμησης Poisson είναι δημοφιλές, ο περιορισμός $E(Y) = Var(Y)$, το κάνει πολλές φορές να μην προσαρμόζεται σε ικανοποιητικό βαθμό στα δεδομένα. Πιο συγκεκριμένα, όταν σε μια μεταβλητή απόκρισης η οποία είναι μία απαρίθμηση η διασπορά της αυξάνεται ταχύτερα από την μέση τιμή, τότε παράγεται το φαινόμενο της *υπερσκέδαση*, ακυρώνοντας έτσι την χρήση της κατανομής Poisson. Το φαινόμενο της *υπερσκέδασης* είναι τόσο κοινό σε μοντέλα παλινδρόμησης για δεδομένα απαρίθμησης και η συνέπειες της είναι προκαλούν δυνητικά σοβαρές συνέπειες στην ανάλυση. Για αυτό τον λόγο υπάρχουν ενναλακτικά μοντέλα τα οποία μπορούν να ξεπεράσουν αυτό το πρόβλημα. Ένα από τα οποία θα δοθεί και περισσότερη σημασία στην επόμενη ενότητα, είναι το μοντέλο Zero Inflated Poisson.

1.4.2 Η παλινδρόμηση Zero Inflated Poisson (ZIP regression)

Η κατανομή ZIP

$$f(y) = P(Y = y) = \pi\, \mathrm{I}(y = 0) + (1 - \pi)e^{-\lambda}\frac{\lambda^y}{y!}, \quad \gamma\iota\alpha\ y = 0, 1, \ldots,$$

$$\begin{cases} \pi + (1 - \pi)e^{-\lambda}, & \gamma\iota\alpha\ y = 0 \\ (1 - \pi)e^{-\lambda}\dfrac{\lambda^y}{y!}, & \gamma\iota\alpha\ \ y = 1, 2, \ldots, \end{cases}$$

όπου

- λ: είναι μία θετική παράμετρος
- π: είναι η πιθανότητα
- $\mathrm{I}(\cdot)$: είναι μία δείκτρια συνάρτηση

Με μέση τιμή:

$$E(Y) = \mu = (1 - \pi)\lambda$$

Και διασπορά:

$$Var(Y) = \mu + \mu^2\frac{\pi}{1 - \pi}$$

Από τις παραπάνω σχέσεις, παρατηρείται εύκολα ό,τι:

$$Var(Y) > E(Y)$$

Παλινδρόμηση ZIP

Ένα ιδιαίτερο είδος του φαινομένου της *υπερσκέδασης*, επικρατεί όταν στα δεδομένα υπάρχουν περισσότερα μηδενικά. Αρκετά είναι τα μοντέλα τα οποία έχουν προταθεί για αυτού του είδους τα δεδομένα, ένα από αυτά είναι το μοντέλο της Zero Inflated Poisson (ZIP) παλινδρόμησης (Lambert, 1992). Το ZIP μοντέλο περιλαμβάνεται από δύο συναρτήσεις. Η πρώτη αποτελείται από ένα δυαδικό μοντέλο λογιστικής παλινδρόμησης, του οποίου η μεταβλητή απόκρισης λαμβάνει αποκλειστικά τιμές 0 και η δεύτερη από ένα μοντέλο Poisson παλινδρόμησης, του οποίου η μεταβλητή απόκρισης μπορεί να είναι 0 ή μία θετική απαρίθμηση. Δηλαδή:

1η Συνάρτηση

$$logit(\pi) = \boldsymbol{z}'\boldsymbol{\gamma} \qquad (1.18)$$

$$\Leftrightarrow$$

$$\pi = \frac{e^{\boldsymbol{z}'\boldsymbol{\gamma}}}{1 + e^{\boldsymbol{z}'\boldsymbol{\gamma}}} \qquad (1.19)$$

όπου

- π: αντιπροσωπεύει την πιθανότητα ό,τι η μεταβλητή απόκρισης Y, θα είναι αναγκαστικά μηδέν
- $\boldsymbol{z}' = (1, z_1, \dots, z_p)$: το διάνυσμα των m $(m = p + 1)$ επεξηγηματικών μεταβλητών του μοντέλου logit, συμπεριλαμβανομένου του σταθερού όρου
- $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \dots, \gamma_p)'$: είναι διάνυσμα με τις άγνωστες παραμέτρους του μοντέλου logit

2η Συνάρτηση

$$\log \lambda = \boldsymbol{x}'\boldsymbol{\beta} \qquad (1.20)$$

$$\Leftrightarrow$$

$$\lambda = e^{\boldsymbol{x}'\boldsymbol{\beta}} \qquad (1.21)$$

όπου

- $\boldsymbol{x}' = (1, x_1, \dots, x_p)$: το διάνυσμα των m $(m = p + 1)$ επεξηγηματικών μεταβλητών του λογαριθμογραμμικού μοντέλου, συμπεριλαμβανομένου του σταθερού όρου
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)'$: είναι διάνυσμα με τις άγνωστες παραμέτρους του λογαριθμογραμμικού μοντέλου

Οπότε με βάση τις σχέσεις (1.19) & (1.20), η μέση τιμή, ισούται:

$$\mu = (1-\pi)\lambda = \frac{e^{\boldsymbol{x'\beta}}}{1+e^{\boldsymbol{z'\gamma}}} \qquad (1.22)$$

Και η διασπορά:

$$\sigma^2 = \mu + \mu^2 \frac{\pi}{1-\pi} = \frac{e^{\boldsymbol{x'\beta}}}{1+e^{\boldsymbol{z'\gamma}}} + e^{\boldsymbol{z'\gamma}} \left(\frac{e^{\boldsymbol{x'\beta}}}{1+e^{\boldsymbol{z'\gamma}}}\right)^2 \qquad (1.23)$$

Σημειώνεται, ό,τι οι επεξηγηματικές μεταβλητές $\boldsymbol{x}$ & $\boldsymbol{z}$, οι οποίες χρησιμοποιούνται για τις παραμέτρους λ & π, αντίστοιχα, δεν είναι υποχρεωτικό να ταυτίζονται, αν και συχνά στις εφαρμογές παίρνονται οι ίδιες επεξηγηματικές μεταβλητές.

Κεφάλαιο 2

2.1 Εισαγωγή

Η παλινδρόμηση αποτελεί ένα από τα βασικότερα εργαλεία της σύγχρονης στατιστικής. Σε κλάδους όπως η οικονομία και οι πολιτικές επιστήμες, σχεδόν όλες οι ποσοτικές έρευνες χρησιμοποιούν μοντέλα παλινδρόμησης, προκειμένου να περιγράψουν σχέσεις σε πολυμεταβλητά δεδομένα και να διατυπώσουν προβλέψεις.

Μια ιδιαίτερα ελκυστική διαδικασία λόγω των περιορισμένων και οικείων υποθέσεων και της ευκολίας με την οποία εκφράζει μια πολυμεταβλητή σχέση με μια γραμμική σχέση μεταξύ πολλών μεταβλητών με μια μόνο μεταβλητή, είναι η παλινδρόμηση των ελαχίστων τετραγώνων (Ordinary Least Square). Οι συντελεστές που εκτιμώνται μέσω της διαδικασία OLS είναι συνήθως εύκολα ερμηνεύσιμες ως η αναμενόμενη αύξηση του αποτελέσματος λόγω αλλαγής μονάδας στην αντίστοιχη επεξηγηματική μεταβλητή.

Αυτή η ευκολία ερμηνείας απλών μοντέλων παλινδρόμησης, ωστόσο, υπονομεύει την πιθανότητα τεράστιας αναλυτικής και ερμηνευτικής πολυπλοκότητας. Η γενικότητα του πλαισίου παλινδρόμησης σημαίνει ό,τι είναι ευρέως γενικευμένη για να εξεταστούν πιο πολύπλοκες σχέσεις, συμπεριλαμβανομένης της εξειδίκευσης των μη γραμμικών σχέσεων μεταξύ των επεξηγηματικών μεταβλητών και της μεταβλητής απόκρισης, την ύπαρξη όρων αλληλεπιδράσεων μεταξύ των πολλαπλών μεταβλητών και των μετασχηματισμών μέσω ενός γενικευμένου γραμμικού μοντέλου (GLM). Με αυτή την ευελιξία να προσδιοριστούν δυνητικά πολύπλοκες σχέσεις πολλών μεταβλητών, υπάρχει κίνδυνος παρερμηνείας, και πράγματι συχνά εσφαλμένος υπολογισμός των ποσοτήτων ενδιαφέροντος.

Για τον σκοπό αυτό πολλοί ερευνητές, αποσκοπώντας στη διεξαγωγή ορθών στατιστικών συμπερασμάτων, προτείνουν την χρήση των Οριακών Επιδράσεων (Marginal Effects). Τα Marginal Effects (ΜΕ), είναι ένα σημαντικό εργαλείο της συμπερασματικής στατιστικής, αποτελώντας ένα μέτρο υπολογισμού της επίδρασης μίας "μοναδιαίας" αλλαγής μιας επεξηγηματικής μεταβλητής στην μεταβλητή απόκρισης, όπου η αλλαγή αυτή μετριέται στις φυσικές της μονάδες. Η θεωρητική προέλευση των ΜΕ, περιλαμβάνουν τυπικούς κανόνες παραγώγισης, με αποτέλεσμα ο υπολογισμός τους να είναι καλά ορισμένος.

Στην ουσία το ΜΕ μιας επεξηγηματικής μεταβλητής ενδιαφέροντος, αποτελεί την κλίση της επιφάνειας παλινδρόμησης σε σχέση με μια συγκεκριμένη επεξηγηματική μεταβλητή. Το ΜΕ εκφράζει τον ρυθμό με τον οποίο η αναμενόμενη τιμή της μεταβλητής απόκρισης μεταβάλλεται σε ένα συγκεκριμένο σημείο στο μεταβλητό χώρο, σε σχέση με μια μονοδιάστατη μεταβλητή, διατηρώντας όλες τις υπόλοιπες επεξηγηματικές μεταβλητές

σταθερές. Αυτή η ποσότητα είναι ιδιαίτερα χρήσιμη επειδή μπορεί να υπολογιστεί από ουσιαστικά οποιοδήποτε σύνολο εκτιμήσεων παλινδρόμησης (Leeper, 2018).

Τα ΜΕ συνοψίζουν την επίδραση μιας επεξηγηματικής μεταβλητής ως προς το μοντέλο πρόβλεψης (Long & Freese, 2014). Είναι ένα πολύ σημαντικό εργαλείο και έχει περισσότερα πλεονεκτήματα από το να βασίζεται κάποιος μόνο από στους εκτιμώμενους συντελεστές ενός μοντέλου, προκειμένου να επιτευχθεί η στατιστική συμπερασματολογία. Πιο συγκεκριμένα:

- Επιτρέπουν την διεξαγωγή στατιστικών συμπερασμάτων όσον αφορά την επίδραση μιας επεξηγηματικής μεταβλητής στην μεταβλητή απόκρισης, ακόμα και όταν υπάρχουν μετασχηματισμένες μεταβλητές ή όροι αλληλεπίδρασης (για παράδειγμα x_1 και x_1^2 ή $x_1 x_2$).
- Αποτρέπουν τα προβλήματα ως προς τα ζητήματα αναγνώρισης (κλίμακας) των συντελεστών σε μοντέλα όπως τα logit, probit & complementary log-log (Long 1997, Long & Freese 2014, Breen, Karlson & Holm 2018).
- Δεδομένου ό,τι βασίζονται στις προβλέψεις ενός μοντέλου, μπορούν να εκφραστούν και να ερμηνευθούν με διαφορετική μέτρηση από ό,τι οι εκτιμώμενοι συντελεστές (Mize, 2019).

Ωστόσο, η πρακτική εφαρμογή των ΜΕ, οδηγεί σε μία πληθώρα ζητημάτων, καθώς η καθοδήγηση μέσω βιβλιογραφίας είναι περιορισμένη. Ορισμένα από τα ζητήματα αυτά, τα οποία θα αναλυθούν στις επόμενες ενότητες είναι:

- Η επιλογή και η ερμηνεία μεταξύ των διάφορων προσεγγίσεων ως προς τον υπολογισμό των ΜΕ.
- Η επιλογή του εκτιμητή για τον υπολογισμό διαστημάτων εμπιστοσύνης.
- Οι περιορισμένες επιλογές λογισμικού για τον υπολογισμό των ΜΕ.

2.1.1 Διαφοροποίηση του όρου *marginal effects*

Είναι σημαντικό να αναφερθεί ό,τι επειδή οι ορολογίες είναι παρόμοιες σε ορισμένους κλάδους, είναι πιθανό να δημιουργείται μία σύγχυση, δηλαδή η γλώσσα που προέρχεται από έναν κλάδο να μην μεταφράζεται καλά σε άλλους Πιο συγκεκριμένα με τον όρο marginal effect, ένας στατιστικός μεταφράζοντάς το ως "περιθωριακές" επιδράσεις, αυτόματα έχει κατά νου ένα ολοκλήρωμα λόγω της περιθωριακής συνάρτησης πυκνότητας πιθανότητας (όπου σε ένα πίνακα συνάφειας, οι πιθανότητες που βρίσκονται στο "περιθώριο" του πίνακα, είναι οι περιθωριακές πιθανότητες). Από τη σκοπιά όμως των οικονομολόγων, μεταφράζοντας τον όρο marginal ως "οριακό", εννοούνε την "οριακή μεταβολή" η οποία είναι η παράγωγος.

Στην παρούσα εργασία ο όρος marginal effect, ενώ είναι εργαλείο της στατιστικής συμπερασματολογίας, θα μεταφράζεται ως οριακές επιδράσεις και θα έχει την έννοια της

οριακής μεταβολής (παράγωγος), δηλαδή θα χρησιμοποιείται όπως και στον τομέα της οικονομίας.

2.2 Εισαγωγή στον υπολογισμό των Οριακών Επιδράσεων

Ανάλογα την φύση των επεξηγηματικών μεταβλητών, τα ΜΕ διαφέρουν στο τρόπο υπολογισμού. Στις επόμενες ενότητες, γίνεται μία ανασκόπηση ως προς τον υπολογισμό των διαφόρων περιπτώσεων. Σημειώνεται, ό,τι για τον υπολογισμό των ΜΕ οι υπόλοιπες επεξηγηματικές μεταβλητές λαμβάνουν σταθερές τιμές.

Έστω ό,τι δίνεται η γενική συνάρτηση παλινδρόμησης από τον τύπο:

$$E(Y|\boldsymbol{x}) = h(\boldsymbol{x}'\boldsymbol{\beta}) \qquad (2.1)$$

όπου

- Y: είναι η μεταβλητή απόκρισης
- $\boldsymbol{x} = (1, x_1, \dots, x_p)'$: το διάνυσμα των p επεξηγηματικών μεταβλητών, συμπεριλαμβανομένου του σταθερού όρου
- x_j: είναι η j-οστή επεξηγηματική μεταβλητή, για $j = 1, \dots, p$ επεξηγηματικές μεταβλητές
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$: είναι το διάνυσμα με τις άγνωστες παραμέτρους
- $h(\cdot)$: είναι μια παραγωγίσιμη συνάρτηση

2.2.1 Υπολογισμός Οριακών Επιδράσεων συνεχών επεξηγηματικών μεταβλητών

Τα ΜΕ για τις συνεχείς επεξηγηματικές μεταβλητές ενός μοντέλου, εκφράζουν το στιγμιαίο ρυθμό μεταβολής. Είναι δημοφιλείς σε πολλούς κλάδους επειδή συχνά παρέχουν καλή προσέγγιση όσον αφορά το ποια θα είναι η μεταβολή της μεταβλητής απόκρισης *Y*, σε μία αλλαγή "μονάδας" της x_j, αν και ορισμένες φορές δεν ισχύει (Williams, 2019).

Στην ουσία το ΜΕ μιας συνεχής επεξηγηματικής μεταβλητής x_j, είναι η μερική παράγωγος της μεταβλητής απόκρισης Y ως προς την x_j (Long, 1997). Συνεπώς για τον υπολογισμό του ΜΕ, ο γενικός τύπος είναι:

$$\text{ME}_{x_j} = \frac{\partial E(Y|\boldsymbol{x})}{\partial x_j} \qquad (2.2)$$

Με άλλα λόγια το ΜΕ της x_j, αντιπροσωπεύει την κλίση της συνάρτησης απόκρισης από ένα μοντέλο κάνοντας "μικρά" βήματα, d σε x_i:

$$\mathrm{ME}_{x_j} = \lim_{d \to 0} \frac{E\left(Y = 1 \middle| x_j + d\right) - E\left(Y = 1 \middle| x_j\right)}{d} \qquad (2.3)$$

Είναι πολύ σημαντική η επιλογή του αριθμού *d*, προκειμένου οι υπολογισμοί να είναι αριθμητικά ακριβείς. Για περισσότερες πληροφορίες σχετικά με την επιλογή του αριθμού *d*, μπορεί να ανατρέξει κάποιος στα βιβλία Gould, Pitblado και Poi (2010), Κεφάλαιο 1 και Greene (2017) στο Παράρτημα Ε.

2.2.2 Υπολογισμός Οριακών Επιδράσεων διακριτών επεξηγηματικών μεταβλητών

Ο υπολογισμός των ΜΕ για τις διακριτές επεξηγηματικές μεταβλητές πραγματοποιείται με βάση τη μέθοδο πεπερασμένων διαφορών. Η μέθοδος πεπερασμένων διαφορών, χρησιμοποιείται για την επίλυση διαφορικών εξισώσεων προσεγγίζοντάς τες με εξισώσεις διαφοράς που οι πεπερασμένες διαφορές προσεγγίζουν τις παραγώγους. Στην πράξη η μέθοδος πεπερασμένων διαφορών, υπολογίζει το ΜΕ της διακριτής επεξηγηματικής μεταβλητής x_j, λαμβάνοντας την διαφορά ανάμεσα στην μεταβλητή απόκρισης όταν η διακριτή επεξηγηματική μεταβλητή x_j μεταβάλλεται "μοναδιαία", με την αρχική μεταβλητή απόκρισης όπου δεν έχει πραγματοποιηθεί καμία αλλαγή (Cameron & Trivedi, 2005). Δηλαδή:

$$\frac{\Delta E(Y|\boldsymbol{x})}{\Delta x_j} = E\left(Y|\boldsymbol{x}, x_j = end\right) - E\left(Y|\boldsymbol{x}, x_j = start\right) \qquad (2.4)$$

Με αυτό τον τρόπο για τις διάφορες κατηγορίες της x_j, υπολογίζονται και στην συνέχεια συγκρίνονται μεταξύ τους οι τιμές που λαμβάνει η μεταβλητή απόκρισης, δεδομένου ό,τι οι υπόλοιπες επεξηγηματικές μεταβλητές του μοντέλου $(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n)$ καθορίζονται σε σταθερές τιμές.

Για παράδειγμα για μία δυαδική επεξηγηματική μεταβλητή (x_j = 0 ,1), το ΜΕ δίνεται από τον παρακάτω τύπο:

$$\mathrm{ME}_{x_j} = E\left(Y\middle|\boldsymbol{x}, x_j = 1\right) - E\left(Y\middle|\boldsymbol{x}, x_j = 0\right) \qquad (2.5)$$

$$= E\left(Y\middle|\, x_1, \dots, x_{j-1}, 1,\ x_{j+1}, \dots, x_p\right) - E\left(Y\middle|x_1, \dots, x_{j-1}, 0,\ x_{j+1}, \dots, x_p\right)$$

Στην περίπτωση όπου μία επεξηγηματική μεταβλητή x_j, λαμβάνει περισσότερες από δύο κατηγορίες, έστω *k*. Τότε στο μοντέλο προκύπτουν $k-1$ δίτιμες επεξηγηματικές μεταβλητές που η κάθε μία κωδικοποιείται με 0 και 1 (0: όχι και 1: ναι). Έπειτα για τον υπολογισμό των ΜΕ των καινούργιων δίτιμων μεταβλητών χρησιμοποιείται ο τύπος (2.5).

2.2.3 Σχόλια

Όταν μία συνεχής επεξηγηματική μεταβλητή x_j λαμβάνει "μικρές" τιμές, (για παράδειγμα το ημερήσιο εισόδημα σε ευρώ), η επίδραση της αύξησης μίας μονάδας (ενός ευρώ) της x_j στη μεταβλητή απόκρισης, είναι πιθανό να ταιριάζει ορθά με το ΜΕ της x_j. Όταν όμως η x_j λαμβάνει "μεγαλύτερες" μονάδες (για παράδειγμα το ετήσιο εισόδημα σε ευρώ), το ΜΕ ενδέχεται να μην είναι μια πολύ καλή προσέγγιση της επίδρασης της x_j, όταν μεταβάλλεται κατά ένα ευρώ. Αυτός είναι και ένας λόγος για τον οποίο οι στιγμιαίες μεταβολές των συνεχών επεξηγηματικών μεταβλητών δεν λαμβάνουν ιδιαίτερη προσοχή στον τομέα της Κοινωνιολογίας. Οι πιο κοινές είναι οι προσεγγίσεις που επικεντρώνονται σε διακριτές αλλαγές (Williams, 2019).

Στα κλασσικά γραμμικά μοντέλα, είναι γνωστό ό,τι το ΜΕ μίας επεξηγηματικής μεταβλητής είναι σταθερή και ερμηνεύεται ως η μεταβολή της μεταβλητής απόκρισης, όταν η επεξηγηματική μεταβλητή ενδιαφέροντος μεταβληθεί "κατά μία μονάδα". Αυτή η ερμηνεία όμως δεν ισχύει και στην περίπτωση όπου δίνεται ένα πιο πολύπλοκο μοντέλο παλινδρόμησης.

Για παράδειγμα έστω ό,τι δίνεται το παρακάτω μοντέλο:

$$E(Y|x_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

Τώρα το ΜΕ δεν ισούται πλέον με τη μεταβολή της x_1 "κατά μία μονάδα". Είναι πολύ σημαντικό να τονισθεί αυτή η διαφορά διό,τι οι περισσότεροι ερμηνεύουν λανθασμένα τα ΜΕ όταν αναφέρονται σε αυτές. Για γίνει περισσότερο κατανοητό, παίρνουμε το παραπάνω μοντέλο και υπολογίζουμε την μεταβλητή απόκρισης στην μία περίπτωση, όταν η συνεχής μεταβλητή μένει ως έχει και στην δεύτερη περίπτωση, υπολογίζουμε την μεταβλητή απόκρισης όταν η x_1 μεταβάλλεται κατά μία μονάδα, δηλαδή για $x_1' = x_1 + 1$. Στη συνέχεια παίρνοντας την διαφορά προκύπτει το παρακάτω αποτέλεσμα.

$$E(Y|x_1') - E(Y|x_1) = \beta_0 + \beta_1(x_1 + 1) + \beta_2(x_1 + 1)^2 - (\beta_0 + \beta_1 x_1 + \beta_2 x_1^2)$$

$$\Leftrightarrow$$

$$E(Y|x_1') - E(Y|x_1) = \beta_1 + 2\beta_2 x_1 + \beta_2 \qquad (2.6)$$

Συνεπώς θα μπόρεσε να ισχυριστεί κάποιος ό,τι το ΜΕ της x_1, όταν μεταβληθεί κατά μία μονάδα, ισούται με $\beta_1 + 2\beta_2 x_1 + \beta_2$. Γεγονός που είναι λάθος, διό,τι, υπολογίζοντας στη συνέχεια το ΜΕ της x_1, με βάση τον τύπο (2.3), τότε προκύπτει ό,τι:

$$\mathrm{ME}_{x_1} = \lim_{d \to 0} \frac{E(Y|x_1 + d) - E(Y|x_1)}{d}$$

$$\Leftrightarrow$$

$$\mathrm{ME}_{x_1} = \lim_{d \to 0} \frac{\beta_0 + \beta_1(x_1 + d) + \beta_2(x_1 + d)^2 - (\beta_0 + \beta_1 x_1 + \beta_2 {x_1}^2)}{d}$$

$$\Leftrightarrow$$

$$\mathrm{ME}_{x_1} = \lim_{d \to 0} \frac{\beta_0 + \beta_1 x_1 + \beta_1 d + \beta_2 ({x_1}^2 + 2x_1 d + d^2) - (\beta_0 + \beta_1 x_1 + \beta_2 {x_1}^2)}{d}$$

$$\Leftrightarrow$$

$$ME_{x_1} = \beta_1 + 2\beta_2 x_1 \qquad (2.7)$$

Από τα παραπάνω συμπεράσματα, είναι προφανές ό,τι (2.7) ≠ (2.6), ως εκ τούτου η σωστή ερμηνεία του ΜΕ της x_1, είναι, ό,τι δηλώνει την μεταβολή της μεταβλητής απόκρισης όταν η x_1 μεταβληθεί κατά μια "αμελητέα ποσότητα" και όχι όταν μεταβάλλεται κατά μία μονάδα.

Κατά συνέπεια, για μη γραμμικά μοντέλα, θα πρέπει να δίνεται ιδιαίτερη προσοχή ό,τι τα ΜΕ για συνεχείς μεταβλητές, ισχύουν για μια "μικρή" αλλαγή της εκάστοτε επεξηγηματικής μεταβλητής.

Συνοψίζοντας, τα ΜΕ μπορούν να αποτελέσουν ένα σημαντικό πληροφοριακό μέσο όσο αφορά το πόσο επηρεάζει την μεταβλητή απόκρισης, μία πιθανή μεταβολή μίας επεξηγηματικής μεταβλητής. Για τις κατηγορικές μεταβλητές τα ΜΕ δείχνουν πόσο η μεταβλητή απόκρισης, αναμένεται να αλλάξει καθώς η x_j αλλάζει από 0 σε 1. Αυτό μπορεί να είναι πολύ χρήσιμο, ενημερωτικό και εύκολο να κατανοηθεί. Ενώ, για συνεχείς επεξηγηματικές μεταβλητές, το ΜΕ μετρά το στιγμιαίο ρυθμό αλλαγής. Εάν ο στιγμιαίος ρυθμός αλλαγής είναι παρόμοιος με την αλλαγή στην μεταβλητή απόκρισης καθώς η x_j αυξάνεται κατά ένα, αυτό μπορεί επίσης να είναι αρκετά χρήσιμο και διαισθητικό. Ωστόσο, αυτό εξαρτάται, εν μέρει, από την κλίμακα της x_j.

2.3 Ειδικές περιπτώσεις

Ορισμένα μοντέλα παλινδρόμησης περιέχουν συντελεστές που μπορούν να ερμηνευτούν άμεσα ως ΜΕ. Ωστόσο, δεν παρέχουν όλα τα μοντέλα τόσο απλές ερμηνείες. Οι εκτιμώμενοι συντελεστές σε πιο σύνθετα μοντέλα δεν προσδίδουν άμεσες ερμηνείες και οι πληροφορίες που παρέχουν μπορεί να οδηγήσουν τους ερευνητές σε μη ορθά στατιστικά συμπεράσματα. Στη παρούσα παράγραφο, διατυπώνονται επιγραμματικά οι διάφοροι τύποι μοντέλων παλινδρόμησης, με σκοπό τονισθεί ό,τι η ερμηνεία και ο υπολογισμός των ΜΕ, είναι απόλυτα συνδεδεμένη με τον τύπο μοντέλου παλινδρόμησης. Σημειώνεται ό,τι για τον εν συνεχεία υπολογισμό των ΜΕ, θεωρείται ό,τι οι επεξηγηματικές μεταβλητές διατηρούνται σε σταθερές τιμές. Οι κατηγορίες των μοντέλων παλινδρόμησης, είναι:

1) Αμιγώς γραμμικά μοντέλα
2) Μοντέλα με μετασχηματισμένες επεξηγηματικές μεταβλητές
3) Γενικευμένα Γραμμικά Μοντέλα

2.3.1 Αμιγώς γραμμικά μοντέλα

Δίνεται ένα κλασικό πολλαπλής γραμμικής παλινδρόμησης, έστω:

$$E(Y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + (\text{όροι ανεξάρτητοι του } x_1)$$

Όπου

- $(\text{όροι ανεξάτητοι του } x_1)$: δηλώνει τον γραμμικό συνδυασμό των παραμέτρων $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_m)'$ εξαιρώντας τους όρους β_0 & β_1 και των υπολοίπων επεξηγηματικών μεταβλητών $\boldsymbol{x}$ εκτός της x_1

Το διάνυσμα των εκτιμώμενων συντελεστών $\boldsymbol{\beta}$ του ακόλουθου μοντέλου παλινδρόμησης, υπολογίζεται με τη μεθόδου ελαχίστων τετραγώνων (OLS).

Σε περίπτωση όπου η x_1 είναι συνεχής επεξηγηματική μεταβλητή, το ΜΕ της με βάση τον τύπο (2.2), υπολογίζεται ως εξής:

$$\text{ME}_{x_1} = \frac{\partial E(Y|\boldsymbol{x})}{\partial x_1} = \frac{\partial(\beta_0 + \beta_1 x_1 + (\text{όροι ανεξάρτητοι του } x_1))}{\partial x_1} = \beta_1$$

Η ερμηνεία του ΜΕ μίας συνεχούς επεξηγηματικής μεταβλητής, είναι άμεσα ερμηνεύσιμη, σταθερή και ισούται με τον αντίστοιχο εκτιμώμενο συντελεστή της εκάστοτε επεξηγηματικής μεταβλητής, η οποία εκφράζει ό,τι σε μία μοναδιαία μεταβολή της x_1, η μεταβλητή απόκρισης Y, αναμένεται να μεταβληθεί κατά σταθερή ποσότητα β_1.

Στην περίπτωση όπου η επεξηγηματική μεταβλητή x_1, είναι διακριτή με δύο κατηγορίες (0/1), τότε το ΜΕ της με βάση τον τύπο (2.5), προκύπτει:

$$\text{ME}_{x_1} = E(Y|\boldsymbol{x}, x_1 = 1) - E(Y|\boldsymbol{x}, x_1 = 0) = \beta_1$$

2.3.2 Μοντέλα με μετασχηματισμένες επεξηγηματικές μεταβλητές

Όπως διαπιστώθηκε προηγουμένως το ΜΕ μίας επεξηγηματικής μεταβλητής (ανεξαρτήτως από τη φύση της) είναι άμεσα ερμηνεύσιμη και σταθερή με τον αντίστοιχο εκτιμώμενο συντελεστή της. Ωστόσο, αυτό το αποτέλεσμα δεν ισχύει σε πιο περίπλοκα μοντέλα τα οποία αποτελούνται από μη γραμμικούς όρους, όπως από όρους αλληλεπιδράσεων, λογαριθμικούς όρους ή όρους ισχύος κ.ο.κ..

2.3.2.1 Μοντέλα με μεταβλητές υψηλότερου όρου

Έστω ό,τι δίνεται το παρακάτω μοντέλο:

$$E(Y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + (\text{όροι ανεξάρτητοι του } x_1)$$

Το ΜΕ τώρα της συνεχούς επεξηγηματικής μεταβλητής x_1, δεν είναι άμεσα ερμηνεύσιμο και ύστερα από τον υπολογισμό του με βάση τον τύπο (2.2) δεν είναι σταθερό ίσο με τον εκτιμώμενο συντελεστή β_1, αλλά εξαρτάται επί πλέον από τις ίδιες τις τιμές της.

$$\text{ME}_{x_1} = \frac{\partial(\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + (\text{όροι ανεξάρτητοι του } x_1))}{\partial x_1} = \beta_1 + 2\beta_2 x_1$$

2.3.2.2 Μοντέλα με όρους αλληλεπίδρασης

Έστω ό,τι δίνεται το παρακάτω μοντέλο:

$$E(Y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + (\text{όροι ανεξάρτητοι του } x_1 \,\&\, x_2)$$

Το ΜΕ της συνεχούς επεξηγηματικής μεταβλητής x_1, δεν είναι άμεσα ερμηνεύσιμο. Όπως φαίνεται και από το ΜΕ της x_1 δεν είναι σταθερό και ίσο με τον εκτιμώμενο συντελεστή β_1, αλλά εξαρτάται από τις τιμές των υπολοίπων επεξηγηματικών μεταβλητών.

$$\text{ME}_{x_1} = \frac{\partial(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + (\text{όροι ανεξάρτητοι του } x_1 \& x_2))}{\partial x_1} = \beta_1 + \beta_3 x_2$$

Όπου

- $(\text{όροι ανεξάτητοι του } x_1 \,\&\, x_2)$: δηλώνει τον γραμμικό συνδυασμό των παραμέτρων $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_m)'$ εξαιρώντας τους όρους β_0, β_1, β_2 & β_3 και των υπολοίπων επεξηγηματικών μεταβλητών $\boldsymbol{x}$ εκτός της $x_1 \& x_2$.

Το ίδιο συμπέρασμα προκύπτει και στην περίπτωση όπου η x_1 είναι διακριτή επεξηγηματική μεταβλητή.

2.3.3 Γενικευμένα Γραμμικά Μοντέλα

Τα γενικευμένα γραμμικά μοντέλα (GLM) είναι και εκείνα που θα δοθεί περισσότερη προσοχή στην παρούσα διπλωματική εργασία. Σε αυτού του είδους τα μοντέλα, οι εκτιμώμενοι συντελεστές δεν μπορούν να ερμηνευτούν άμεσα ως ΜΕ (όπως στα μοντέλα OLS), για αυτόν τον λόγο η χρήση των τύπων (2.2) & (2.5) για των υπολογισμό των ΜΕ, κρίνεται αναγκαία.

Με αυτόν τον τρόπο επιτυγχάνεται καλύτερη ερμηνεία, όσον αφορά τις σχέσεις μεταξύ των επεξηγηματικών μεταβλητών και της μεταβλητής απόκρισης. Η ιδιαιτερότητα-πολυπλοκότητα αυτών των μοντέλων, όπως θα δειχθεί και στην συνέχεια, είναι ό,τι το ΜΕ μιας επεξηγηματικής μεταβλητής, εξαρτάται τόσο από την συνάρτηση σύνδεσης του συγκεκριμένου GLM, όσο και από τις υπόλοιπες επεξηγηματικές μεταβλητές του.

Σε αυτή την ενότητα, διατυπώνεται ο γενικός τύπος υπολογισμού του ΜΕ για τα γενικευμένα γραμμικά μοντέλα. Σημειώνεται, ό,τι για τον υπολογισμό των ΜΕ, σταθεροποιούνται οι υπόλοιπες επεξηγηματικές μεταβλητές σε συγκεκριμένες τιμές.

Έστω ό,τι δίνεται ένα GLM:

$$g\big(E(Y|\boldsymbol{x})\big) = \boldsymbol{x}'\boldsymbol{\beta} \qquad (2.8)$$

Όπου, η συνάρτηση σύνδεσης $g(\cdot)$ μετασχηματίζει την αναμενόμενη μεταβλητή απόκριση σε μία γραμμική εξίσωση μεταξύ των παραμέτρων $\boldsymbol{\beta} = \big(\beta_0, \beta_1, \beta_2, \dots, \beta_p\big)'$ και των επεξηγηματικών μεταβλητών $\boldsymbol{x} = \big(1, x_1, x_2, \dots, x_p\big)'$.

Η συνάρτηση σύνδεσης είναι αντίστροφή, συνεπώς ο παραπάνω τύπος γράφεται ως εξής:

$$E(Y|\boldsymbol{x}) = g^{-1}(\boldsymbol{x}'\boldsymbol{\beta})$$

Θέτοντας όπου $g^{-1}(\cdot) = h(\cdot)$, προκύπτει:

$$E(Y|\boldsymbol{x}) = h(\boldsymbol{x}'\boldsymbol{\beta}) \qquad (2.9)$$

Επομένως, το GLM (2.9), θεωρείται ως το μοντέλο γραμμικής παλινδρόμησης που βρίσκεται μέσα σε έναν μη γραμμικό μετασχηματισμό της μεταβλητής απόκρισης. Η επιλογή της συνάρτησης σύνδεσης $g(\cdot)$, (όπως είναι γνωστό από το 1ο κεφάλαιο) εξαρτάται από την κατανομή της μεταβλητής απόκρισης Y. Όπως προαναφέρθηκε, επειδή ένα GLM συνήθως υποδηλώνει το γραμμικό μοντέλο μέσα σε μια μη γραμμική συνάρτηση, δεν είναι εφικτό να παρθούν άμεσα τα ΜΕ από τους εκτιμώμενους συντελεστές.

Ενναλακτικά, με βάση την εξίσωση (2.2), ο γενικός τύπος υπολογισμού του ΜΕ μίας συνεχούς επεξηγηματικής μεταβλητής x_j, σε ένα GLM, είναι:

$$\mathrm{ME}_{x_j} = \frac{\partial E(Y|\boldsymbol{x})}{\partial x_j} = \frac{\partial h(\boldsymbol{x}'\boldsymbol{\beta})}{\partial x_j}\,\beta_j = h'(\boldsymbol{x}'\boldsymbol{\beta})\beta_j \qquad (2.10)$$

όπου

- $h'(\cdot)$: η πρώτη παράγωγος της συνάρτησης $h(\cdot)$

Όπως, φαίνεται από τον παραπάνω τύπο, προκύπτει ό,τι λόγω της μη γραμμικότητας της συνάρτησης σύνδεσης $h(\cdot)$, το ΜΕ της συνεχής επεξηγηματικής μεταβλητής x_j, εξαρτάται από την μερική παράγωγο της αντίστροφης της συνάρτησης σύνδεσης, μέσα στην οποία

εμπεριέχονται όλες οι υπόλοιπες επεξηγηματικές μεταβλητές καθώς και οι εκτιμώμενοι συντελεστές τους.

Είναι σημαντικό να αναφερθεί ό,τι ο παραπάνω τύπος (2.10), δεν μπορεί να εφαρμοστεί σε περίπτωση που η επεξηγηματική μεταβλητή x_j, είναι διωνυμική (για παράδειγμα η μεταβλητή φύλο). Αυτό οφείλεται στο γεγονός ό,τι υπολογίζει το ΜΕ για μία απειροστά μικρή μεταβολή της x_j και όχι για μία μεταβολή της τάξεως μίας μονάδας (από το 0 στο 1) (Fernihough, 1989). Για αυτό τον λόγο με βάση των τύπο (2.5) το ΜΕ, για μία δυαδική μεταβλητή x_j ενός GLM, ισούται:

$$\mathrm{ME}_{x_j} = \frac{\Delta E(Y|\boldsymbol{x})}{\Delta x_j} = E\big(Y\big|\boldsymbol{x}, x_j = 1\big) - E\big(Y\big|\boldsymbol{x}, x_j = 0\big) \Leftrightarrow$$

$$\mathrm{ME}_{x_j} = h\big(\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)} + \beta_j\big) - h\big(\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)}\big) \qquad (2.11)$$

όπου

- $\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)}$: δηλώνει τον γραμμικό συνδυασμό των παραμέτρων $\boldsymbol{\beta}$ (εξαιρώντας την παράμετρο β_j) και των υπολοίπων επεξηγηματικών μεταβλητών $\boldsymbol{x}$ εκτός της x_j.

2.4 Οριακές Επιδράσεις για τα γενικευμένα γραμμικά μοντέλα logit, probit & complementary log-log

Τα γενικευμένα γραμμικά μοντέλα logit, probit & complementary log-log, είναι μοντέλα των οποίων η μεταβλητή απόκρισης είναι κατηγορική με δύο κατηγορίες, ως εκ τούτου είναι μη γραμμικά. Τέτοιου είδους μοντέλα εκτιμώνται χρησιμοποιώντας τη μέθοδο μέγιστης πιθανότητας. Η κατανόηση των επιπτώσεων της μη γραμμικότητας είναι θεμελιώδους σημασίας για την ορθή ερμηνεία αυτών των μοντέλων.

2.4.1 Οριακές Επιδράσεις για μοντέλα logit

Δίνεται το μοντέλο logit:

$$\pi = \Lambda(\boldsymbol{x}'\boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}'\boldsymbol{\beta}}}$$

Το ΜΕ μίας συνεχούς επεξηγηματικής μεταβλητής έστω x_j, με βάση των τύπο (2.10), υπολογίζεται ως εξής:

$$\mathrm{ME}_{x_j} = \frac{\partial \pi}{\partial\, x_j} = \Lambda(\boldsymbol{x}'\boldsymbol{\beta})\big(1 - \Lambda(\boldsymbol{x}'\boldsymbol{\beta})\big)\beta_j = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{(1 + e^{\boldsymbol{x}'\boldsymbol{\beta}})^2}\beta_j \qquad (2.12)$$

Ενώ το ΜΕ της x_j, η οποία είναι δίτιμη μεταβλητή, με βάση τον τύπο (2.11), ισούται με:

$$\mathrm{ME}_{x_j} = \frac{\Delta \pi}{\Delta x_j} = \Lambda\big(\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)} + \beta_j\big) - \Lambda\big(\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)}\big) \Leftrightarrow$$

$$\mathrm{ME}_{x_j} = \frac{e^{\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)} + \beta_j}}{1 + e^{\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)} + \beta_j}} - \frac{e^{\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)}}}{1 + e^{\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)}}} \qquad (2.13)$$

2.4.2 Οριακές Επιδράσεις για μοντέλα probit

Έστω ό,τι δίνεται ένα μοντέλο probit:

$$\pi = \Phi(\boldsymbol{x}'\boldsymbol{\beta}) = \int_{-\infty}^{\boldsymbol{x}'\boldsymbol{\beta}} \varphi(z)dz$$

Το ΜΕ μίας συνεχούς επεξηγηματικής μεταβλητής έστω x_j, με βάση την εξίσωση (2.10), υπολογίζεται ως εξής:

$$\mathrm{ME}_{x_j} = \frac{\partial \pi}{\partial x_j} = \varphi(\boldsymbol{x}'\boldsymbol{\beta})\beta_j \qquad (2.14)$$

Διαφορετικά όταν η j-οστή επεξηγηματική μεταβλητή ενός μοντέλου probit, είναι δίτιμη (0/1), ακολουθώντας τον τύπο (2.11), το ΜΕ είναι ίσο με:

$$\mathrm{ME}_{x_j} = \frac{\Delta \pi}{\Delta x_j} = \Phi\big(\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)} + \beta_j\big) - \Phi\big(\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)}\big) \qquad (2.15)$$

2.4.3 Οριακές Επιδράσεις για μοντέλα complementary log-log

Έστω ό,τι δίνεται ένα μοντέλο complementary log-log:

$$\pi = W(\boldsymbol{x}'\boldsymbol{\beta})$$

Το ΜΕ μίας συνεχούς επεξηγηματικής μεταβλητής έστω x_j, με βάση την εξίσωση (2.10), υπολογίζεται ως εξής:

$$\mathrm{ME}_{x_j} = \frac{\partial \pi}{\partial x_j} = w(\boldsymbol{x}'\boldsymbol{\beta})\beta_j = e^{\boldsymbol{x}'\boldsymbol{\beta}}\big(1 - W(\boldsymbol{x}'\boldsymbol{\beta})\big)\beta_j = e^{\boldsymbol{x}'\boldsymbol{\beta} - e^{\boldsymbol{x}'\boldsymbol{\beta}}}\beta_j \qquad (2.16)$$

Διαφορετικά όταν η j-οστή επεξηγηματική μεταβλητή ενός μοντέλου complementary log-log, είναι δίτιμη (0/1), ακολουθώντας τον τύπο (2.11), το ΜΕ είναι ίσο με:

$$\mathrm{ME}_{x_j} = \frac{\Delta\pi}{\Delta x_j} = W\left(\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)} + \beta_j\right) - W\left(\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)}\right) \Leftrightarrow$$

$$\mathrm{ME}_{x_j} = \left(1 - e^{-e^{\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)} + \beta_j}}\right) - \left(1 - e^{-e^{\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)}}}\right) \Leftrightarrow$$

$$\mathrm{ME}_{x_j} = e^{-e^{\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)}}} - e^{-e^{\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)} + \beta_j}} \qquad (2.17)$$

Πίνακας 2.2: Διωνυμικά μοντέλα παλινδρόμησης: Τα πιο ευρέως κοινά χρησιμοποιούμενα

Model	**Probability**	**Marginal Effects for x_j continuous explanatory variable**	**Marginal Effects for x_j discrete explanatory variable**
Logit	$\Lambda(\boldsymbol{x}'\boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}'\boldsymbol{\beta}}}$	$\frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{(1 + e^{\boldsymbol{x}'\boldsymbol{\beta}})^2}\beta_j$	$\frac{e^{\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)} + \beta_j}}{1 + e^{\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)} + \beta_j}} - \frac{e^{\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)}}}{1 + e^{\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)}}}$
Probit	$\Phi(\boldsymbol{x}'\boldsymbol{\beta}) = \int_{-\infty}^{\boldsymbol{x}'\boldsymbol{\beta}} \varphi(z)dz$	$\varphi(\boldsymbol{x}'\boldsymbol{\beta})\beta_j$	$\Phi\left(\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)} + \beta_j\right) - \Phi\left(\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)}\right)$
Complementary log-log	$\mathrm{W}(\boldsymbol{x}'\boldsymbol{\beta}) = 1 - e^{-e^{\boldsymbol{x}'\boldsymbol{\beta}}}$	$e^{\boldsymbol{x}'\boldsymbol{\beta} - e^{\boldsymbol{x}'\boldsymbol{\beta}}}\beta_j$	$e^{-e^{\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)}}} - e^{-e^{\boldsymbol{x}'^{(-j)}\,\boldsymbol{\beta}^{(-j)} + \beta_j}}$
Linear Probability	$\boldsymbol{x}'\boldsymbol{\beta}$	β_j	β_j

2.5 Οριακές Επιδράσεις για τα γενικευμένα γραμμικά μοντέλα Poisson και ZIP

2.5.1 Οριακές Επιδράσεις στην παλινδρόμηση Poisson

Δίνεται το μοντέλο παλινδρόμησης Poisson:

$$\mu = e^{\boldsymbol{x'\beta}}$$

Το ΜΕ μίας συνεχούς επεξηγηματικής μεταβλητής έστω x_j, με βάση των τύπο (2.10), υπολογίζεται ως εξής:

$$\mathrm{ME}_{x_j} = \frac{\partial \mu}{\partial\, x_j} = \; e^{\boldsymbol{x'\beta}}\, \beta_j \qquad (2.18)$$

Ενώ το ΜΕ της x_j, η οποία είναι δίτιμη μεταβλητή, με βάση τον τύπο (2.11), ισούται με:

$$\mathrm{ME}_{x_j} = \frac{\Delta \mu}{\Delta x_j} = \; e^{\boldsymbol{x'^{(-j)}\beta^{(-j)}}+\beta_j} \; - \; e^{\boldsymbol{x'^{(-j)}\beta^{(-j)}}} \qquad (2.19)$$

2.5.2 Οριακές Επιδράσεις στην παλινδρόμηση ZIP

Δίνεται το μοντέλο παλινδρόμησης ZIP, το οποίο αποτελείται από παρακάτω δύο συναρτήσεις:

$$\lambda = e^{\boldsymbol{x'\beta}} \qquad \& \qquad \pi = \frac{e^{\boldsymbol{z'\gamma}}}{1 + e^{\boldsymbol{z'\gamma}}}$$

Για το οποίο, η μέση τιμή όπως έχει δειχθεί ισούται:

$$\mu = (1 - \pi)\lambda = \frac{e^{\boldsymbol{x'\beta}}}{1 + e^{\boldsymbol{z'\gamma}}}$$

Θεωρώντας τώρα ό,τι οι επεξηγηματικές μεταβλητές $\boldsymbol{x}$ & $\boldsymbol{z}$ είναι οι ίδιες, δηλαδή $\boldsymbol{x} = \boldsymbol{z}$, η μέση τιμή θα ισούται:

$$\mu = \frac{e^{\boldsymbol{x'\beta}}}{1 + e^{\boldsymbol{x'\gamma}}}$$

Οπότε, το ΜΕ μίας συνεχούς επεξηγηματικής μεταβλητής έστω x_j, με βάση των τύπο (2.10), υπολογίζεται ως εξής:

$$\mathrm{ME}_{x_j} = \frac{\partial \mu}{\partial\, x_j} = \frac{e^{\boldsymbol{x'\beta}+\boldsymbol{x'\gamma}}\left(\beta_j - \gamma_j\right) + e^{\boldsymbol{x'\beta}}\,\beta_j}{\left(1 + e^{\boldsymbol{x'\gamma}}\right)^2} \qquad (2.20)$$

Ενώ το ΜΕ της x_j, η οποία είναι δίτιμη μεταβλητή, με βάση τον τύπο (2.11), ισούται με:

$$\mathrm{ME}_{x_j} = \frac{\Delta \mu}{\Delta x_j} = \frac{e^{\boldsymbol{x'^{(-j)}\beta^{(-j)}}+\beta_j}}{1 + e^{\boldsymbol{x'^{(-j)}\gamma^{(-j)}}+\gamma_j}} - \frac{e^{\boldsymbol{x'^{(-j)}\beta^{(-j)}}}}{1 + e^{\boldsymbol{x'^{(-j)}\gamma^{(-j)}}}} \qquad (2.21)$$

2.6 Διαφορές των Odds Ratios, Relative Risks και Marginal Effects ως προς την ερμηνεία των συντελεστών ενός μοντέλου Logit

Οι ερευνητές, έρχονται συχνά σε δίλημμα, όσον αφορά με ποιο τρόπο να ερμηνεύσουν τους εκτιμώμενους συντελεστές ενός μοντέλου με δυαδική {0/1} μεταβλητή απόκρισης (logit, probit). (Norton & Dowd, 2018). Αν και η χρήση αυτών των μοντέλων μπορεί να είναι κατάλληλη σε πολλές εφαρμογές, η δυσκολία έγκειται στην ερμηνεία. Στην πραγματικότητα, οι εκτιμώμενοι συντελεστές σε αυτά τα μοντέλα έχουν μικρή χρησιμότητα, εκτός από το να δείξουν εάν μια επεξηγηματική μεταβλητή έχει θετικό ή αρνητικό αποτέλεσμα. Για αυτόν τον λόγο οι ερευνητές, έχουν διάφορα στατιστικά εργαλεία πέρα από τα ΜΕ, με σκοπό να ερμηνεύσουν τους εκτιμώμενους συντελεστές του μοντέλου. Σημειώνεται ό,τι προτιμάται η εκτίμηση μοντέλων logit και όχι μοντέλων probit, εξαιτίας της ερμηνείας του λόγου των πιθανοτήτων (*OR*) ή του σχετικού κινδύνου *(RR*) χρησιμοποιώντας τους εκτιμώμενους συντελεστές ενός μοντέλου logit. Στην συνέχεια γίνεται μία σύντομη ανασκόπηση ως προς αυτές τις ποσότητες. Έπειτα αναφέρονται τα μειονεκτήματα που παρουσιάζουν σε σχέση με τα ΜΕ.

Odds Ratios - OR

$$\mathrm{OR} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{odds_1}{odds_2}, \qquad \mathrm{OR}\,\epsilon\,(0, +\infty)$$

Η τιμή των OR σε ένα μοντέλο logit, είναι:

$$\mathrm{OR} = e^{\beta}$$

Relative Risk - RR

$$\mathrm{RR} = \frac{p_1}{p_2}, \qquad \mathrm{RR}\,\epsilon\,(0, +\infty)$$

Οι επιδράσεις σε ένα μοντέλο logit, παρουσιάζονται συχνά σε σχετικούς όρους, ως αναλογίες πιθανοτήτων ή σχετικού κινδύνου. Το προσδιοριστικό σημείο αυτών των ποσοτήτων είναι το 1 ενώ των ΜΕ είναι 0 στην περίπτωση γραμμικής διαφοράς. Επί πλέον, οι ποσότητες OR και RR παίρνουν πάντα θετικές τιμές, ενώ τα ΜΕ εκτείνονται σε ολόκληρη την πραγματική ευθεία η οποία περιλαμβάνει θετικές και αρνητικές τιμές. Οι OR και RR, ποσοτικοποιούν τις σχετικές διαφορές μεταξύ των καθορισμένων ομάδων μιας έρευνας, ενώ το ΜΕ ποσοτικοποιεί τη σταδιακή διαφορά στα αποτελέσματα μεταξύ καθορισμένων ομάδων. Ταυτόχρονα, οι ποσότητες OR και RR, δεν εκφράζονται σε μονάδες μέτρησης και δεν αντιπροσωπεύουν το μεγέθους του δείγματος (Onukwugha, Bergtold, & Jain, 2015).

Τόσο ο λόγος πιθανοτήτων, όσο και ο σχετικός κίνδυνος, είναι ποσότητες οι οποίες μπορούν να υπολογιστούν εξ ίσου εύκολα και για συνεχείς και για διακριτές επεξηγηματικές μεταβλητές, Παρ' όλα αυτά, έχουν πολλά μειονεκτήματα. Αυτό θα γίνει πιο κατανοητό με το επόμενο παράδειγμα. Σημειώνεται, ό,τι στο επόμενο παράδειγμα για λόγους απλότητας δεν χρησιμοποιούνται επεξηγηματικές μεταβλητές ενός μοντέλου παλινδρόμησης και ό,τι οι παρακάτω πιθανότητες δεν αποτελούν πραγματικά δεδομένα. Τα ακόλουθα συμπεράσματα γενικεύονται και σε πιο πολύπλοκες περιπτώσεις.

Παράδειγμα

Έστω ό,τι η πιθανότητα να κολλήσει covid-19 ένας ενήλικας είναι 70%, ενώ η πιθανότητα να κολλήσει covid-19 ένας ενήλικας ο οποίος έχει εμβολιστεί είναι 20% (70%-20% = 50% απόκλιση). Τότε, ο λόγος των πιθανοτήτων των δυο ενδεχομένων, ισούται περίπου με 9.4, η ερμηνεία του οποίου σημαίνει ό,τι η σχετική πιθανότητα να κολλήσει ένας ενήλικας covid-19, ο οποίος δεν έχει εμβολιαστεί είναι περίπου 9.4 φορές μεγαλύτερη από εκείνον που έχει εμβολιαστεί. Θα μπορούσε να ισχυριστεί κανείς ό,τι ο στόχος επιτευχθεί, από την σκοπιά της διεξαγωγής στατιστικών συμπερασμάτων. Ωστόσο προκύπτουν ορισμένα ζητήματα.

Πρώτον στην τιμή του OR δεν μεταδίδεται το μέγεθος του δείγματος που ερευνάται. Δεύτερον, είναι πιθανό, πολλά ζεύγη πιθανοτήτων να δίνουν την ίδια ακριβώς τιμή OR. Για παράδειγμα, αν η πιθανότητα να κολλήσει covid-19 ένας ενήλικας ήταν 99% και η πιθανότητας να κολλήσει covid-19 ένας ενήλικας ο οποίος έχει εμβολιστεί ήταν 91.3%, πάλι ο λόγος πιθανοτήτων θα ισούται με 9.4, παρ' όλο που η απόκλιση στα ποσοστά είναι μόνο 7.7%, σε σύγκριση με το 50%. Η απόκλιση της τάξης του 50% στην πρώτη περίπτωση, δείχνει ό,τι το εμβόλιο έχει επίδραση ως προς το αν νοσήσει ένα ενήλικας, από την άλλη πλευρά, η απόκλιση 7.7% στην δεύτερη περίπτωση, σημαίνει ό,τι το εμβόλιο δεν έχει καμία σημαντική επίδραση ως προς το αν νοσήσει ή όχι ένας ενήλικας. Ωστόσο, και στις δύο περιπτώσεις, ερμηνεύεται ό,τι η σχετική πιθανότητα να κολλήσει ένας ενήλικας covid-19, ο οποίος δεν έχει εμβολιαστεί είναι περίπου 9.4 φορές μεγαλύτερη από εκείνον που έχει εμβολιαστεί. Τέλος, ο λόγος πιθανοτήτων λαμβάνει πολύ μεγάλες τιμές σε περιπτώσεις που είναι σπάνιες/ ακραίες. Για παράδειγμα έστω ό,τι οι παραπάνω πιθανότητες ήταν 2% και 0.29% αντίστοιχα, τότε αποδίδουν λόγο πιθανοτήτων 9.4, αρκετά μεγάλη τιμή, παρόλο που επηρεάζονται μόνο 2 στους 100 ενήλικες από το εμβόλιο.

Τα ίδια συμπεράσματα, εμφανίζονται και όταν υπολογίζεται ο σχετικός κίνδυνος. Δηλαδή, εάν η πιθανότητα να κολλήσει covid-19 κάποιος που δεν έχει εμβολιαστεί είναι 40%, ενώ εκείνος που έχει εμβολιαστεί είναι 20%, τότε ο σχετικός κίνδυνος ισούται με 2, με την ερμηνεία ό,τι η πιθανότητα να κολλήσει κάποιος που δεν έχει κάνει εμβόλιο είναι διπλάσια από εκείνον που έχει κάνει. Ωστόσο η τιμή του σχετικού κινδύνου θα ήταν ακριβώς η ίδια αν οι παραπάνω πιθανότητες ήταν 4% και 2% αντίστοιχα. Συνεπώς θα μπορούσε να πει κανείς ό,τι το εμβόλιο μειώνει την πιθανότητα να νοσήσει κάποιος κατά το μισό, το οποίο συμπέρασμα είναι παραπλανητικό, διό,τι έχει τεράστια διαφορά η μείωση της τάξης του 2% (4%-2%), από την προηγούμενη της τάξης του 20% (40%-20%).

Σχόλια

Όπως είναι φανερό τόσο ο λόγος πιθανοτήτων, όσο και ο σχετικός κίνδυνος, οδηγούν σε παραπλανητικά συμπεράσματα, γεγονός που τους καθιστά έναν μη αξιόπιστο τρόπο ερμηνείας των αποτελεσμάτων. Ιδανικά, ο ερευνητής θα ήθελε να ερμηνεύσει ένα δυαδικό μοντέλο (0/1) (logit, probit), στην κλίμακα πιθανότητας (0,1) και όχι στην κλίμακα αναλογιών (LO, OR & RR). Άλλα όπως έχει αναφερθεί στην κλίμακα πιθανότητας, όλες επιδράσεις είναι μη γραμμικές. Λύση στο παραπάνω πρόβλημα έρχονται να δώσουν οι Οριακές Επιδράσεις. Στο μοντέλο logit είναι ευρέως χρησιμοποιούμενα τα ΜΕ, διό,τι ο ερευνητής επιθυμεί να μελετήσει και να ερμηνεύσει τις επιδράσεις στην κλίμακα πιθανότητας, παρ' όλο που οι συντελεστές του μοντέλου εκτιμώνται στην κλίμακα log odds. Ωστόσο, τα ΜΕ εφαρμόζονται σε κάθε είδους γενικευμένο γραμμικό μοντέλο. Εν κατακλείδι, τα μοντέλα παλινδρόμησης εκτιμώνται με σκοπό να παραχθούν προβλέψεις ώστε με την χρήση των ΜΕ να επιτευχθεί όσο τον δυνατόν καλύτερη ερμηνεία του μοντέλου στην κλίμακα που έχει περισσότερο νόημα (Coca, 2019).

Κεφάλαιο 3

3.1 Προσεγγίσεις

Εκτός από τα μοντέλα γραμμικής παλινδρόμησης ίσως το πιο ευρέως χρησιμοποιούμενο εργαλείο ενός ερευνητή, είναι τα γενικευμένα γραμμικά μοντέλα (GLM). Ακόμη και στα πιο απλά γενικευμένα γραμμικά μοντέλα με πολλαπλές επεξηγηματικές μεταβλητές, ο επιστημονικός χώρος δεν παρέχει σαφή καθοδήγηση σχετικά με τον καλύτερο τρόπο υπολογισμού των ΜΕ, τα οποία είναι τα απαραίτητα εργαλεία για την διεξαγωγή ορθών στατιστικών συμπερασμάτων. Στόχος αυτού του κεφαλαίου είναι να παράσχει τέτοια σαφήνεια επισημαίνοντας μια προσέγγιση που παρέχει αποτελέσματα που συμβαδίζουν περισσότερο με τους στόχους της θεωρητικά κατευθυνόμενης εμπειρικής έρευνας.

Όπως διατυπώθηκε στο προηγούμενο κεφάλαιο, τα ΜΕ για μοντέλα παλινδρόμησης, τα οποία είναι μη γραμμικά, εξαρτώνται από όλες τις τιμές των επεξηγηματικών μεταβλητών που περιέχει το μοντέλο. Ως εκ τούτου, υπάρχουν τρεις προσεγγίσεις, συμφωνά με τις οποίες μπορούν να υπολογιστούν τα ΜΕ, από τις οποίες οι δύο τελευταίες είναι και οι προσεγγίσεις που χρησιμοποιούνται με μεγαλύτερη συχνότητα. Οι προσεγγίσεις αυτές, είναι:

1) Οριακές Επιδράσεις σε Αντιπροσωπευτικές Τιμές (Marginal Effects at Representative values, MER)
2) Οριακές Επιδράσεις στο Μέσο (Marginal Effects at Means, MEM)
3) Μέσες Οριακές Επιδράσεις (Average Marginal Effects, AME)

Για καλύτερη κατανόηση, ας θεωρήσουμε το μοντέλο παλινδρόμησης

$$E(Y|\boldsymbol{x}) = h(\boldsymbol{x}'\boldsymbol{\beta})$$

όπου

- Y: είναι η μεταβλητή απόκρισης
- $\boldsymbol{x} = (1, x_1, \dots, x_p)'$: το διάνυσμα των p επεξηγηματικών μεταβλητών, συμπεριλαμβανομένου του σταθερού όρου
- x_j: είναι η j-οστή επεξηγηματική μεταβλητή, για $j = 1, \dots, p$ επεξηγηματικές μεταβλητές
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$: είναι το διάνυσμα με τις άγνωστες παραμέτρους
- $h(\cdot)$: είναι μια παραγωγίσιμη συνάρτηση

3.1.1 Οριακές Επιδράσεις σε Αντιπροσωπευτικές Τιμές (MER)

Η προσέγγιση MER, υπολογίζει το ΜΕ μίας επεξηγηματικής μεταβλητής, επιλέγοντας συγκεκριμένες τιμές $\boldsymbol{x_0}$ για το διάνυσμα $\boldsymbol{x}$ των επεξηγηματικών μεταβλητών του μοντέλου (Leeper, 2018). Οι ερευνητές μπορούν να δημιουργήσουν αρκετές περιπτώσεις, αλλά συνήθως χρησιμοποιούν μόνο μία περίπτωση, αναλόγως με το τι επιθυμούν να μελετήσουν.

Στη συνέχεια, διατυπώνονται οι τύποι υπολογισμού των ΜΕ, με βάση την προσέγγιση MER, για διακριτές και συνεχείς επεξηγηματικές μεταβλητές.

Η MER για μία συνεχή επεξηγηματική μεταβλητή x_j, δίνεται από τύπο:

$$\text{MER}_{x_j} = \frac{\partial h(\boldsymbol{x_0'\beta})}{\partial x_j} = h'(\boldsymbol{x_0'\beta})\beta_j \qquad (3.1)$$

όπου

- $\boldsymbol{x_0}$: δηλώνει το διάνυσμα με τις συγκεκριμένες τιμές για το διάνυσμα $\boldsymbol{x}$ των επεξηγηματικών μεταβλητών
- $h'(\cdot)$: δηλώνει την πρώτη παράγωγο της συνάρτησης $h(\cdot)$

Η MER για μία δυαδική διακριτή επεξηγηματική μεταβλητή x_j, η οποία λαμβάνει τιμές 0 και 1, υπολογίζεται ως εξής:

$$\text{MER}_{x_j} = h\big(\boldsymbol{x_0'\beta}|x_j = 1\big) - h\big(\boldsymbol{x_0'\beta}|x_j = 0\big) \qquad (3.2)$$

3.1.2 Οριακές Επιδράσεις στο Μέσο (MEM)

Η προσέγγιση MEM μπορεί να θεωρηθεί ειδική περίπτωση της MER. Η διαφορά έγκειται στο γεγονός ό,τι η προσέγγιση MEM, υπολογίζει το ΜΕ μίας επεξηγηματικής μεταβλητής θεωρώντας ό,τι οι επεξηγηματικές μεταβλητές λαμβάνουν την μέση τους τιμή (Greene, 2012).

Στη συνέχεια, δίνονται οι τύποι υπολογισμού των ΜΕ, με βάση την προσέγγιση MEM, για διακριτές και συνεχείς επεξηγηματικές μεταβλητές.

Η MEM για μία συνεχή επεξηγηματική μεταβλητή x_j, υπολογίζεται ως εξής:

$$\text{MEM}_{x_j} = \frac{\partial h(\overline{\boldsymbol{x}}'\boldsymbol{\beta})}{\partial x_j} = h'(\overline{\boldsymbol{x}}'\boldsymbol{\beta})\beta_j \qquad (3.3)$$

όπου

- $\bar{\boldsymbol{x}}' = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_i}$: εκφράζει το διάνυσμα το οποίο περιέχει τις μέσες τιμές των επεξηγηματικών μεταβλητών.

Η MEM για μία δυαδική διακριτή επεξηγηματική μεταβλητή x_j, η οποία λαμβάνει τιμές 0 και 1, υπολογίζεται ως εξής:

$$\text{MEM}_{x_j} = h\big(\bar{\boldsymbol{x}}'\boldsymbol{\beta}|x_j = 1\big) - h\big(\bar{\boldsymbol{x}}'\boldsymbol{\beta}|x_j = 0\big) \qquad (3.4)$$

Σημειώνεται ό,τι για τις διακριτές επεξηγηματικές μεταβλητές, στους τύπους μπαίνει η παρατηρηθήσα μέση τιμή. Για παράδειγμα, έστω ό,τι δίνεται μία δίτιμη επεξηγηματική μεταβλητή (0/ 1), σε περίπτωση που η μέση της τιμής ήταν 0.4 τότε στον τύπο θα έμπαινε η τιμή 0.4.

3.1.3 Μέσες Οριακές Επιδράσεις (AME)

Για τον υπολογισμό της Μέσης Οριακής Επίδρασης (AME) μίας επεξηγηματικής μεταβλητής, αρχικά υπολογίζονται οι ΜΕ ξεχωριστά για κάθε παρατήρηση και στη συνέχεια λαμβάνεται ο μέσος όρος τους (Greene, 2012).

Η AME για μία συνεχή επεξηγηματική μεταβλητή x_j, υπολογίζεται ως εξής:

$$\text{AME}_{x_j} = \beta_j \frac{1}{n}\sum_{i=1}^{n} \frac{\partial h(\boldsymbol{x_i'\beta})}{\partial x_j} = \beta_j \frac{1}{n}\sum_{i=1}^{n} h'(\boldsymbol{x_i'\beta}) \qquad (3.5)$$

Η AME για μία δυαδική διακριτή επεξηγηματική μεταβλητή x_j, η οποία λαμβάνει τιμές 0 και 1, υπολογίζεται ως εξής:

$$\text{AME}_{x_j} = \frac{1}{n}\sum_{i=1}^{n}\big\{h\big(\boldsymbol{x_i'\beta}|x_j = 1\big) - h\big(\boldsymbol{x_i'\beta}|x_j = 0\big)\big\} \qquad (3.6)$$

3.1.4 Σχόλια

Η προσέγγιση MEM εκτιμάει το ΜΕ για μία επεξηγηματική μεταβλητή με μέσα χαρακτηριστικά ή μπορεί να υπολογιστεί σε άλλες συγκεκριμένες τιμές των επεξηγηματικών μεταβλητών, παρέχοντας εκτιμήσεις των ΜΕ για διαφορετικές αντιπροσωπευτικές παρατηρήσεις. Οι εκτιμήσεις των ΜΕ με βάση την προσέγγιση MEM, μπορεί να μη είναι πάντα ρεαλιστικές σε όλες τις επεξηγηματικές μεταβλητές που περιλαμβάνονται στο μοντέλο παλινδρόμησης, ή είναι πιθανό να μην υπάρχουν δεδομένα που να αντιπροσωπεύουν τις

συγκεκριμένες μέσες τιμές. Για παράδειγμα, όταν το μοντέλο αποτελείται από μία κατηγορική επεξηγηματική μεταβλητή, η οποία χωρίζεται σε δύο κατηγορίες (π.χ. 0: άντρας, 1: γυναίκα), τότε με βάση την προσέγγιση MEM, παίρνοντας τη μέση τιμή είναι πιθανό να μην είναι ακέραιος αριθμός. Ως εκ τούτου, δεν θα υπήρχε παρατήρηση που να αντιπροσωπεύει τη συγκεκριμένη μέση τιμή. Θεωρητικά, ο υπολογισμός της τιμής MEM δεν είναι επιθυμητός για μία δυαδική επεξηγηματική μεταβλητή επειδή, όπως σημειώνεται, "*κανείς στο σύνολο δεδομένων δεν θα είναι 60% γυναίκα ή 20% έγκυος*" (Dowd, Greene & Norton, 2014). Ως εναλλακτική λύση για τον υπολογισμό του ΜΕ στη μέση τιμή των επεξηγηματικών κατηγορικών μεταβλητών, οι ερευνητές μπορούν να ορίσουν την κατηγορική μεταβλητή στη τιμή που εμφανίζεται πιο συχνά.

Αντιθέτως η προσέγγιση AME υπολογίζει το μέσο ΜΕ για όλες τις παρατηρήσεις του δείγματος (Greene, 2012). Για αυτόν τον λόγο, ένα μεγάλο πλεονέκτημα της χρήσης του AME είναι η διαθεσιμότητα του ΜΕ που υπάρχει για κάθε παρατήρηση του δείγματος. Ουσιαστικά η προσέγγιση AME μπορεί να ερμηνευτεί ως το ΜΕ της αλλαγής της επεξηγηματικής μεταβλητής x_i από μια συγκεκριμένη τιμή κατά μέσο όρο σε όλο το δείγμα. Παρά την προτίμηση της προσέγγισης AME ως προεπιλεγμένη επιλογή, η σωστή επιλογή ως προς το εάν θα χρησιμοποιηθεί η προσέγγιση MEM ή AME θα πρέπει να καθοδηγείται από τις ανάγκες του εκάστοτε ερευνητή. Όμοια, ανάλογο συμπέρασμα προκύπτει και με το αν θα χρησιμοποιηθεί ολόκληρο το δείγμα ή ένα υποσύνολο για τους υπολογισμούς κάθε ΜΕ (Long & Freese, 2014 ˙ Long & Mustillo, 2018).

Η προσέγγιση AME είναι ιδιαίτερα χρήσιμη επειδή, σε αντίθεση με την MEM, παράγει μια ενιαία συνοπτική ποσότητα που αντικατοπτρίζει την πλήρη κατανομή των ΜΕ της επεξηγηματικής μεταβλητής x_i αντί μιας αυθαίρετης πρόβλεψης. Όπως η προσέγγιση MER, έτσι και η AME δίνει ενδεχομένως τη δυνατότητα μεταφοράς σημαντικών πληροφοριών σχετικά με την επίδραση κάθε μεταβλητής επί του αποτελέσματος.

Παρ' όλα αυτά, για μεγάλα δείγματα οι δύο αυτές προσεγγίσεις δίνουν περίπου τα ίδια αποτελέσματα. Αντιθέτως σε μικρά μεγέθους δείγματα, ενδέχεται οι δύο προσεγγίσεις να δώσουν διαφορετικά αποτελέσματα (Greene, 2012). Στην επόμενη ενότητα, παρουσιάζονται οι λόγοι, οι οποίοι μεγαλώνουν την διαφορά ανάμεσα στις δυο προσεγγίσεις (AME και MEM).

3.2 Οι προσεγγίσεις MEM & AME για τα Γενικευμένα Γραμμικά Μοντέλα με Δίτιμες Μεταβλητές Απόκρισης

3.2.1 Οι προσεγγίσεις MEM & AME για τα μοντέλα logit

Δίνεται το μοντέλο logit:

$$\pi = \Lambda(\boldsymbol{x}'\boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}'\boldsymbol{\beta}}}$$

Η προσέγγιση MEM για τα μοντέλα logit

Η MEM για μία συνεχή επεξηγηματική μεταβλητή x_j, ισούται με:

$$\text{MEM}_{x_j} = \Lambda(\bar{\boldsymbol{x}}'\boldsymbol{\beta})\big(1 - \Lambda(\bar{\boldsymbol{x}}'\boldsymbol{\beta})\big)\beta_j = \frac{e^{\bar{\boldsymbol{x}}'\boldsymbol{\beta}}}{(1 + e^{\bar{\boldsymbol{x}}'\boldsymbol{\beta}})^2}\beta_j \qquad (3.7)$$

ενώ, η MEM για μία δυαδική διακριτή επεξηγηματική μεταβλητή x_j, η οποία λαμβάνει τιμές 0 και 1, υπολογίζεται ως εξής:

$$\text{MEM}_{x_j} = \Lambda\big(\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)} + \beta_j\big) - \Lambda\big(\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)}\big) \Leftrightarrow$$

$$\text{MEM}_{x_j} = \frac{e^{\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)} + \beta_j}}{1 + e^{\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)} + \beta_j}} - \frac{e^{\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)}}}{1 + e^{\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)}}} \qquad (3.8)$$

όπου

- $\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)}$: είναι τα διανύσματα $\bar{\boldsymbol{x}}$ και $\boldsymbol{\beta}$ παραλείποντας το στοιχείο $\bar{x}_j$ και β_j, αντίστοιχα.

Η προσέγγιση AME για τα μοντέλα logit

Η AME για μία συνεχή επεξηγηματική μεταβλητή x_j, υπολογίζεται ως εξής:

$$\text{AME}_{x_j} = \beta_j \frac{1}{n}\sum_{i=1}^{n} \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})\big(1 - \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})\big) = \beta_j \frac{1}{n}\sum_{i=1}^{n} \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}})^2} \qquad (3.9)$$

Η AME για μία δυαδική διακριτή επεξηγηματική μεταβλητή x_j, η οποία λαμβάνει τιμές 0 και 1, υπολογίζεται ως εξής:

$$\text{AME}_{x_j} = \frac{1}{n}\sum_{i=1}^{n}\left\{\Lambda\left(\boldsymbol{x}_i'^{(-j)}\boldsymbol{\beta}^{(-j)} + \beta_j\right) - \Lambda\left(\boldsymbol{x}_i'^{(-j)}\boldsymbol{\beta}^{(-j)}\right)\right\} \quad \Leftrightarrow$$

$$AME_{x_j} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{e^{\boldsymbol{x}_i'^{(-j)}\boldsymbol{\beta}^{(-j)}+\beta_j}}{1+e^{\boldsymbol{x}_i'^{(-j)}\boldsymbol{\beta}^{(-j)}+\beta_j}} - \frac{e^{\boldsymbol{x}_i'^{(-j)}\boldsymbol{\beta}^{(-j)}}}{1+e^{\boldsymbol{x}_i'^{(-j)}\boldsymbol{\beta}^{(-j)}}}\right\} \quad (3.10)$$

όπου

- $\boldsymbol{x}_i'^{(-j)}\boldsymbol{\beta}^{(-j)}$: δηλώνει τον γραμμικό συνδυασμό των παραμέτρων $\boldsymbol{\beta}$ (εξαιρώντας την παράμετρο β_j) με το διάνυσμα με τις τιμές των p επεξηγηματικών μεταβλητών της i-οστής παρατήρησης εξαιρώντας την j-οστή επεξηγηματική μεταβλητή

3.2.2 Οι προσεγγίσεις MEM & AME για τα μοντέλα probit

Έστω ό,τι δίνεται ένα μοντέλο probit:

$$\pi = \Phi(\boldsymbol{x}'\boldsymbol{\beta}) = \int_{-\infty}^{\boldsymbol{x}'\boldsymbol{\beta}} \varphi(z)dz$$

Η προσέγγιση MEM για τα μοντέλα probit

Η MEM για μία συνεχή επεξηγηματική μεταβλητή x_j, ισούται με:

$$\text{MEM}_{x_j} = \varphi(\bar{\boldsymbol{x}}'\boldsymbol{\beta})\beta_j \quad (3.11)$$

Ενώ, η MEM για μία δυαδική διακριτή επεξηγηματική μεταβλητή x_j, η οποία λαμβάνει τιμές 0 και 1, υπολογίζεται ως εξής:

$$\text{MEM}_{x_j} = \Phi\left(\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)} + \beta_j\right) - \Phi\left(\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)}\right) \quad (3.12)$$

Η προσέγγιση AME για τα μοντέλα probit

Η AME για μία συνεχή επεξηγηματική μεταβλητή x_j, υπολογίζεται ως εξής:

$$\text{AME}_{x_j} = \beta_j \frac{1}{n}\sum_{i=1}^{n}\varphi(\boldsymbol{x}_i'\boldsymbol{\beta}) \quad (3.13)$$

Η *AME* για μία δυαδική διακριτή επεξηγηματική μεταβλητή x_j, η οποία λαμβάνει τιμές 0 και 1, υπολογίζεται ως εξής:

$$AME_{x_j} = \frac{1}{n}\sum_{i=1}^{n}\left\{\Phi\left(\boldsymbol{x}_i'^{(-j)}\boldsymbol{\beta}^{(-j)} + \beta_j\right) - \Phi\left(\boldsymbol{x}_i'^{(-j)}\boldsymbol{\beta}^{(-j)}\right)\right\} \qquad (3.14)$$

3.2.3 Οι προσεγγίσεις MEM & AME για τα γενικευμένα γραμμικά μοντέλα complementary log-log

Έστω ό,τι δίνεται ένα μοντέλο complementary log-log:

$$\pi = W(\boldsymbol{x}'\boldsymbol{\beta}) = 1 - e^{-e^{\boldsymbol{x}'\boldsymbol{\beta}}}$$

Η προσέγγιση MEM για τα μοντέλα complementary log-log

Η MEM για μία συνεχή επεξηγηματική μεταβλητή x_j, ισούται με:

$$\text{MEM}_{x_j} = w(\bar{\boldsymbol{x}}'\boldsymbol{\beta})\beta_j = e^{\bar{\boldsymbol{x}}'\boldsymbol{\beta} - e^{\bar{\boldsymbol{x}}'\boldsymbol{\beta}}}\beta_j \qquad (3.15)$$

Ενώ, η MEM για μία δυαδική διακριτή επεξηγηματική μεταβλητή x_j, η οποία λαμβάνει τιμές 0 και 1, υπολογίζεται ως εξής:

$$\text{MEM}_{x_j} = W\left(\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)} + \beta_j\right) - W\left(\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)}\right) \qquad \Leftrightarrow$$

$$\text{ME}_{x_j} = e^{-e^{\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)}}} - e^{-e^{\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)} + \beta_j}} \qquad (3.16)$$

Η προσέγγιση AME για τα μοντέλα complementary log-log

Η AME για μία συνεχή επεξηγηματική μεταβλητή x_j, υπολογίζεται ως εξής:

$$\text{AME}_{x_j} = \beta_j \frac{1}{n}\sum_{i=1}^{n} e^{\boldsymbol{x}_i'\boldsymbol{\beta} - e^{\boldsymbol{x}_i'\boldsymbol{\beta}}} \qquad (3.17)$$

Η AME για μία δυαδική διακριτή επεξηγηματική μεταβλητή x_j, η οποία λαμβάνει τιμές 0 και 1, υπολογίζεται ως εξής:

$$\mathrm{AME}_{x_j} = \frac{1}{n}\sum_{i=1}^{n}\left\{W\left(\boldsymbol{x}_i'^{(-j)}\boldsymbol{\beta}^{(-j)} + \beta_j\right) - W\left(\boldsymbol{x}_i'^{(-j)}\boldsymbol{\beta}^{(-j)}\right)\right\} \quad \Leftrightarrow$$

$$AME_{x_j} = \frac{1}{n}\sum_{i=1}^{n}\left\{e^{-e^{x_i'^{(-j)}\boldsymbol{\beta}^{(-j)}}} - e^{-e^{x_i'^{(-j)}\boldsymbol{\beta}^{(-j)}+\beta_j}}\right\} \quad (3.18)$$

3.3 Οι προσεγγίσεις MEM & AME για τα Γενικευμένα Γραμμικά Μοντέλα με Μεταβλητές Απόκρισης Απαριθμήσεων

3.3.1 Οι προσεγγίσεις MEM & AME για τα μοντέλα Poisson

Έστω ό,τι δίνεται ένα μοντέλο Poisson:

$$\mu = e^{\boldsymbol{x}'\boldsymbol{\beta}}$$

Η προσέγγιση MEM για τα μοντέλα Poisson

Η MEM για μία συνεχή επεξηγηματική μεταβλητή x_j, ισούται με:

$$\mathrm{MEM}_{x_j} = e^{\bar{\boldsymbol{x}}'\boldsymbol{\beta}}\,\beta_j \quad (3.19)$$

Ενώ, η MEM για μία δυαδική διακριτή επεξηγηματική μεταβλητή x_j, η οποία λαμβάνει τιμές 0 και 1, υπολογίζεται ως εξής:

$$\mathrm{MEM}_{x_j} = e^{\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)}+\beta_j} - e^{\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)}} \quad (3.20)$$

Η προσέγγιση AME για τα μοντέλα Poisson

Η AME για μία συνεχή επεξηγηματική μεταβλητή x_j, υπολογίζεται ως εξής:

$$\mathrm{AME}_{x_j} = \beta_j\,\frac{1}{n}\sum_{i=1}^{n} e^{x_i'\boldsymbol{\beta}} \quad (3.21)$$

Η AME για μία δυαδική διακριτή επεξηγηματική μεταβλητή x_j, η οποία λαμβάνει τιμές 0 και 1, υπολογίζεται ως εξής:

$$\text{AME}_{x_j} = \frac{1}{n}\sum_{i=1}^{n}\left\{e^{\boldsymbol{x}_i'^{(-j)}\boldsymbol{\beta}^{(-j)}+\beta_j} - e^{\boldsymbol{x}_i'^{(-j)}\boldsymbol{\beta}^{(-j)}}\right\} \qquad (3.22)$$

3.3.2 Οι προσεγγίσεις MEM & AME για τα μοντέλα ZIP

Έστω ό,τι δίνεται ένα μοντέλο ZIP:

$$\lambda = e^{\boldsymbol{x}'\boldsymbol{\beta}} \quad \& \quad \pi = \frac{e^{\boldsymbol{z}'\boldsymbol{\gamma}}}{1+e^{\boldsymbol{z}'\boldsymbol{\gamma}}}$$

Για το οποίο, η μέση τιμή όπως έχει δειχθεί ισούται:

$$\mu = (1-\pi)\lambda = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{1+e^{\boldsymbol{z}'\boldsymbol{\gamma}}}$$

Θεωρώντας τώρα ό,τι οι επεξηγηματικές μεταβλητές $\boldsymbol{x}$ & $\boldsymbol{z}$ είναι οι ίδιες, δηλαδή $\boldsymbol{x} = \boldsymbol{z}$, η μέση τιμή θα ισούται:

$$\mu = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}'\boldsymbol{\gamma}}}$$

Η προσέγγιση MEM για τα μοντέλα ZIP

Η MEM για μία συνεχή επεξηγηματική μεταβλητή x_j, ισούται με:

$$\text{MEM}_{x_j} = \frac{e^{\bar{\boldsymbol{x}}'\boldsymbol{\beta}+\bar{\boldsymbol{x}}'\boldsymbol{\gamma}}\left(\beta_j-\gamma_j\right)+e^{\bar{\boldsymbol{x}}'\boldsymbol{\beta}}\,\beta_j}{\left(1+e^{\bar{\boldsymbol{x}}'\boldsymbol{\gamma}}\right)^2} \qquad (3.23)$$

Ενώ το ΜΕ της x_j, η οποία είναι δίτιμη μεταβλητή, με βάση τον τύπο (2.11), ισούται με:

$$\text{MEM}_{x_j} = \frac{e^{\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)}+\beta_j}}{1+e^{\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\gamma}^{(-j)}+\gamma_j}} - \frac{e^{\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\beta}^{(-j)}}}{1+e^{\bar{\boldsymbol{x}}'^{(-j)}\boldsymbol{\gamma}^{(-j)}}} \qquad (3.24)$$

Η προσέγγιση AME για τα μοντέλα ZIP

Η AME για μία συνεχή επεξηγηματική μεταβλητή x_j, υπολογίζεται ως εξής:

$$\mathrm{AME}_{x_j} = \beta_j \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{e^{x_i'\boldsymbol{\beta}+x_i'\boldsymbol{\gamma}}\left(\beta_j-\gamma_j\right)+e^{x_i'\boldsymbol{\beta}}\beta_j}{\left(1+e^{x_i'\boldsymbol{\gamma}}\right)^2}\right\} \qquad (3.25)$$

Η AME για μία δυαδική διακριτή επεξηγηματική μεταβλητή x_j, η οποία λαμβάνει τιμές 0 και 1, υπολογίζεται ως εξής:

$$\mathrm{AME}_{x_j} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{e^{x_i'^{(-j)}\boldsymbol{\beta}^{(-j)}+\beta_j}}{1+e^{x_i'^{(-j)}\boldsymbol{\gamma}^{(-j)}+\gamma_j}} - \frac{e^{x_i'^{(-j)}\boldsymbol{\beta}^{(-j)}}}{1+e^{x_i'^{(-j)}\boldsymbol{\gamma}^{(-j)}}}\right\} \qquad (3.26)$$

3.4 Σύγκριση των προσεγγίσεων AME και MEM

Αν και υποστηρίχθηκε ό,τι η προσέγγιση MEM είναι μια καλή προσέγγιση της AME (Greene, 1997), παρόλα αυτά υπάρχουν ορισμένες ενδείξεις ό,τι οι προσεγγίσεις MEM και AME μπορεί να οδηγήσουν σε διαφορετικά αποτελέσματα, ακόμα και αν το δείγμα είναι πολύ μεγάλο (Bockarjova & Hazans, 2000). Στη συγκεκριμένη ενότητα, θα καταγραφούν οι συνθήκες, κάτω από τις οποίες, η προσέγγιση MEM δεν μπορεί να θεωρηθεί ως καλή εκτίμηση της προσέγγισης AME (Bartus, 2005).

Παίρνοντας τις εξισώσεις (3.5) και (3.3), για την περίπτωση όπου η επεξηγηματική μεταβλητή x_j είναι συνεχής, η διαφορά ανάμεσα στις προσεγγίσεις AME και MEM, γράφεται ως εξής:

$$\mathrm{AME}_{x_j} - \mathrm{MEM}_{x_j} = \beta_j \frac{1}{n}\sum_{i=1}^{n}\{h'(\boldsymbol{x_i'\beta}) - h'(\bar{\boldsymbol{x}}'\boldsymbol{\beta})\} \qquad (3.27)$$

Χρησιμοποιώντας τη σειρά Taylor 2ης τάξης γύρω από το σημείο $\bar{\boldsymbol{x}}'$, η ποσότητα $h'(\boldsymbol{x_i'\beta}) - h'(\bar{\boldsymbol{x}}'\boldsymbol{\beta})$, μπορεί να εκτιμηθεί ως εξής:

$$h'(\boldsymbol{x_i'\beta}) - h'(\bar{\boldsymbol{x}}'\boldsymbol{\beta}) \cong h''(\bar{\boldsymbol{x}}'\boldsymbol{\beta})(\boldsymbol{x_i'\beta} - \bar{\boldsymbol{x}}'\boldsymbol{\beta}) + \frac{1}{2}h'''(\bar{\boldsymbol{x}}'\boldsymbol{\beta})(\boldsymbol{x_i'\beta} - \bar{\boldsymbol{x}}'\boldsymbol{\beta})^2 \qquad (3.28)$$

Όπου

- $h''(\cdot)$ & $h'''(\cdot)$, είναι η δεύτερη και τρίτη παράγωγος αντίστοιχα της συνάρτησης $h(\cdot)$

Αντικαθιστώντας τις εξισώσεις (3.5) και (3.3), προκύπτει:

$$\mathrm{AME}_{x_j} - \mathrm{MEM}_{x_j} \cong \beta_j \frac{1}{n}\sum_{i=1}^{n}\left\{h''(\bar{\boldsymbol{x}}'\boldsymbol{\beta})(\boldsymbol{x_i'\beta} - \bar{\boldsymbol{x}}'\boldsymbol{\beta}) + \frac{1}{2}h'''(\bar{\boldsymbol{x}}'\boldsymbol{\beta})(\boldsymbol{x_i'\beta} - \bar{\boldsymbol{x}}'\boldsymbol{\beta})^2\right\} \qquad (3.29)$$

$$\Leftrightarrow$$

$$\text{AME}_{x_j} - \text{MEM}_{x_j} \cong \beta_j \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{2} h'''(\bar{\boldsymbol{x}}'\boldsymbol{\beta})(\boldsymbol{x_i}'\boldsymbol{\beta} - \bar{\boldsymbol{x}}'\boldsymbol{\beta})^2\right\}$$

<u>Παρατηρήσεις</u>

1) Από την εξίσωση (3.29), διαπιστώνεται ό,τι η ποσότητα $\sum_{i=1}^{n}(\boldsymbol{x_i}'\boldsymbol{\beta} - \bar{\boldsymbol{x}}'\boldsymbol{\beta}) = 0$, διό,τι, $\sum_{i=1}^{n}(\boldsymbol{x_i}' - \bar{\boldsymbol{x}}') = \sum_{i=1}^{n}\boldsymbol{x_i}' - n\bar{\boldsymbol{x}}' = n\bar{\boldsymbol{x}}' - n\bar{\boldsymbol{x}}' = 0$.

2) Επί πλέον, παρατηρούμε ό,τι η ποσότητα $\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x_i}'\boldsymbol{\beta} - \bar{\boldsymbol{x}}'\boldsymbol{\beta})^2$, δίνει την διασπορά του $\boldsymbol{x}'\boldsymbol{\beta}$, η οποία συμβολίζεται καταχρηστικά με $Var(\boldsymbol{x}'\boldsymbol{\beta})$, δηλαδή, $Var(\boldsymbol{x}'\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x_i}'\boldsymbol{\beta} - \bar{\boldsymbol{x}}'\boldsymbol{\beta})^2$.

Συνεπώς, η διαφορά μεταξύ των δύο προσεγγίσεων AME και MEM, δίνεται από τον ακόλουθο τύπο:

$$\text{AME}_{x_j} - \text{MEM}_{x_j} \cong \frac{1}{2}\beta_j h'''(\bar{\boldsymbol{x}}'\boldsymbol{\beta})\, Var(\boldsymbol{x}'\boldsymbol{\beta}) \qquad (3.30)$$

Παρόμοια αποτέλεσμα προκύπτουν και στην περίπτωση όπου η x_j, είναι διακριτή επεξηγηματική μεταβλητή.

3.5.1 Συμπεράσματα

Όπως διαπιστώθηκε και προηγουμένως, η διαφορά μεταξύ των δύο προσεγγίσεων δεν εξαρτάται από τον πρώτο όρο της σειράς Taylor (βλέπε παρατήρηση 1). Αυτό έχει ως συνέπεια το ΜΕ μιας επεξηγηματικής μεταβλητής, για τον υπολογισμό της οποίας ο ερευνητής δεν σταθεροποιεί τις υπόλοιπες επεξηγηματικές μεταβλητές στον δειγματικό τους μέσο, αλλά λαμβάνει ένα άλλο διάνυσμα τιμών, έστω $\boldsymbol{x_0}$, τότε η ποσότητα $\sum_{i=1}^{n}(\boldsymbol{x_i}'\boldsymbol{\beta} - \boldsymbol{x_0}\boldsymbol{\beta})$, να μην μηδενίζεται. Άρα η σχέση μεταξύ των δύο προσεγγίσεων γίνεται πιο πολύπλοκη καθώς η διαφορά τους θα εξαρτάται και από τον πρώτο όρο της σειράς Taylor. Επομένως, ο υπολογισμός του ΜΕ μιας επεξηγηματικής μεταβλητής, διατηρώντας τις υπόλοιπες επεξηγηματικές μεταβλητές στη μέση τους τιμή, είναι καλύτερη προσέγγιση της AME, σε σχέση με τις προσεγγίσεις που υπολογίζουν τα ΜΕ, διατηρώντας τις υπόλοιπες επεξηγηματικές μεταβλητές σε άλλες σταθερές τιμές.

Η διαφορά ανάμεσα στις δύο προσεγγίσεις, όπως διαπιστώνεται και στην εξίσωση (3.30), αυξάνεται όσο αυξάνεται η διασπορά της γραμμικής πρόβλεψης ($Var(\boldsymbol{x}'\boldsymbol{\beta})$). Η διασπορά της γραμμικής πρόβλεψης, είναι μεγάλη όταν οι εκτιμήσεις των παραμέτρων β,

λαμβάνουν μεγάλες τιμές. Συνεπώς, κρατώντας σταθερούς όλους τους παράγοντες της εξίσωσης (3.30), η διαφορά των AME και MEM, είναι μεγάλη όταν οι εκτιμήσεις των παραμέτρων παίρνουν μεγάλες τιμές και μικρή όταν οι εκτιμήσεις των παραμέτρων παίρνουν μικρές τιμές. Ως εκ τούτου, για μεγάλες τιμές των εκτιμημένων παραμέτρων η προσέγγιση MEM δεν μπορεί να θεωρηθεί καλή εκτίμηση της προσέγγισης AME, αντιθέτως όσο πιο μικρές τιμές λαμβάνουν οι εκτιμήσεις των παραμέτρων, τόσο πιο καλή εκτίμηση της AME μπορεί να θεωρηθεί η προσέγγιση MEM.

Η προσέγγιση MEM, είναι πιθανό να είναι μεγαλύτερη ή μικρότερη από την προσέγγιση AME, καθώς εξαρτάται μόνο από την ποσότητα $h'''(\cdot)$, δηλαδή από την τρίτη παράγωγο της συνάρτησης $h(\cdot)$. Για παράδειγμα, σε ένα μοντέλο probit είναι γνωστό ό,τι η ποσότητα $\Phi(x)$ δηλώνει την αθροιστική συνάρτηση της κανονικής κατανομής, ενώ η $\varphi(x)$ εκφράζει την συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής. Η δεύτερη παράγωγος της $\varphi(x)$, ισούται με $\varphi''(x) = (x^2 - 1)\varphi(x)$, η οποία μηδενίζεται καθώς το x λαμβάνει τιμές 1 ή -1, Επί πλέον ισχύει ό,τι, $\varphi''(x) > 0, \forall\, x \in (-\infty, -1) \cap (1, +\infty)$ και $\varphi''(x) < 0, \forall\, x \in (-1,1)$. Επιπροσθέτως από τον στατιστικό πίνακα της τυπικής κανονικής κατανομής είναι γνωστό ό,τι $\Phi(-1) = 0.1587$ και $\Phi(1) = 0.8413$. Ως εκ τούτου, όταν η προβλεπόμενη πιθανότητα π, του μοντέλου probit, η οποία εκτιμάται με βάση τις μέσες τιμές των επεξηγηματικών μεταβλητών, βρίσκεται ανάμεσα στον σύνολο τιμών (0.1587, 0.8413), η διαφορά των δύο προσεγγίσεων καταλήγει σε αρνητικό αριθμό, κατά συνέπεια ισχύει ό,τι $\text{AME}_{x_j} - \text{MEM}_{x_j} < 0 \Rightarrow \text{AME}_{x_j} < \text{MEM}_{x_j}$, δηλαδή η προσέγγιση MEM είναι μεγαλύτερη από την προσέγγιση AME. Ωστόσο εάν η προβλεπόμενη πιθανότητα π, λαμβάνει τιμές κάτω από 0.1587 ή πάνω από 0.8413, τότε ισχύει το αντίστροφο, δηλαδή η προσέγγιση MEM είναι μικρότερη από την προσέγγιση AME. Τέλος για τιμές της προβλεπόμενης πιθανότητας π, που είναι είτε κοντά στο 0.1587, είτε κοντά στο 0.8413, η διαφορά μεταξύ των δύο προσεγγίσεων θεωρείται αμελητέα. Στο ίδιο συμπέρασμα καταλήγουμε και στις περιπτώσεις των μοντέλων logit, complementary log-log, Poisson & ZIP.

3.5 Βιβλιογραφική Ανασκόπηση

Οι Hanushek & Jackson (1977) και οι Aldrich & Nelson (1984) μέσα από τα έργα τους, ξεχωρίζουν τα μοντέλα περιορισμένης εξαρτώμενης μεταβλητής για την πολιτική επιστήμη. Τα έργα αυτά αναφέρονται στις διαφορές που προκύπτουν μεταξύ γραμμικών και μη γραμμικών μοντέλων, τις οποίες επεξηγούν μέσω ενός εμπειρικού παραδείγματος που δείχνει τις προβλέψεις που προκύπτουν για ένα εύρος τιμών, που ενδέχεται να λάβουν οι υπόλοιπες επεξηγηματικές μεταβλητές. Οι Aldrich & Nelson (1984) προτείνουν ό,τι τα ΜΕ θα μπορούσαν να εκτιμηθούν επιλέγοντας διάφορες "ενδιαφέρουσες τιμές" για τις άλλες επεξηγηματικές μεταβλητές. Για εκείνη την εποχή, δεδομένου ό,τι υπήρχαν ελλιπείς γνώσεις για τα μοντέλα αυτά, οι μελέτες των παραπάνω αντιπροσωπεύουν σημαντικές προόδους. Δυστυχώς, όμως

σταματούν να συζητούν εναλλακτικές προσεγγίσεις με το πώς θα μπορούσε κάποιος να καθορίσει ποιες τιμές θεωρούνται ενδιαφέρουσες.

Ο King (1998), εκφράζει πόσο σημαντικό είναι το να αξιολογεί κάποιος την καμπύλη του διαγράμματος διασποράς, που προκύπτει από τα σημεία της μεταβλητής απόκρισης και να είναι προσεκτικός για να υποστηρίξει τη μελέτη του αποτελέσματος των προβλέψεων για διάφορους συνδυασμούς των επεξηγηματικών μεταβλητών. Μέσα από το έργο του, διατυπώνει ό,τι : "*Πιθανώς η πιο προφανής τιμή στη χρήση είναι και η πιο τετριμμένη, η μέση τιμή δηλαδή κάθε επεξηγηματικής μεταβλητής*".

Ο Long (1997 διαδραμάτισε σημαντικό ρόλο για τους εμπειρικούς πολιτικούς επιστήμονες. Η εργασία του δεν προσφέρει σαφείς συμβουλές για το ποιος είναι ο καλύτερος τρόπος προκειμένου να υπολογιστούν τα ΜΕ, παρόλο που αναφέρεται σε πληθώρα προσεγγίσεων. Παρ' όλα αυτά αναφέρει όσον αφορά τα ΜΕ, ό,τι ενώ η προσέγγιση MEM είναι μια δημοφιλής προσέγγιση, ο καθορισμός των άλλων επεξηγηματικών μεταβλητών στις μέσες τους τιμές είναι δυνητικά προβληματικός, καθώς όπως έχει αναφερθεί, μπορεί να μην υπάρχει η μέση περίπτωση μίας επεξηγηματικής μεταβλητής μέσα στον πληθυσμό. Επί πλέον τονίζει ό,τι οι προσεγγίσεις AME και MEM θα μπορούσαν να αποφέρουν διαφορετικά αποτελέσματα, ωστόσο δεν διατυπώνει μία σταθερή άποψη για το ποια είναι η καλύτερη προσέγγιση, λέγοντας μόνο ό,τι η *προσέγγιση AME* "*μπορεί να προτιμάται*". Δεδομένου όμως ό,τι τα παραδείγματα του βασίζονται σχεδόν αποκλειστικά στην προσέγγιση MEM, καθώς και το λογισμικό SPost (Long & Freese, 2005) χρησιμοποιεί τις μέσες τιμές των επεξηγηματικών μεταβλητών ως προεπιλογή και δεν παρέχουν εκτιμήσεις χρησιμοποιώντας την προσέγγιση AME, κρίνεται λογικό να συμπεράνει κανείς ό,τι προτιμάται η προσέγγιση MEM.

Παράλληλα, ενώ υπάρχουν μελέτες (Mccolley & Wolfinger, 1980), οι οποίες χρησιμοποιούν την προσέγγιση AME, η καθοδήγηση μέσω βιβλιογραφικής έρευνας, οδηγεί στο συμπέρασμα ό,τι η προσέγγιση MEM, κυριαρχεί κυρίως στον τομέα των πολιτικών επιστημών, γεγονός που διαπιστώνεται και από την ανάλυση περιεχομένου των άρθρων American Political Science Review, American Journal of Political Science και Journal of Politics, τα οποία ενισχύουν αυτή την πεποίθηση από το 2006 (Hanmer & Ozan Kalkan, 2013). Το 68% αυτών των άρθρων στα οποία χρησιμοποιούνται μοντέλα δυαδικής επιλογής, χρησιμοποιούν την προσέγγιση MEM και μόλις το 1% των άρθρων χρησιμοποιούν την προσέγγιση AME. Δυστυχώς για το 15% των σχετικών άρθρων, οι συγγραφείς δεν αναφέρουν ποια προσέγγιση χρησιμοποίησαν προκειμένου να παράγουν τα ΜΕ, ενώ τα υπόλοιπα άρθρα (15%) δεν αναφέρουν αποτελέσματα πέραν των συντελεστών. Η υπεροχή της προσέγγισης MEM ίσως επηρεάστηκε από την ευρεία διαθεσιμότητα λογισμικών τα οποία όρισαν ως προεπιλογή την προσέγγιση MEM και είναι εύκολα στην χρήση όπως διαπιστώνεται (Tomz, Wittenberg & King, 2001˙ Long & Freese, 2006).

Ωστόσο, συνεχίζεται να υπάρχουν ενδοιασμοί όσον αφορά ποια προσέγγιση θεωρείται "καλύτερη". Αν και πλέον υπάρχουν αρκετές μελέτες οι οποίες καλύπτουν το γεγονός ό,τι τα ΜΕ στα μη γραμμικά μοντέλα δεν είναι σταθερά και εξαρτώνται εν μέρει από τις τιμές για τις

άλλες επεξηγηματικές μεταβλητές, οι συμβουλές που παρέχονται σχετικά με τις τιμές που επιλέγονται είναι συχνά ελλιπείς ή ασυνεπείς.

3.6 Καταλληλότητα της προσέγγισης AME

Όπως διατυπώθηκε σε προηγούμενα κεφάλαια, στην περίπτωση γραμμικών μοντέλων παλινδρόμησης (χωρίς όρους αλληλεπίδρασης ή μετασχηματισμένες επεξηγηματικές μεταβλητές), με τη μέθοδο OLS, το ΜΕ της επεξηγηματικής μεταβλητής ενδιαφέροντος είναι σταθερό και δεν εξαρτάται από τις τιμές των υπολοίπων μεταβλητών, συνεπώς και οι δύο προσεγγίσεις καταλήγουν σε ίδια αποτελέσματα. Ωστόσο, για τα γενικευμένα γραμμικά μοντέλα (GLM) απαιτείται μεγαλύτερη προσοχή κατά την ερμηνεία των οριακών επιδράσεων. Η επιλογή μιας ενιαίας περίπτωσης ως σημείο αναφοράς (για παράδειγμα MEM) κατά την οποία θα υπολογιστούν τα ΜΕ της επεξηγηματικής μεταβλητής ενδιαφέροντος μπορεί να είναι προβληματική. Δηλαδή, θέτοντας συγκεκριμένες τιμές μόνο σε μία περίπτωση (για παράδειγμα στη μέση τιμή), υπάρχει κίνδυνος αφενός να προκύψει μια περίπτωση που δεν εμπεριέχεται στον πληθυσμό ή να εμφανίζεται σπάνια, αφετέρου μπορεί το ΜΕ για τη συγκεκριμένη περίπτωση να μην είναι αντιπροσωπευτικές για την συνολική εικόνα των ΜΕ. Συνεπώς, η ερμηνεία των ΜΕ με την προσέγγιση MEM παρέχει μια λιγότερο εμπεριστατωμένη εφαρμογή της θεωρίας και μια αναποτελεσματική χρήση των δεδομένων, μειώνοντας έτσι ενδεχομένως την ικανότητα να προκύψουν στατιστικά συμπεράσματα σχετικά με τον πληθυσμό ενδιαφέροντος.

Η θεωρία των πολιτικών επιστημών δεν έχει επιδιώξει να αναπτύξει θεωρίες σχετικά με το ποια προσέγγιση θεωρείται καλύτερη. Παρ' όλα αυτά αντί να προσπαθεί κάποιος να καταλάβει το αποτέλεσμα για τη μέση περίπτωση, ο στόχος είναι να ληφθεί μια εκτίμηση του μέσου αποτελέσματος στον πληθυσμό.

Για παράδειγμα, από τα στοιχεία των ερευνών από τις Ηνωμένες Πολιτείες, με την προσέγγιση *MEM* και θέτοντας τις υπόλοιπες επεξηγηματικές μεταβλητές στις μέσες τους τιμές, η περίπτωση συγκεκριμένα για τους ανθρώπους οι οποίοι είναι κατά μέσο όρο 48 χρονών, είναι ανεξάρτητα πολιτικά και έχουν εισόδημα κατά μέσο όρο μεταξύ 40.000-45.000, δεν εμπεριέχονται στο πληθυσμό. Επί πλέον σε ένα άλλο παράδειγμα με συγκεντρωτικά στοιχεία, δεν γνωρίζουμε τις θεωρίες σχετικά με χώρες με ποσοστό ανεργίας 9% και πληθυσμό 18 εκατομμυρίων ατόμων, το 17% των οποίων είναι άνω των 65 ετών κλπ. Με έναν τόσο περιορισμένο πληθυσμό, θα ήταν πιθανό να υπάρχουν περιγραφικά στατιστικά στοιχεία, αποφεύγοντας έτσι το ζήτημα του καλύτερου τρόπου απόκτησης αποτελεσμάτων από ένα μοντέλο περιορισμένης εξαρτώμενης μεταβλητής.

Ωστόσο, η προσέγγιση AME παρέχει εκτιμήσεις οι οποίες αναφέρονται άμεσα με τις ποσότητες ενδιαφέροντος στον πληθυσμό. Για παράδειγμα, αν κάποιος ενδιαφέρεται για το ΜΕ της εκλογικής μεταρρύθμισης στην προσέλευση στις εκλογές τις Αμερικής, η προσέγγιση της AME παρέχει το μέσο ΜΕ του δείγματος που μπορεί κανείς να συναγάγει από τον πληθυσμό.

Αντιθέτως, η προσέγγιση ΜΕΜ υπολογίζει το ΜΕ για μία μόνο περίπτωση, π.χ. 50 χρόνων, λευκές γυναίκες, με εισόδημα 40.000$ - 45.000$, κλπ., ως εκ τούτου αντιπροσωπεύουν μόνο το ΜΕ συγκεκριμένα στο ευρύτερο πληθυσμό. Έτσι, εκτός αν η υπόθεση σχετίζεται με το μικρό υποσύνολο που αποτελεί τη μέση περίπτωση, η μέση περίπτωση επαρκεί για την πλήρη αξιολόγηση της υπόθεσης ενδιαφέροντος σχετικά με την ουσιαστική και στατιστική σημασία της επίδρασης. Δηλαδή, το αποτέλεσμα για τη μέση περίπτωση μπορεί να μην είναι γενικευμένο στον ευρύτερο πληθυσμό, ειδικά εάν η μέση περίπτωση δεν υπάρχει στον πληθυσμό ή είναι σπάνια.

Για μελέτες που βασίζονται σε ένα νομοθετικό σώμα, σε χώρες και ούτω καθεξής, τότε οι μεμονωμένες περιπτώσεις μπορεί να αποτελούν ένα ειδικό ενδιαφέρον. Για παράδειγμα, ίσως θελήσει κάποιος να εξετάσει τι θα μπορούσε να έχει επηρεάσει την πιθανότητα, ό,τι η Γερουσιαστής Χίλαρι Κλίντον θα είχε υποστηρίξει ένα νομοσχέδιο μετανάστευσης, τον αριθμό των περιβαλλοντικών κανονισμών που θα εφαρμόσει η Ιρλανδία, σε ποιες χώρες ενδέχεται να συμβεί εμφύλιος πόλεμος ή πώς τα ιστορικά γεγονότα μπορεί να είχαν εξελιχθεί διαφορετικά. Οι Signorino και Tarar (2006) παρουσιάζουν στην μελέτη τους ένα ωραίο παράδειγμα, μέσα από το οποίο αναλύουν πώς η μεσοπρόθεσμη και βραχυπρόθεσμη ισορροπία των στρατιωτικών δυνάμεων θα μπορούσε να έχει τροποποιήσει τον αποκλεισμό του Βερολίνου του 1948 και τη Σοβιετική - Ιαπωνική σύγκρουση του 1937- 38 πάνω από το Manchukuo. Για καθένα από αυτά τα παραδείγματα, όπου οι υπολογισμοί επικεντρώνονται για όλες τις παρατηρήσεις του πληθυσμού, αποτελούν μέρος της προσέγγισης ΑΜΕ, καθώς αρχικά υπολογίζεται το ΜΕ για κάθε παρατήρηση ξεχωριστά και στη συνέχεια λαμβάνεται ο μέσος όρος των ΜΕ. Ως εκ τούτου, αν επιθυμεί κάποιος να επικεντρωθεί στην ερμηνεία για μία μόνο συγκεκριμένη περίπτωση, τότε θα πρέπει να υπάρχουν θεωρητικά επιχειρήματα, προκειμένου να αναφερθεί για ποιο λόγο θέλει να εξετάσει αυτή την συγκεκριμένη περίπτωση. Συνεπώς συνιστάται στους συγγραφείς να δηλώνουν τους λόγους για τους οποίους θεωρούν σημαντικό να μελετήσουν μία συγκεκριμένη περίπτωση (Hanmer & Ozan Kalkan, 2013).

Ενώ μπορεί να υπάρχει ένας καλός λόγος για να μελετήσει κάποιος εις βάθος μία συγκεκριμένη περίπτωση, όπως αναφέρθηκε προηγουμένως φαίνεται απίθανο να υπάρχει μια ορθή θεωρητική δικαιολογία για τη μελέτη της μέσης περίπτωσης. Για παράδειγμα, η εξέταση του ΜΕ της αύξησης του ισπανικού πληθυσμού στην πιθανότητα υποστήριξης της Χίλαρι Κλίντον για ένα νομοσχέδιο μετανάστευσης θα είχε μεγαλύτερο θεωρητικό ενδιαφέρον από ό,τι η εξέταση του ΜΕ για τον μέσο γερουσιαστή, ο οποίος μπορεί να μην υπάρχει στον πληθυσμό. Ακόμα κι αν ο μέσος γερουσιαστής υπάρχει στον πληθυσμό, η περίπτωση αυτή μπορεί να είναι εξαιρετικά σπάνια. Οι θεωρητικοί λόγοι, μπορεί να υπάρχουν για να επικεντρωθεί κάποιος σε μια συγκεκριμένη περίπτωση, ωστόσο οι Hanmer & Ozan Kalkan (2013), υποστηρίζουν ό,τι οι πολιτικοί επιστήμονες δεν πρέπει να σταματούν εκεί. Ως εκ τούτου, κάθε ενδελεχής έρευνα του ΜΕ μιας συγκεκριμένης μεμονωμένης περίπτωσης (ή συνδυασμό περιπτώσεων), προτείνεται να συνοδεύεται επί πλέον από το ΜΕ που υπολογίζεται μέσω της προσέγγισης ΑΜΕ, προκειμένου να επιτευχθεί μία σφαιρική εικόνα των αποτελεσμάτων που προκύπτουν από την αντίστοιχη ανάλυση.

Ωστόσο, δημιουργούνται ερωτηματικά σχετικά με την προσέγγιση ΜΕΜ, όπου αναλύονται μέσα από την μελέτη των Hanmer & Ozan Kalkan (2013). Τα ερωτηματικά αυτά αφορούν καταστάσεις όπου η μη ορθολογική χρήση της προσέγγισης ΜΕΜ, οδηγεί στη δημιουργία λανθασμένων, παραπλανητικών περιπτώσεων ως σημεία αναφοράς, από την οποία κανείς προσπαθεί να καταλήξει σε γενικευμένα συμπεράσματα. Παραδόξως, σχεδόν το 20% των άρθρων που χρησιμοποιούν την προσέγγιση ΜΕΜ, αναγνωρίζεται ό,τι θέτει όλες τις άλλες επεξηγηματικές μεταβλητές στις μέσες του τιμές, ακόμη και όταν αυτές ήταν κατηγορικές μεταβλητές. Η ανάθεση μιας "κεντρικής τιμής" (για παράδειγμα μέσης τιμής) σε δυαδικές επεξηγηματικές μεταβλητές μπορεί να κάνει "*την έννοια της κεντρικής τιμής, λιγότερο ουσιαστική* "(Gelman & Hill, 2007). Για παράδειγμα, μία δυαδική επεξηγηματική μεταβλητή, η οποία δηλώνει αν κάποιος έχει πτυχίο ή όχι, ο καθορισμός της τιμής στην μέση τιμή, δεν έχει νόημα καθώς ο υπολογισμός μιας πρόβλεψης με βάση την μέση τιμή αυτής της μεταβλητής, η οποία έχει οριστεί στο δείγμα να λαμβάνει τιμές 0 και 1, οποιαδήποτε τιμή διαφορετική από 0 ή 1, δεν έχει νόημα. Παρ' όλα αυτά οι απόψεις πολλών ερευνητών διίστανται. Για παράδειγμα ο Wooldridge (2002) προτείνει την επιλογή μιας ενδεικτικής τιμής για τις κατηγορικές μεταβλητές, λέγοντας ό,τι η τιμή αυτή " *is really based on taste*", ενώ οι Hanmer & Ozan Kalkan (2013) διαφωνούν. Παρόμοια προβλήματα προκύπτουν και με μοντέλα παλινδρόμησης, τα οποία περιλαμβάνουν τετραγωνικούς όρους ή όρους αλληλεπίδρασης. Για παράδειγμα θεωρώντας τετραγωνικούς όρους, στο *American National Election Studies (ANES)* του 2004 ο μέσος όρος της επεξηγηματικής μεταβλητής που δηλώνει την ηλικία, είναι 47, αλλά ο μέσος όρος της επεξηγηματικής μεταβλητής του μοντέλου, η οποία δηλώνει την ηλικία υψωμένη στο τετράγωνο είναι 2528, η τετραγωνική ρίζα του οποίου ισούται περίπου με 50. Γεγονός που σημαίνει ό,τι σε ένα μοντέλο που περιλαμβάνεται μία επεξηγηματική μεταβλητή καθώς και το τετράγωνό της, εξετάζοντας τα ΜΕ για όλες τις επεξηγηματικές μεταβλητές στις μέσες τους τιμές, είναι σύνηθες να καταλήξει κάποιος σε ένα μη-ορθολογικό αποτέλεσμα, καθώς όπως διαπιστώθηκε στο προηγούμενο παράδειγμα, η μέση τιμή της ηλικίας δεν μπορεί να είναι ταυτόχρονα 47 και 50 ετών.

Συνεπώς, η χρήση της προσέγγισης ΑΜΕ, εξυπηρετεί καλύτερα το στόχο των θεωρητικών εμπειρικών ερευνών που συνάπτουν συμπεράσματα σχετικά με τον πληθυσμό που ενδιαφέρει από το δείγμα. Όπως αναφέρουν και οι Hanushek and Jackson: "*Ας τονίσουμε εκ νέου ό,τι η ουσιαστική εμπειρική εργασία πρέπει να βασίζεται σε ρητές υποθέσεις και δηλώσεις σχετικά με την προβλεπόμενη συμπεριφορά*". Οι εκτιμήσεις από την προσέγγιση ΑΜΕ συνδέονται άμεσα με τις αρχικές υποθέσεις και πιο εύκολα επιτρέπουν στους ερευνητές για την αξιολόγηση των ουσιαστικών επιπτώσεων της θεωρίας τους.

Ωστόσο, είναι πολύ σημαντικό να σημειωθεί ό,τι η προσέγγιση ΑΜΕ δεν είναι αλάνθαστη. Αν και η μεγαλύτερη προσοχή δίνεται στον καθορισμό των τιμών των επεξηγηματικών μεταβλητών που δε διαχειρίζονται εύκολα, ο καθορισμός των τιμών των επεξηγηματικών μεταβλητών που διαχειρίζονται με μεγαλύτερη ευκολία, είναι εξίσου σημαντικός. Δηλαδή, οι ερευνητές πρέπει να δίνουν ιδιαίτερη προσοχή στον ορισμό των αντικρουόμενων τους γεγονότων, ώστε να είναι ρεαλιστικά για τον πληθυσμό που τους ενδιαφέρει (Gelman & Pardoe, 2007· King & Zeng, 2006).

Κεφάλαιο 4

4.1 Τα τυπικά σφάλματα των εκτιμημένων Οριακών Επιδράσεων

Προκειμένου να υλοποιηθούν στατιστικοί έλεγχοι υποθέσεων και να κατασκευαστούν διαστήματα εμπιστοσύνης για τα ΜΕ, απαιτείται πρώτα η εκτίμηση των ασυμπτωτικών τυπικών σφαλμάτων. Στις επόμενες ενότητες γίνεται αρχικά μία αναφορά στα τυπικά σφάλματα των ΜΕ για τα αμιγώς γραμμικά μοντέλα παλινδρόμησης και στην συνέχεια για τα πιο πολύπλοκα μη γραμμικά μοντέλα.

4.1.1 Τυπικά σφάλματα Οριακών Επιδράσεων για αμιγώς γραμμικά μοντέλα

Έστω ό,τι δίνεται ένα αμιγώς γραμμικό μοντέλο παλινδρόμησης:

$$Y_i = \boldsymbol{x_i'\beta} + \varepsilon_i \qquad (4.1)$$

όπου

- Y_i: τιμή της μεταβλητής απόκρισης για την i-οστή παρατήρηση
- $\boldsymbol{x_i'}$: είναι το διάνυσμα με τις τιμές των m επεξηγηματικών μεταβλητών της i-οστής παρατήρησης, συμπεριλαμβανομένου του σταθερού όρου
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)'$: είναι διάνυσμα με τις άγνωστες παραμέτρους $(m = p + 1)$
- ε_i: τυχαία σφάλματα με μέση τιμή 0 και σταθερή διασπορά σ^2

Θέτοντας:

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}_{n \text{ x } 1}, \qquad \boldsymbol{X} = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix}_{n \text{ x } m}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{\substack{m \text{ x } 1 \\ (m=p+1)}}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n \text{ x } 1}$$

Η σχέση (4.1), γράφεται ως εξής:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

όπου

- $\boldsymbol{X}$: πίνακας τάξης m

Οι συντελεστές $\boldsymbol{\beta}$ εκτιμώνται με την μέθοδο ελαχίστων τετραγώνων (OLS) και υπολογίζονται ως εξής:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \qquad (4.2)$$

όπου

- $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p\right)'$: είναι το διάνυσμα με τους εκτιμώμενους συντελεστές

Τα τυπικά σφάλματα των εκτιμώμενων συντελεστών $\widehat{\boldsymbol{\beta}}$, συμβολίζονται $\widehat{\boldsymbol{\sigma}}_{\widehat{\boldsymbol{\beta}}}$ και είναι τα διαγώνια στοιχεία του πίνακα διασπορών-συνδιασπορών:$\sqrt{\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}}$. Όπου η τυπική απόκλιση σ^2, εκτιμάται:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}}{n-m} = \frac{\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})^2}{n-m}$$

Συνεπώς, οι εκτιμητές των τυπικών σφαλμάτων των $\widehat{\boldsymbol{\beta}}$, ισούνται με την τετραγωνική ρίζα των διαγώνιων στοιχείων του εκτιμημένου πίνακα διασπορών-συνδιασπορών:

$$\widehat{\boldsymbol{\sigma}}_{\widehat{\boldsymbol{\beta}}} = \sqrt{\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}} \qquad (4.3)$$

Επιστρέφοντας στα ΜΕ, όπως είναι ήδη γνωστό από το 2ο κεφάλαιο, το ΜΕ μίας επεξηγηματικής μεταβλητής του παραπάνω μοντέλου (4.1) είναι σταθερό και ίσο με τον αντίστοιχο εκτιμώμενο συντελεστή (ανεξάρτητα από την φύση της μεταβλητής και των τιμών των υπολοίπων επεξηγηματικών μεταβλητών). Κατά συνέπεια, επειδή σε τέτοιου είδους μοντέλα τα ΜΕ ταυτίζονται με τους αντίστοιχους εκτιμώμενους συντελεστές $\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p\right)$, τα εκτιμώμενα τυπικά σφάλματα των ΜΕ είναι τα διαγώνια στοιχεία του πίνακα (4.3). Δηλαδή:

$$\widehat{\boldsymbol{\sigma}}_{\widehat{\boldsymbol{ME}}} = \widehat{\boldsymbol{\sigma}}_{\widehat{\boldsymbol{\beta}}} = \sqrt{\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}} \qquad (4.4)$$

4.1.2 Τυπικά σφάλματα Οριακών Επιδράσεων για μη γραμμικά μοντέλα

Ένα μοντέλο παλινδρόμησης, μπορεί να είναι μη γραμμικό με δύο τρόπους. Είτε όταν αποτελείται από μετασχηματισμένες επεξηγηματικές μεταβλητές είτε όταν αποτελείται από μία μετασχηματισμένη μεταβλητή απόκρισης.

4.1.2.1 Τυπικά σφάλματα Οριακών Επιδράσεων για μοντέλα με μετασχηματισμένες επεξηγηματικές μεταβλητές

Έστω ό,τι δίνεται το παρακάτω μοντέλο:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + (\text{όροι ανεξάρτητοι του } x_1) + \varepsilon$$

Όπως έχει ήδη υπολογιστεί, το ΜΕ της συνεχούς επεξηγηματικής μεταβλητής x_1, διατηρώντας τις υπόλοιπες επεξηγηματικές μεταβλητές σε σταθερές τιμές, εκτιμάται:

$$\widehat{\mathrm{ME}}_{x_1} = \hat{\beta}_1 + 2\hat{\beta}_2 x_1 \qquad (4.5)$$

Συνεπώς, το εκτιμώμενο τυπικό σφάλμα του ΜΕ της x_1, με βάση τις βασικές ιδιότητες της διασποράς, υπολογίζεται ως εξής:

$$\widehat{\boldsymbol{\sigma}}_{\widehat{\mathrm{ME}}_{x_1}} = \sqrt{Var(\hat{\beta}_1) + 4x_1^2 Var(\hat{\beta}_2) + 4x_1 Cov(\hat{\beta}_1, \hat{\beta}_2)} \qquad (4.6)$$

Έστω τώρα ό,τι δίνεται το μοντέλο:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + (\text{όροι ανεξάρτητοι του } x_1 \& x_2) + \varepsilon$$

Το ΜΕ της συνεχούς επεξηγηματικής μεταβλητής x_1, διατηρώντας τις υπόλοιπες επεξηγηματικές μεταβλητές σε σταθερές τιμές, ισούται:

$$\widehat{\mathrm{ME}}_{x_1} = \hat{\beta}_1 + \hat{\beta}_3 x_2 \qquad (4.7)$$

Συνεπώς, το εκτιμώμενο τυπικό σφάλμα του ΜΕ της x_1, ισούται με:

$$\widehat{\boldsymbol{\sigma}}_{\widehat{\mathrm{ME}}_{x_1}} = \sqrt{Var(\hat{\beta}_1) + x_2^2 Var(\hat{\beta}_3) + 2x_2 Cov(\hat{\beta}_1, \hat{\beta}_3)} \qquad (4.8)$$

Συνοψίζοντας, για αυτού του είδους τα μοντέλα ο τρόπος εκτίμησης των τυπικών σφαλμάτων των ΜΕ, πραγματοποιείται χρησιμοποιώντας τις βασικές ιδιότητες της διασποράς.

4.1.2.2 Τυπικά σφάλματα Οριακών Επιδράσεων για τα Γενικευμένα Γραμμικά Μοντέλα

Έστω ό,τι δίνεται ένα μοντέλο GLM,

$$E(Y|\boldsymbol{x}) = h(\boldsymbol{x}'\boldsymbol{\beta}) \qquad (4.9)$$

όπου

- $h(\cdot)$: είναι μια παραγωγίσιμη συνάρτηση
- $\boldsymbol{x}' = (1, x_1, \dots, x_p)$: το διάνυσμα των m επεξηγηματικών μεταβλητών, συμπεριλαμβανομένου του σταθερού όρου
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)'$: είναι διάνυσμα με τις άγνωστες παραμέτρους

Τότε, το ΜΕ μιας συνεχούς επεξηγηματικής μεταβλητής x_j σε ένα μοντέλο GLM, με βάση την προσέγγιση MEM, εκτιμάται:

$$\widehat{\mathrm{MEM}}_{x_j} = \frac{\partial h(\bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}})}{\partial x_j} = h'(\bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}})\hat{\beta}_j \qquad (4.10)$$

όπου

- $h'(\cdot)$: η πρώτη παράγωγος της συνάρτησης $h(\cdot)$
- $\bar{\boldsymbol{x}}'$: εκφράζει το διάνυσμα το οποίο περιέχει τις μέσες τιμές των επεξηγηματικών μεταβλητών.

Ενώ, το ΜΕ μίας συνεχής επεξηγηματικής μεταβλητής x_j, σε ένα μοντέλο GLM, με βάση την προσέγγιση AME, εκτιμάται:

$$\widehat{\mathrm{AME}}_{x_j} = \hat{\beta}_j \frac{1}{n}\sum_{i=1}^{n} \frac{\partial h(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})}{\partial x_j} = \hat{\beta}_j \frac{1}{n}\sum_{i=1}^{n} h'(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) \qquad (4.11)$$

όπου

- $\boldsymbol{x}_i' = (1, x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{ip})$: είναι το διάνυσμα με τις τιμές των m επεξηγηματικών μεταβλητών της i-οστής παρατήρησης

Σημειώνεται ό,τι τα ΜΕ διαφέρουν ανάλογα με την προσέγγιση που θα χρησιμοποιεί για την εκτίμησή τους (MEM ή AME). Κατά συνέπεια και οι εκτιμήσεις των τυπικών σφαλμάτων των ΜΕ, θα διαφέρουν. Επί πλέον, ο υπολογισμός των τυπικών σφαλμάτων των ΜΕ για τα γενικευμένα γραμμικά μοντέλα, όπως διαπιστώνεται και από τις εξισώσεις (4.10) & (4.11), είναι πιο πολύπλοκος και περιλαμβάνει περισσότερα βήματα από ό,τι εφαρμογή των βασικών ιδιοτήτων της διασποράς. Αυτό αναλύεται στην επόμενη ενότητα.

4.2 Μέθοδοι υπολογισμού τυπικών σφαλμάτων των Οριακών Επιδράσεων

Υπάρχουν αρκετοί μέθοδοι υπολογισμού τυπικών σφαλμάτων των ΜΕ, μερικές από τις οποίες, είναι:

1) Η μέθοδος δέλτα
2) Η μέθοδος των Krinsky & Robb (K-R)
3) Η μέθοδος bootstrap

Η επιλογή για το ποια μέθοδος θα χρησιμοποιηθεί, εξαρτάται συχνά από την εφαρμογή. Σε γενικές γραμμές δεν υπάρχει προτίμηση για το ποια από τις τρεις μεθόδους είναι πιο

κατάλληλη. Στις περισσότερες εφαρμογές, η επιλογή μεταξύ των τριών μεθόδων βασίζεται ως επί τον πλείστον στον προγραμματισμό ή στην υπολογιστική ευκολία. Ωστόσο, οι παραδοχές στις οποίες βασίζεται κάθε μέθοδος είναι σημαντικές. Οι ιδιότητες των εκτιμητών που συζητούνται σε αυτήν την ενότητα είναι ασυμπτωτικές και η συμπεριφορά των εκτιμητών σε πεπερασμένα δείγματα μπορεί να τεθεί σε κίνδυνο εάν οι παραδοχές που έχουν γίνει δεν ικανοποιούνται.

4.3 Μέθοδος δέλτα

Η μέθοδος δέλτα είναι η μέθοδος που χρησιμοποιούν κατά κόρον τα περισσότερα πακέτα λογισμικού για την εκτίμηση των τυπικών σφαλμάτων των ΜΕ. Βασίζεται στο προσεγγιστικό πολυώνυμο Taylor πρώτης τάξης για τη συνάρτηση $h(\cdot)$ αναπτύσσοντας την $h(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}})$ για $\widehat{\boldsymbol{\beta}}$ γύρω από το πραγματικό $\boldsymbol{\beta}$, προκειμένου να εκτιμηθούν τα τυπικά σφάλματα. (William H. Greene, 2012). Ο γενικός τύπος υπολογισμού, ανεξαρτήτως από το ποια προσέγγιση εκτιμήθηκαν τα ΜΕ, είναι:

$$Var(\widehat{\mathbf{ME}}) = \nabla\widehat{\mathbf{ME}}^T Var(\widehat{\boldsymbol{\beta}})\nabla\widehat{\mathbf{ME}} \qquad (4.12)$$

όπου

- $\nabla\widehat{\mathbf{ME}}$: είναι ένας p x m πίνακας, που περιέχει τις μερικές παραγώγους των ΜΕ ως προς τους συντελεστές του μοντέλου (Υπενθυμίζεται ό,τι p είναι το πλήθος των επεξηγηματικών μεταβλητών και κατ' επέκταση το πλήθος των ΜΕ του μοντέλου και m είναι το πλήθος των συντελεστών του μοντέλου).
- $Var(\widehat{\boldsymbol{\beta}})$: είναι ο πίνακας διασπορών-συνδιασπορών των εκτιμώμενων συντελεστών $\widehat{\boldsymbol{\beta}}$

Έστω ό,τι δίνεται ένα μοντέλο παλινδρόμησης το οποίο αποτελείται από p επεξηγηματικές μεταβλητές και από m συντελεστές. Αρχικά πρέπει να εκτιμηθούν τα ΜΕ των p επεξηγηματικών μεταβλητών (με όποια από τις τρεις προσεγγίσεις του κεφαλαίου 3 επιθυμεί ο εκάστοτε αναλυτής), κατά συνέπεια θα υπολογιστούν p το πλήθος ΜΕ. Οπότε σύμφωνα με τον τύπο (4.12), ο πίνακας p x p πίνακας διασπορών-συνδιασπορών των εκτιμώμενων **ΜΕ**, $\widehat{Var}(\widehat{\mathbf{ME}})$, προκύπτει από τις παρακάτω πράξεις μεταξύ των πινάκων:

$$\begin{bmatrix} \frac{\partial\widehat{\mathrm{ME}}_{x_1}}{\partial\hat{\beta}_0} & \cdots & \frac{\partial\widehat{\mathrm{ME}}_{x_1}}{\partial\hat{\beta}_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial\widehat{\mathrm{ME}}_{x_p}}{\partial\hat{\beta}_0} & \cdots & \frac{\partial\widehat{\mathrm{ME}}_{x_p}}{\partial\hat{\beta}_p} \end{bmatrix}_{p\,\mathrm{x}\,m} \begin{bmatrix} Var(\hat{\beta}_0) & \cdots & Cov(\hat{\beta}_0,\hat{\beta}_p) \\ \vdots & \ddots & \vdots \\ Cov(\hat{\beta}_p,\hat{\beta}_0) & \cdots & Var(\hat{\beta}_p) \end{bmatrix}_{m\,\mathrm{x}\,m} \begin{bmatrix} \frac{\partial\widehat{\mathrm{ME}}_{x_1}}{\partial\hat{\beta}_0} & \cdots & \frac{\partial\widehat{\mathrm{ME}}_{x_p}}{\partial\hat{\beta}_0} \\ \vdots & \ddots & \vdots \\ \frac{\partial\widehat{\mathrm{ME}}_{x_1}}{\partial\hat{\beta}_p} & \cdots & \frac{\partial\widehat{\mathrm{ME}}_{x_p}}{\partial\hat{\beta}_p} \end{bmatrix}_{m\,\mathrm{x}\,p}$$

Εφόσον γίνουν οι παραπάνω πράξεις μεταξύ των πινάκων, προκύπτει ο p x p πίνακας $\widehat{Var}(\widehat{\mathbf{ME}})$, ο οποίος θα έχει την μορφή:

$$\widehat{Var}(\widehat{\mathbf{ME}}) = \begin{bmatrix} \widehat{Var}(\widehat{\mathrm{ME}}_{x_1}) & \cdots & \widehat{Cov}(\widehat{\mathrm{ME}}_{x_1}, \widehat{\mathrm{ME}}_{x_p}) \\ \vdots & \ddots & \vdots \\ \widehat{Cov}(\widehat{\mathrm{ME}}_{x_1}, \widehat{\mathrm{ME}}_{x_p}) & \cdots & \widehat{Var}(\widehat{\mathrm{ME}}_{x_p}) \end{bmatrix}_{p \text{ x } p} \quad (4.13)$$

Συνεπώς, η εκτίμηση των τυπικών σφαλμάτων των ΜΕ, ισούται με την τετραγωνική ρίζα των διαγώνιων στοιχείων του πίνακα $\widehat{Var}(\widehat{\mathbf{ME}})$.

Ένα από το πλεονεκτήματα της συγκεκριμένης μεθόδου είναι ό,τι το μοντέλο εκτιμάται μόνο μια φορά. Αντιθέτως ένα μειονέκτημα που προκύπτει, είναι ό,τι για ορισμένα μοντέλα και ορισμένα πακέτα λογισμικού η διαδικασία αυτή δεν γίνεται αυτόματα αλλά χρειάζεται ο αναλυτής να κατασκευάσει χειροκίνητα τον κώδικα. Πιο συγκεκριμένα, σε αυτές τις περιπτώσεις απαιτείται από τον αναλυτή να παράσχει τις μερικές παραγώγους για τα *p* ΜΕ σε σχέση με τους *m* εκτιμημένους συντελεστές και στη συνέχεια να εφαρμόσει τον τύπο (4.13), γεγονός που καθιστά τη διαδικασία χρονοβόρα και εμπεριέχει μεγάλη πιθανότητα προγραμματιστικού λάθους. Ωστόσο, στα περισσότερα σύγχρονα πακέτα λογισμικού, η παραπάνω διαδικασία γίνεται αυτοματοποιημένα (Dowd, Greene & Norton, 2014).

4.3.1 Μέθοδος δέλτα σε μοντέλα GLM

Έστω ό,τι δίνεται ένα μοντέλο GLM,

$$E(Y|\boldsymbol{x}) = h(\boldsymbol{x_0'\beta})$$

όπου

- $h(\cdot)$: είναι μια παραγωγίσιμη συνάρτηση
- $\boldsymbol{x_0}$: είναι το διάνυσμα με τις *m* επεξηγηματικές μεταβλητές (συμπεριλαμβανομένου του σταθερού όρου) οι οποίες είναι "κεντραρισμένες" σε σταθερές τιμές
- x_j: είναι η *j*-οστή επεξηγηματική μεταβλητή, για $j = 1, \ldots, m$ επεξηγηματικές μεταβλητές
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)'$: είναι διάνυσμα με τις άγνωστες παραμέτρους

Η εκτίμηση του πίνακα διασπορών-συνδιασπορών της συνάρτησης ενδιαφέροντος, $h(\boldsymbol{x}_0'\boldsymbol{\beta})$, με βάση την μέθοδο δέλτα, δίνεται από τον παρακάτω τύπο:

$$Var\left(\hat{h}(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}})\right) = \nabla\hat{h}(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}})^T Var(\widehat{\boldsymbol{\beta}}) \nabla\hat{h}(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}) \quad (4.14)$$

Ο πίνακας $\nabla\hat{h}\left(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}\right)$ υπολογίζεται ως εξής:

$$\left(\frac{\partial\hat{h}\left(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}\right)}{\partial\widehat{\boldsymbol{\beta}}}\right) = \hat{h}'\left(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}\right)\boldsymbol{x}_0$$

Οπότε, από τις παραπάνω σχέσεις, ο γενικός τύπος εκτίμησης του πίνακα διασπορών-συνδιασπορών της προβλεπόμενης πιθανότητας με βάση την μέθοδο δέλτα, θα δίνεται από τον πίνακα (Greene, 2003):

$$Var\left(\hat{h}\left(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}\right)\right) = \hat{h}'\left(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}\right)^2\boldsymbol{x}_0'Var\left(\widehat{\boldsymbol{\beta}}\right)\boldsymbol{x}_0' \qquad (4.15)$$

Η σχέση (4.15), είναι πολύ χρήσιμη στην περίπτωση που θέλουμε να εκτιμήσουμε τα τυπικά σφάλματα του ΜΕ μίας διακριτής επεξηγηματικής μεταβλητής.

Πιο συγκεκριμένα, για την περίπτωση μίας δίτιμης επεξηγηματικής μεταβλητής έστω x_j, είναι γνωστό ό,τι το ΜΕ, εκτιμάται με βάση τον τύπο:

$$\widehat{\mathrm{ME}}_{x_j} = \hat{h}\left(\boldsymbol{x}_0'^{(-j)}\,\widehat{\boldsymbol{\beta}}^{(-j)} + \hat{\beta}_j\right) - \hat{h}\left(\boldsymbol{x}_0'^{(-j)}\,\widehat{\boldsymbol{\beta}}^{(-j)}\right)$$

Οπότε η ασυμπτωτική διασπορά του ME_{x_j}, εκτιμάται ως εξής:

$$\widehat{Var}\left(\widehat{\mathrm{ME}}_{x_j}\right) = \left[\frac{\partial\widehat{\mathrm{ME}}_{x_j}}{\partial\widehat{\boldsymbol{\beta}}}\right]^T Var\left(\widehat{\boldsymbol{\beta}}\right)\left[\frac{\partial\widehat{\mathrm{ME}}_{x_j}}{\partial\widehat{\boldsymbol{\beta}}}\right] \qquad (4.16)$$

Όπου

$$\left[\frac{\partial\widehat{\mathrm{ME}}_{x_j}}{\partial\widehat{\boldsymbol{\beta}}}\right] = \hat{h}'\left(\boldsymbol{x}_0'^{(-j)}\,\widehat{\boldsymbol{\beta}}^{(-j)} + \hat{\beta}_j\right)\begin{bmatrix}1\\1\\\boldsymbol{x}_{(j)}\end{bmatrix} - \hat{h}'\left(\boldsymbol{x}_0'^{(-j)}\,\widehat{\boldsymbol{\beta}}^{(-j)}\right)\begin{bmatrix}1\\0\\\boldsymbol{x}_{(j)}\end{bmatrix} \qquad (4.17)$$

όπου

- $\boldsymbol{x}_{(j)} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)'$: είναι ένας πίνακας με τις "κεντραρισμένες" τιμές των επεξηγηματικών μεταβλητών, εξαιρώντας την τιμή της επεξηγηματικής μεταβλητής x_j

Για την εκτίμηση του πίνακα διασπορών-συνδιασπορών των εκτιμώμενων ΜΕ όσον αφορά αποκλειστικά τις συνεχείς επεξηγηματικές μεταβλητές, ακολουθείται η παρακάτω γενικευμένη διαδικασία. Πιο συγκεκριμένα, για την περίπτωση μίας συνεχούς επεξηγηματικής μεταβλητής έστω x_j, είναι γνωστό ό,τι το ΜΕ, υπολογίζεται με βάση τον τύπο:

$$\mathrm{ME}_{x_j} = \frac{\partial h\left(\boldsymbol{x}_0'\boldsymbol{\beta}\right)}{\partial x_j} = h'\left(\boldsymbol{x}_0'\boldsymbol{\beta}\right)\beta_j$$

Οπότε ο ασυμπτωτικός πίνακας διασπορών-συνδιασπορών των ΜΕ των m συνεχών επεξηγηματικών μεταβλητών, με βάση τον τύπο (4.12), εκτιμάται:

$$\widehat{Var}\left(\widehat{\mathbf{ME}}\right) = \nabla\hat{h}\left(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}\right)^T Var\left(\widehat{\boldsymbol{\beta}}\right)\nabla\hat{h}\left(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}\right) \qquad (4.18)$$

όπου

$$\nabla\hat{h}(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}) = \hat{h}'(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}})\boldsymbol{I} + \left(\frac{\partial\hat{h}'(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{x}\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}}\right)\boldsymbol{x}_0\widehat{\boldsymbol{\beta}}'$$

όπου

- $\boldsymbol{I}$: ο ταυτοτικός πίνακας

Συνεπώς, από τις παραπάνω σχέσεις, ο γενικός τύπος εκτίμησης του πίνακα διασπορών-συνδιασπορών των ΜΕ των m συνεχών επεξηγηματικών μεταβλητών, $\widehat{Var}(\widehat{\mathbf{ME}})$, ισούται:

$$\left[\hat{h}'(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}})\boldsymbol{I} + \left(\frac{\partial\hat{h}'(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}}\right)\widehat{\boldsymbol{\beta}}\boldsymbol{x}_0'\right]Var(\widehat{\boldsymbol{\beta}})\left[\hat{h}'(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}})\boldsymbol{I} + \left(\frac{\partial\hat{h}'(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}})}{\partial\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}}}\right)\boldsymbol{x}_0\widehat{\boldsymbol{\beta}}'\right] \qquad (4.19)$$

Προηγουμένως, παρουσιάστηκε η μέθοδος δέλτα θεωρώντας δεδομένο ό,τι οι επεξηγηματικές μεταβλητές της συνάρτηση ενδιαφέροντος $h(\boldsymbol{x}_0'\boldsymbol{\beta})$, είναι σταθερές τιμές (συνήθως οι μέσες τους τιμές). Ωστόσο, χρειάζεται ιδιαίτερη προσοχή όταν πρόκειται να χρησιμοποιηθεί η μέθοδος δέλτα για την μέση συνάρτηση ενδιαφέροντος $\overline{h(\boldsymbol{x}'\boldsymbol{\beta})}$. Για αυτόν το λόγο σημειώνεται ό,τι η μέθοδος δέλτα αντιμετωπίζει την συνάρτηση $\overline{h(\boldsymbol{x}'\boldsymbol{\beta})}$, ως εξής:

$$\frac{\partial\left(\frac{1}{n}\sum_{i=1}^{n}h(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})\right)}{\partial\widehat{\boldsymbol{\beta}}} = \frac{1}{n}\frac{\partial\sum_{i=1}^{n}h(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})}{\partial\widehat{\boldsymbol{\beta}}} = \overline{\left[\frac{\partial h(\boldsymbol{x}'\boldsymbol{\beta})}{\partial\widehat{\boldsymbol{\beta}}}\right]}$$

Συνεπώς, η εκτίμηση του πίνακα διασπορών-συνδιασπορών $\overline{h(\boldsymbol{x}'\boldsymbol{\beta})}$, με βάση την μέθοδο δέλτα, δίνεται από τον παρακάτω τύπο:

$$Var\left(\overline{\hat{h}(\boldsymbol{x}'\boldsymbol{\beta})}\right) = \overline{\nabla\hat{h}(\boldsymbol{x}'\boldsymbol{\beta})}^{T}Var(\widehat{\boldsymbol{\beta}})\overline{\nabla\hat{h}(\boldsymbol{x}'\boldsymbol{\beta})} \qquad (4.20)$$

Κατ' επέκταση η εκτίμηση του πίνακα διασπορών-συνδιασπορών των ΑΜΕ, θα δίνεται από τον τύπο:

$$\widehat{Var}(\widehat{\mathbf{ME}}) = \overline{\left[\frac{\partial\widehat{\mathbf{ME}}}{\partial\widehat{\boldsymbol{\beta}}}\right]}^{T}Var(\widehat{\boldsymbol{\beta}})\overline{\left[\frac{\partial\widehat{\mathbf{ME}}}{\partial\widehat{\boldsymbol{\beta}}}\right]} \qquad (4.21)$$

4.3.1.1 Μέθοδος Δέλτα σε μοντέλα logit

Για την εκτίμηση των τυπικών σφαλμάτων των ΜΕ συνεχών επεξηγηματικών μεταβλητών ενός μοντέλου logit, με βάση τον τύπο (4.12), ισχύει:

$$\widehat{Var}(\widehat{\mathbf{ME}}) = \left[\frac{\partial\Lambda(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}})[1-\Lambda(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}})]\widehat{\boldsymbol{\beta}}}{\partial\widehat{\boldsymbol{\beta}}}\right]^{T}Var(\widehat{\boldsymbol{\beta}})\left[\frac{\partial\Lambda(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}})[1-\Lambda(\boldsymbol{x}_0'\widehat{\boldsymbol{\beta}})]\widehat{\boldsymbol{\beta}}}{\partial\widehat{\boldsymbol{\beta}}}\right] \qquad (4.22)$$

Επί πλέον αποδεικνύεται εύκολα οτι:

$$\frac{\partial \Lambda(x)\big(1-\Lambda(x)\big)}{\partial x} = \Lambda(x)\big(1-\Lambda(x)\big)\big(1-2\Lambda(x)\big) \qquad (4.23)$$

Οπότε με βάση τον τύπο (4.23), ο πίνακας $\nabla\widehat{\mathbf{ME}}$, ισούται:

$$\left[\frac{\partial \Lambda\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\left(1-\Lambda\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\right)\widehat{\boldsymbol{\beta}}}{\partial \widehat{\boldsymbol{\beta}}}\right] = \Lambda\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\left(1-\Lambda\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\right)\left[\boldsymbol{I}+\left(1-2\Lambda\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\right)\left(\boldsymbol{x_0}\widehat{\boldsymbol{\beta}}'\right)\right]$$

Θέτοντας,

$$l\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right) = \Lambda\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\left(1-\Lambda\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\right)$$

τα εκτιμώμενα τυπικά σφάλματα των ΜΕ, ισούνται με τις τετραγωνικές ρίζες των διαγώνιων στοιχείων του πίνακα

$$l(\boldsymbol{x_0'}\boldsymbol{\beta})^2\left[\boldsymbol{I}+\left(1-2\Lambda\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\right)\left(\widehat{\boldsymbol{\beta}}\boldsymbol{x_0'}\right)\right]Var\left(\widehat{\boldsymbol{\beta}}\right)\left[\boldsymbol{I}+\left(1-2\Lambda\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\right)\left(\boldsymbol{x_0}\widehat{\boldsymbol{\beta}}'\right)\right] \qquad (4.24)$$

4.3.1.2 Μέθοδος Δέλτα σε μοντέλα probit

Για την εκτίμηση των τυπικών σφαλμάτων των ΜΕ συνεχών επεξηγηματικών μεταβλητών ενός μοντέλου probit, με βάση τον τύπο (4.12), ισχύει:

$$\widehat{Var}\left(\widehat{\mathbf{ME}}\right) = \left[\frac{\partial\varphi\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\widehat{\boldsymbol{\beta}}}{\partial\widehat{\boldsymbol{\beta}}}\right]^T Var\left(\widehat{\boldsymbol{\beta}}\right)\left[\frac{\partial\varphi\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\widehat{\boldsymbol{\beta}}}{\partial\widehat{\boldsymbol{\beta}}}\right] \qquad (4.25)$$

Από το οποίο αποδεικνύεται εύκολα ό,τι ο πίνακας $\nabla\widehat{\mathbf{ME}}$, ισούται με:

$$\left[\frac{\partial\varphi\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\widehat{\boldsymbol{\beta}}}{\partial\widehat{\boldsymbol{\beta}}}\right] = \varphi\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\left[\boldsymbol{I}-\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\left(\boldsymbol{x_0}\widehat{\boldsymbol{\beta}}'\right)\right]$$

Συνεπώς, από τις παραπάνω σχέσεις, συνεπάγεται ό,τι:

$$\widehat{Var}\left(\widehat{\mathbf{ME}}\right) = \varphi\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)^2\left[\boldsymbol{I}-\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\left(\widehat{\boldsymbol{\beta}}\boldsymbol{x_0'}\right)\right]Var\left(\widehat{\boldsymbol{\beta}}\right)\left[\boldsymbol{I}-\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\left(\boldsymbol{x_0}\widehat{\boldsymbol{\beta}}'\right)\right] \qquad (4.26)$$

4.3.1.3 Μέθοδος Δέλτα σε μοντέλα complementary log-log

Για την εκτίμηση των τυπικών σφαλμάτων των ΜΕ συνεχών επεξηγηματικών μεταβλητών ενός μοντέλου complementary log-log, με βάση τον τύπο (4.12), ισχύει:

$$\widehat{Var}\left(\widehat{\mathbf{ME}}\right) = \left[\frac{\partial e^{x_0'\widehat{\boldsymbol{\beta}}}\left(1 - W(\boldsymbol{x_0'\beta})\right)\widehat{\boldsymbol{\beta}}}{\partial \widehat{\boldsymbol{\beta}}}\right]^T Var\left(\widehat{\boldsymbol{\beta}}\right)\left[\frac{\partial e^{x_0'\widehat{\boldsymbol{\beta}}}\left(1 - W(\boldsymbol{x_0'\beta})\right)\widehat{\boldsymbol{\beta}}}{\partial \widehat{\boldsymbol{\beta}}}\right] \qquad (4.27)$$

Από το οποίο αποδεικνύεται εύκολα ό,τι ο πίνακας $\nabla\widehat{\mathbf{ME}}$, ισούται με:

$$\left[\frac{\partial e^{x_0'\widehat{\boldsymbol{\beta}}}\left(1 - W(\boldsymbol{x_0'\beta})\right)\widehat{\boldsymbol{\beta}}}{\partial \widehat{\boldsymbol{\beta}}}\right] = e^{x_0'\widehat{\boldsymbol{\beta}}}\left(1 - W\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\right)\left[\boldsymbol{I} + \left(1 - e^{x_0'\widehat{\boldsymbol{\beta}}}\right)\left(\boldsymbol{x_0}\widehat{\boldsymbol{\beta}}'\right)\right]$$

Θέτοντας,

$$w\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right) = e^{x_0'\widehat{\boldsymbol{\beta}}}\left(1 - W\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\right)$$

ο παραπάνω πίνακας $\widehat{Var}\left(\widehat{\mathbf{ME}}\right)$, διατυπώνεται ως εξής:

$$\widehat{Var}\left(\widehat{\mathbf{ME}}\right) = \left(w\left(\boldsymbol{x_0'}\widehat{\boldsymbol{\beta}}\right)\right)^2\left[\boldsymbol{I} + \left(1 - e^{x_0'\widehat{\boldsymbol{\beta}}}\right)\left(\widehat{\boldsymbol{\beta}}\boldsymbol{x_0'}\right)\right]Var\left(\widehat{\boldsymbol{\beta}}\right)\left[\boldsymbol{I} + \left(1 - e^{x_0'\widehat{\boldsymbol{\beta}}}\right)\left(\boldsymbol{x_0}\widehat{\boldsymbol{\beta}}'\right)\right] \qquad (4.28)$$

4.3.1.4 Μέθοδος Δέλτα σε μοντέλα Poisson

Για την εκτίμηση των τυπικών σφαλμάτων των ΜΕ συνεχών επεξηγηματικών μεταβλητών ενός μοντέλου Poisson, με βάση τον τύπο (4.12), ισχύει:

$$\widehat{Var}\left(\widehat{\mathbf{ME}}\right) = \left[\frac{\partial e^{x_0'\widehat{\boldsymbol{\beta}}}\widehat{\boldsymbol{\beta}}}{\partial \widehat{\boldsymbol{\beta}}}\right]^T Var\left(\widehat{\boldsymbol{\beta}}\right)\left[\frac{\partial e^{x_0'\widehat{\boldsymbol{\beta}}}\widehat{\boldsymbol{\beta}}}{\partial \widehat{\boldsymbol{\beta}}}\right] \qquad (4.29)$$

Από το οποίο αποδεικνύεται ό,τι ο πίνακας $\nabla\widehat{\mathbf{ME}}$, ισούται με:

$$\left[\frac{\partial e^{x_0'\widehat{\boldsymbol{\beta}}}\widehat{\boldsymbol{\beta}}}{\partial \widehat{\boldsymbol{\beta}}}\right] = e^{x_0'\widehat{\boldsymbol{\beta}}}\left[\boldsymbol{I} + \boldsymbol{x_0}\widehat{\boldsymbol{\beta}}'\right]$$

Συνεπώς, από τις παραπάνω σχέσεις, συνεπάγεται ό,τι:

$$\widehat{Var}\left(\widehat{\mathbf{ME}}\right) = e^{x_0'\widehat{\boldsymbol{\beta}}^2}\left[\boldsymbol{I} + \widehat{\boldsymbol{\beta}}\boldsymbol{x_0'}\right]Var\left(\widehat{\boldsymbol{\beta}}\right)\left[\boldsymbol{I} + \boldsymbol{x_0}\widehat{\boldsymbol{\beta}}'\right] \qquad (4.30)$$

4.3.1.5 Μέθοδος Δέλτα σε μοντέλα ZIP

Για την εκτίμηση των τυπικών σφαλμάτων των ΜΕ συνεχών επεξηγηματικών μεταβλητών ενός μοντέλου ZIP, η διαδικασία είναι πιο πολύπλοκη και δεν δίνεται άμεσα από τον τύπο (4.12). Προκείμενου να υπολογιστεί ο πίνακας $\nabla\widehat{\mathbf{ME}}$, αρκεί αρχικά να υπολογιστούν οι δύο επιμέρους πίνακες $\nabla\widehat{\mathbf{ME}}_{(\boldsymbol{\beta})}$ και $\nabla\widehat{\mathbf{ME}}_{(\boldsymbol{\gamma})}$ (οι οποίοι αποτελούνται από τις μερικές παραγώγους των ΜΕ ως προς τους συντελεστές $\boldsymbol{\beta}$ & $\boldsymbol{\gamma}$ αντίστοιχα):

$$\nabla\widehat{\mathbf{ME}}_{(\boldsymbol{\beta})} = A[\left(\boldsymbol{I} + \boldsymbol{x}\widehat{\boldsymbol{\beta}}'\right) + e^{\boldsymbol{x}'\widehat{\boldsymbol{\gamma}}}\left(\boldsymbol{I} + \boldsymbol{x}(\widehat{\boldsymbol{\beta}}' - \widehat{\boldsymbol{\gamma}}')\right)]$$

&

$$\nabla\widehat{\mathbf{ME}}_{(\boldsymbol{\gamma})} = B[\left(\boldsymbol{I} + \boldsymbol{x}(\widehat{\boldsymbol{\beta}}' + \widehat{\boldsymbol{\gamma}}')\right) + e^{\boldsymbol{x}'\widehat{\boldsymbol{\gamma}}}\left(\boldsymbol{I} + \boldsymbol{x}(\widehat{\boldsymbol{\beta}}' - \widehat{\boldsymbol{\gamma}}')\right)]$$

Όπου

$$A = \frac{\mu}{1 + e^{\boldsymbol{x}'\widehat{\boldsymbol{\gamma}}}} \qquad \& \qquad B = -\frac{\mu\,\pi}{1 + e^{\boldsymbol{x}'\widehat{\boldsymbol{\gamma}}}}$$

$$\pi = \frac{e^{\boldsymbol{x}_0'\widehat{\boldsymbol{\gamma}}}}{1 + e^{\boldsymbol{x}'\widehat{\boldsymbol{\gamma}}}} \qquad \& \qquad \mu = \frac{e^{\boldsymbol{x}'\widehat{\boldsymbol{\beta}}}}{1 + e^{\boldsymbol{x}'\widehat{\boldsymbol{\gamma}}}}$$

Οπότε ο πίνακας $\nabla\widehat{\mathbf{ME}}$, θα υπολογίζεται, ως εξής:

$$\nabla\widehat{\mathbf{ME}} = \begin{bmatrix} \nabla\widehat{\mathbf{ME}}_{(\boldsymbol{\beta})} \\ \nabla\widehat{\mathbf{ME}}_{(\boldsymbol{\gamma})} \end{bmatrix}$$

Συνεπώς, από τις παραπάνω σχέσεις, συνεπάγεται ό,τι:

$$\widehat{Var}\left(\widehat{\mathbf{ME}}\right) = \left[\nabla\widehat{\mathbf{ME}}_{(\boldsymbol{\beta})}{}^{T}\ \nabla\widehat{\mathbf{ME}}_{(\boldsymbol{\gamma})}{}^{T}\right] Var\left(\widehat{\boldsymbol{\beta}}\right) \begin{bmatrix} \nabla\widehat{\mathbf{ME}}_{(\boldsymbol{\beta})} \\ \nabla\widehat{\mathbf{ME}}_{(\boldsymbol{\gamma})} \end{bmatrix} \qquad (4.31)$$

Παρόμοια πολυπλοκότητα, παρουσιάζεται και για τον υπολογισμό των τυπικών σφαλμάτων των ΜΕ διακριτών επεξηγηματικών μεταβλητών. Ο τύπος (4.17) σε αυτή την περίπτωση δεν είναι αρκετός και η διαδικασία που ακολουθείται είναι η εξής:

Έστω ό,τι η x_j είναι μία διακριτή δίτιμη επεξηγηματική μεταβλητή, αρχικά πρέπει να υπολογιστούν οι δύο επιμέρους πίνακες $\nabla\widehat{\mathrm{ME}}_{x_j(\boldsymbol{\beta})}$ και $\nabla\widehat{\mathrm{ME}}_{x_j(\boldsymbol{\gamma})}$:

$$\nabla\widehat{\mathrm{ME}}_{x_j(\boldsymbol{\beta})} = \mu\left(x_j = 1\right)\begin{bmatrix} 1 \\ 1 \\ \boldsymbol{x}_{(j)} \end{bmatrix} - \mu\left(x_j = 0\right)\begin{bmatrix} 1 \\ 0 \\ \boldsymbol{x}_{(j)} \end{bmatrix}$$

&

$$\nabla\widehat{\mathrm{ME}}_{x_j(\boldsymbol{\gamma})} = -\mu(x_j=1)\pi(x_j=1)\begin{bmatrix}1\\1\\\boldsymbol{x}_{(j)}\end{bmatrix} + \mu(x_j=0)\pi(x_j=0)\begin{bmatrix}1\\0\\\boldsymbol{x}_{(j)}\end{bmatrix}$$

Όπου

$$\mu(x_j=1) = \frac{e^{\boldsymbol{x}'^{(-j)}\boldsymbol{\beta}^{(-j)}+\beta_j}}{1+e^{\boldsymbol{x}'^{(-j)}\boldsymbol{\gamma}^{(-j)}+\gamma_j}} \qquad \& \qquad \mu(x_j=0) = \frac{e^{\boldsymbol{x}'^{(-j)}\boldsymbol{\beta}^{(-j)}}}{1+e^{\boldsymbol{x}'^{(-j)}\boldsymbol{\gamma}^{(-j)}}}$$

$$\pi(x_j=1) = \frac{e^{\boldsymbol{x}'^{(-j)}\boldsymbol{\gamma}^{(-j)}+\gamma_j}}{1+e^{\boldsymbol{x}'^{(-j)}\boldsymbol{\gamma}^{(-j)}+\gamma_j}} \qquad \& \qquad \pi(x_j=0) = \frac{e^{\boldsymbol{x}'^{(-j)}\boldsymbol{\gamma}^{(-j)}}}{1+e^{\boldsymbol{x}'^{(-j)}\boldsymbol{\gamma}^{(-j)}}}$$

Οπότε ο πίνακας $\nabla\widehat{\mathrm{ME}}_{x_j}$, θα υπολογίζεται, ως εξής:

$$\nabla\widehat{\mathrm{ME}}_{x_j} = \begin{bmatrix}\nabla\widehat{\mathrm{ME}}_{x_j(\boldsymbol{\beta})}\\\nabla\widehat{\mathrm{ME}}_{x_j(\boldsymbol{\gamma})}\end{bmatrix}$$

Συνεπώς, από τις παραπάνω σχέσεις, συνεπάγεται ό,τι:

$$\widehat{Var}\left(\widehat{\mathrm{ME}}_{x_j}\right) = \left[\nabla\widehat{\mathrm{ME}}_{x_j(\boldsymbol{\beta})}{}^{T}\ \nabla\widehat{\mathrm{ME}}_{x_j(\boldsymbol{\gamma})}{}^{T}\right] Var(\widehat{\boldsymbol{\beta}}) \begin{bmatrix}\nabla\widehat{\mathrm{ME}}_{x_j(\boldsymbol{\beta})}\\\nabla\widehat{\mathrm{ME}}_{x_j(\boldsymbol{\gamma})}\end{bmatrix} \qquad (4.32)$$

4.4 Μέθοδος των Krinsky & Robb (K-R)

Η μέθοδος K-R βασίζεται στην υπόθεση ό,τι οι εκτιμώμενοι συντελεστές του μοντέλου είναι σταθεροί και προέρχονται από την ασυμπτωτική πολυδιάστατη κανονική κατανομή. Η μέθοδος K-R παράγει πολλαπλά διανύσματα συντελεστών $\boldsymbol{\beta} = \boldsymbol{\beta}^s$ για s = 1, …, S από τη πολυδιάστατη κανονική κατανομή με μέσο ίσο με το διάνυσμα των εκτιμώμενων συντελεστών $\widehat{\boldsymbol{\beta}}$ και πίνακα διασπορών-συνδιασπορών $Var(\widehat{\boldsymbol{\beta}})$. Έπειτα, κάθε ένα καινούργιο διάνυσμα συντελεστών $\boldsymbol{\beta}^s$, χρησιμοποιείται για τον υπολογισμό των τιμών των ΜΕ της μεταβλητής. Η τυπική απόκλιση αυτών των εκτιμώμενων ΜΕ, αποτελεί έναν εκτιμητή του τυπικού σφάλματος του ΜΕ της μεταβλητής. Στην ουσία είναι μία Monte Carlo προσομοίωση για την εκτίμηση του $\widehat{Var}(\mathbf{ME})$ η οποία επικαλείται το νόμο των μεγάλων αριθμών (Onukwugha, Bergtold, & Jain, 2014).

Σημειώνεται ό,τι τα πακέτα λογισμικού χρησιμοποιούν διαφορετικό τρόπο για την δημιουργία δειγμάτων, κατά συνέπεια είναι σχεδόν αδύνατον να αναπαράγονται ακριβώς

παρόμοια αποτελέσματα (Dowd, Greene & Norton, 2014). Τέλος ένα μειονέκτημα της μεθόδου των Krinsky και Robb σε σχέση με τη μέθοδο Δέλτα είναι ό,τι είναι πιο απαιτητική ως προς τους υπολογισμούς, δεδομένου ό,τι απαιτεί έναν μεγάλο αριθμό προσομοιώσεων. Λαμβάνοντας υπόψη όμως την ταχύτητα των σύγχρονων υπολογιστών, δεν αποτελεί πλέον σημαντικό εμπόδιο (Hole, 2008).

4.5 Μέθοδος bootstrap

Όπως η μέθοδος K-R, έτσι και η μέθοδος bootstrap (Efron, 1979) εφαρμόζει πολλά κατασκευασμένα διανύσματα συντελεστών $\boldsymbol{\beta}^s$, στις δειγματοληπτικές παρατηρήσεις, με την διαφορά ό,τι η διασπορά στα διανύσματα των συντελεστών λαμβάνονται με εκ νέου εκτίμηση του μοντέλου πολλές φορές σε διαφορετικά δείγματα. Κάθε νέο δείγμα λαμβάνεται παίρνοντας *n* παρατηρήσεις με αντικατάσταση από το αρχικό δείγμα. Το μέγεθος του νέου δείγματος είναι ίσο με το μέγεθος του αρχικού δείγματος. Το μοντέλο στη συνέχεια επανεκτιμάται στο νέο δείγμα, με αποτέλεσμα να προκύψει ένα καινούργιο διάνυσμα με εκτιμημένους συντελεστές. Έπειτα, για το κάθε νέο διάνυσμα υπολογίζονται τα ΜΕ της x_j μεταβλητής. Συνεπώς η τυπική απόκλιση που προκύπτει από την κατανομή των εκτιμώμενων ΜΕ που έχουν υπολογιστεί για την x_j μεταβλητή, για όλα τα διανύσματα $\widehat{\boldsymbol{\beta}}$, αποτελεί έναν εκτιμητή του τυπικού σφάλματος της ΜΕ της x_j μεταβλητής.

Συνοπτικά, η μέθοδος bootstrap, πραγματοποιεί μία προσομοίωση κατασκευάζοντας ψευδοτυχαία δείγματα (εφαρμόζοντας δειγματοληψία με αντικατάσταση), με σκοπό να εκτιμήσει τα διανύσματα των συντελεστών για το κάθε ψευδό-τυχαίο δείγμα αντίστοιχα, υποθέτοντας ό,τι το πρωτότυπο δείγμα το οποίο χρησιμοποιείται για την εκτίμηση του μοντέλου είναι μία προσέγγιση του πληθυσμού. Είναι μία μέθοδος εύκολη στην υλοποίησή της και είναι προγραμματισμένη σε πολλά πακέτα λογισμικού (Hannachi, 2006).

Κεφάλαιο 5

5.1 Απλοποίηση των Οριακών Επιδράσεων στα GLM

Οι Anderson & Newell (2003), μέσα από το έργο τους υπέδειξαν πώς από μια απλή τυποποίηση των επεξηγηματικών μεταβλητών, τα ΜΕ των επεξηγηματικών μεταβλητών στα γενικευμένα γραμμικά μοντέλα logit και probit, μπορούν να απλοποιηθούν δραματικά. Σε αυτό το κεφάλαιο γίνεται παρουσίαση της ερευνάς τους. Επί πλέον, διατυπώνονται παρόμοιες απλοποιήσεις για την εκτίμηση των τυπικών σφαλμάτων των ΜΕ, βάσει της μεθόδου δέλτα. Ωστόσο, παρόλο που οι Anderson & Newell (2003), στην εργασία τους επικεντρώθηκαν κυρίως στην απλοποίηση των ΟΕ για τα μοντέλα logit και probit, το ίδιο συμπέρασμα ως προς τις απλοποιήσεις των ΟΕ, προκύπτει και στις περιπτώσεις των μοντέλων complementary log-log, Poisson και ZIP.

5.2 Τυποποίηση των επεξηγηματικών μεταβλητών

Για την τυποποίηση έστω μίας x_j επεξηγηματικής μεταβλητής, οι οποία αποτελείται από *n* στοιχεία, δηλαδή $x_j = (x_{1j}, \dots, x_{ij}, \dots, x_{nj})'$ ακολουθείται η εξής διαδικασία:

1ο Βήμα Υπολογίζεται η μέση τιμή της x_j

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

2ο Βήμα Στην συνέχεια, υπολογίζεται η τυπική της απόκλιση

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}{n-1}}$$

3ο Βήμα Συνεπώς, η τυποποιημένη μεταβλητή της x_j, προκύπτει από τον ακόλουθο τύπο

$$x_j^* = \frac{x_j - \bar{x}_j}{\sigma_j}$$

Από τον οποίον προκύπτουν τα εξής πολύ σημαντικά συμπεράσματα:

$$E(x_j^*) = 0 \qquad (5.2)$$

$$Var(x_j^*) = 1 \qquad (5.3)$$

Με παρόμοιο τρόπο, αντιμετωπίζονται και οι διακριτές επεξηγηματικές μεταβλητές, αρκεί να κωδικοποιούνται με ορθό τρόπο (δηλαδή, 0/1) έτσι ώστε ο σταθερός όρος να αντιστοιχεί στην επιθυμητή ομάδα αναφοράς για την οποία θα υπολογιστούν τα ΜΕ.

5.3 Οριακές Επιδράσεις τυποποιημένων επεξηγηματικών μεταβλητών

Έστω ό,τι η προβλεπόμενη πιθανότητα, ενός διωνυμικού μοντέλου, δίνεται από το παρακάτω μοντέλο παλινδρόμησης:

$$E(Y|\boldsymbol{x}) = F(\boldsymbol{x}'\boldsymbol{\beta}) \qquad (5.4)$$

Όπου

- x_j: είναι η j-οστή επεξηγηματική μεταβλητή, για $j = 1, \dots, m$ επεξηγηματικές μεταβλητές
- $\boldsymbol{x}' = (1, x_1, \dots, x_p)$: το διάνυσμα των m επεξηγηματικών μεταβλητών, συμπεριλαμβανομένου του σταθερού όρου
- $\boldsymbol{x_i} = (1, x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{ip})'$: είναι το διάνυσμα με τις τιμές των m επεξηγηματικών μεταβλητών της i-οστής παρατήρησης
- $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \beta_2, \dots, \beta_p\right)'$: είναι διάνυσμα με τις άγνωστες παραμέτρους

5.3.1 Οριακές Επιδράσεις τυποποιημένων συνεχών επεξηγηματικών μεταβλητών

Όπως έχει προαναφερθεί στα προηγούμενα κεφάλαια, για τον υπολογισμό του ΜΕ, μίας έστω x_j συνεχής επεξηγηματικής μεταβλητής (διατηρώντας τις υπόλοιπες επεξηγηματικές μεταβλητές σταθερές), χρησιμοποιείται ο παρακάτω τύπος:

$$ME_{x_j} = \frac{\partial F(\boldsymbol{x}'\boldsymbol{\beta})}{\partial x_j} = f(\boldsymbol{x}'\boldsymbol{\beta})\beta_j \qquad (5.5)$$

όπου

- $f(\cdot)$: δηλώνει την αντίστοιχη συνάρτηση πυκνότητας πιθανότητας

Για το μοντέλο logit η συνάρτηση πυκνότητας πιθανότητας $f(\cdot)$, δίνεται από την $l(\cdot)$ την λογιστική συνάρτηση κατανομής:

$$l(\boldsymbol{x}'\boldsymbol{\beta}) = \Lambda(\boldsymbol{x}'\boldsymbol{\beta})[1 - \Lambda(\boldsymbol{x}'\boldsymbol{\beta})] \qquad (5.6)$$

Ενώ για το μοντέλο probit η συνάρτηση πυκνότητας πιθανότητας $f(\cdot)$, δίνεται από την $\varphi(\cdot)$ την συνάρτηση πυκνότητας πιθανότητας της τυπικής κανονικής κατανομής:

$$\varphi(\boldsymbol{x}'\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi}}\, e^{-(\boldsymbol{x}'\boldsymbol{\beta})^2/2} \qquad (5.7)$$

Εφαρμόζοντας τυποποίηση σε όλες τις συνεχείς επεξηγηματικές μεταβλητές του μοντέλου και σταθεροποιώντας τες στις μέσες τιμές τους τιμές (προσέγγιση MEM), τότε με βάση τη συνθήκη (5.2), ο γραμμικός συνδυασμός $\boldsymbol{x}^{*\prime}\boldsymbol{\beta}$ ($\boldsymbol{x}^*$ οι τυποποιημένες συνεχείς επεξηγηματικές μεταβλητές), απλοποιείται και ισούται με τον εκτιμώμενο σταθερό όρος του μοντέλου β_0. Κατά συνέπεια η συνάρτηση $f(\boldsymbol{x}'\boldsymbol{\beta})$, απλοποιείται σε $f(\beta_0)$. Συνεπώς για τον υπολογισμό των ΜΕ (με βάση την προσέγγιση MEM) των συνεχών επεξηγηματικών μεταβλητών, αρκεί ο εκάστοτε αναλυτής να πολλαπλασιάσει το διάνυσμα των εκτιμώμενων συντελεστών $\boldsymbol{\beta}$, με την ποσότητα $f(\beta_0)$.

Σημειώνεται ό,τι επειδή και οι δύο κατανομές (logit & probit) είναι συμμετρικές ισχύει $f(-\beta_0) = f(\beta_0)$ και για -β_0 η προβλεπόμενη πιθανότητα ισούται με 1- $F(\beta_0)$.

Πιο συγκεκριμένα το ΜΕ μίας συνεχής επεξηγηματικής μεταβλητής, έστω x_j, υπολογίζεται ως εξής:

$$\mathrm{ME}_{x_j} = f(\beta_0)\beta_j \qquad (5.8)$$

Παρατηρείται, ό,τι όσο ο σταθερός όρος β_0 λαμβάνει μεγάλες θετικές τιμές, τόσο η ποσότητα $F(\beta_0)$ προσεγγίζει την μονάδα, κατά συνέπεια η ποσότητα $f(\beta_0)$ τείνει στο μήδεν, επομένως με βάση τον τύπο (5.5) και τα αντίστοιχα ΜΕ τείνουν στο μηδέν. Το ίδιο συμπέρασμα προκύπτει όταν η σταθερά β_0 λαμβάνει μεγάλες αρνητικές τιμές.

5.3.2 Οριακές Επιδράσεις τυποποιημένων διακριτών επεξηγηματικών μεταβλητών

Το ΜΕ για μία δυαδική διακριτή επεξηγηματική μεταβλητή x_j, η οποία λαμβάνει τιμές 0 και 1, υπολογίζεται ως εξής:

$$\mathrm{ME}_{x_j} = F\big(\boldsymbol{x}'\boldsymbol{\beta}|x_j = 1\big) - F\big(\boldsymbol{x}'\boldsymbol{\beta}|x_j = 0\big) \qquad (5.9)$$

Δεδομένων των τυποποιημένων επεξηγηματικών μεταβλητών που πραγματοποιήθηκαν παραπάνω, η ΜΕ μίας διακριτής επεξηγηματικής μεταβλητής (με βάση την προσέγγιση MEM), έστω x_j, υπολογίζεται ως εξής:

$$ME_{x_j} = F\big(\beta_0 + \beta_j\big) - F(\beta_0) \qquad (5.10)$$

Και σε αυτή την περίπτωση σημειώνεται, ό,τι όσο ο σταθερός όρος β_0 λαμβάνει μεγάλες θετικές τιμές, τόσο ο πρώτος όρος όσο και ο δεύτερος του τύπου (5.8), προσεγγίζουν την μονάδα, κατά συνέπεια η ποσότητα $f(\beta_0)$ τείνει στο μηδέν, επομένως και οι αντίστοιχες

τείνουν στο μηδέν. Το ίδιο συμπέρασμα προκύπτει όταν η σταθερά β_0 λαμβάνει μεγάλες αρνητικές τιμές.

Τέλος το ΜΕ μίας διακριτής επεξηγηματικής μεταβλητής για οποιαδήποτε τιμή των β_0 και β_j μπορεί να βρεθεί εύκολα, αντικαθιστώντας απλά την ποσότητα $\beta_0 + \beta_j$ με το β_0 ώστε να βρεθεί η τιμή της συνάρτησης $F(\beta_0 + \beta_j)$.

5.4 Απλοποίηση των διασπορών των ΜΕ τυποποιημένων μεταβλητών

Όπως ήδη είναι γνωστό από το προηγούμενο κεφάλαιο, ο ασυμπωτικός πίνακας διασπορών-συνδιασπορών $\widehat{Var}(\widehat{\mathbf{ME}})$ για ένα μοντέλο logit, δίνεται από τον τύπο (4.21):

$$\widehat{Var}(\widehat{\mathbf{ME}}) = l(\boldsymbol{x}'\boldsymbol{\beta})^2 \left[\boldsymbol{I} + \left(1 - 2\Lambda(\boldsymbol{x}'\widehat{\boldsymbol{\beta}})\right)(\boldsymbol{x}\widehat{\boldsymbol{\beta}}')\right] Var(\widehat{\boldsymbol{\beta}}) \left[\boldsymbol{I} + \left(1 - 2\Lambda(\boldsymbol{x}'\widehat{\boldsymbol{\beta}})\right)(\boldsymbol{x}\widehat{\boldsymbol{\beta}}')\right]$$

Ενώ σε ένα μοντέλο probit, με βάση τον τύπο (4.23), είναι:

$$\widehat{Var}(\widehat{\mathbf{ME}}) = \varphi(\boldsymbol{x}'\widehat{\boldsymbol{\beta}})^2 \left[\boldsymbol{I} - (\boldsymbol{x}'\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}\boldsymbol{x}')\right] Var(\widehat{\boldsymbol{\beta}}) \left[\boldsymbol{I} - (\boldsymbol{x}'\widehat{\boldsymbol{\beta}})(\boldsymbol{x}\widehat{\boldsymbol{\beta}}')\right]$$

5.4.1 Απλοποίηση των διασπορών των ΜΕ συνεχών τυποποιημένων μεταβλητών

Λαμβάνοντας υπόψη τις παραπάνω τυποποιήσεις, αποδεικνύεται ό,τι η ασυμπτωτική διασπορά του ΜΕ μιας συγκεκριμένης συνεχής επεξηγηματικής μεταβλητής, έστω x_j, σε ένα μοντέλο logit, απλοποιείται ως εξής:

$$\widehat{Var}\left(\widehat{\mathrm{ME}}_{x_j}\right) = l(\hat{\beta}_0)^2 \left(\sigma^2_{\hat{\beta}_j} + 2\left(1 - 2\Lambda(\hat{\beta}_0)\right)\hat{\beta}_j \sigma_{\hat{\beta}_j\hat{\beta}_0} + \left(1 - 2\Lambda(\hat{\beta}_0)\right)^2 \hat{\beta}_j{}^2 \sigma^2_{\hat{\beta}_0}\right) \qquad (5.13)$$

Ενώ σε ένα μοντέλο probit:

$$\widehat{Var}\left(\widehat{\mathrm{ME}}_{x_j}\right) = \varphi(\hat{\beta}_0)^2 \left(\sigma^2_{\hat{\beta}_j} + \hat{\beta}_0{}^2 \hat{\beta}_j{}^2 \sigma^2_{\hat{\beta}_0} - 2\hat{\beta}_0 \hat{\beta}_j \sigma_{\hat{\beta}_j\hat{\beta}_0}\right) \qquad (5.14)$$

Όπου

- $\sigma^2_{\hat{\beta}_0}$: είναι η ασυμπτωτική διασπορά των σταθερό όρου $\hat{\beta}_0$
- $\sigma^2_{\hat{\beta}_j}$: είναι η ασυμπτωτική διασπορά του εκτιμώμενου συντελεστή $\hat{\beta}_j$
- $\sigma_{\hat{\beta}_j\hat{\beta}_0}$: είναι η ασυμπτωτική συν διασπορά μεταξύ του σταθερού όρου $\hat{\beta}_0$ και του εκτιμώμενου συντελεστή $\hat{\beta}_j$

Ως εκ τούτου, η εκτίμηση της ασυμπτωτικής διασποράς του ΜΕ μιας συγκεκριμένης συνεχής επεξηγηματικής μεταβλητής, απλοποιείται αρκετά και είναι εύκολο να πραγματοποιηθεί.

5.4.2 Απλοποίηση των διασπορών των ΜΕ διακριτών τυποποιημένων μεταβλητών

Στην εργασία των Anderson & Newell (2003), ο αναλυτικός υπολογισμός για τις ασυμπτωτικές διασπορές των ΜΕ διακριτών τυποποιημένων μεταβλητών εμπεριέχεται από ένα λάθος, το οποίο οδηγεί σε λάθος αποτέλεσμα. Ωστόσο οι Carlevaro & Senegas (2006), με την σειρά τους έρχονται για να διορθώσουν το παραπάνω λάθος. Η διόρθωση των οποίων αναλύεται στην συνέχεια.

Για τον υπολογισμό της ασυμπτωτικής διασποράς του ΜΕ μιας διακριτής επεξηγηματικής μεταβλητής, αρχικά απαιτείται ο υπολογισμός του ασυμπωτικού πίνακα διασπορών-συνδιασπορών της προβλεπόμενης πιθανότητας, για τον οποίο χρησιμοποιώντας τη μέθοδο Δέλτα είναι πλέον γνωστό ό,τι δίνεται από τον τύπο (4.15):

$$Var\left(\hat{F}\left(\boldsymbol{x}'\widehat{\boldsymbol{\beta}}\right)\right) = \hat{f}\left(\boldsymbol{x}'\widehat{\boldsymbol{\beta}}\right)^2 \boldsymbol{x}' Var\left(\widehat{\boldsymbol{\beta}}\right)\boldsymbol{x} \qquad (4.15)$$

Όταν μία δίτιμη επεξηγηματική μεταβλητή έστω x_j, τυποποιηθεί με τέτοιο τρόπο ώστε η τιμή της να είναι μηδενική κατά το σημείο αναφοράς στο οποίο υπολογίζεται το ΜΕ, τότε το ΜΕ της θα υπολογίζεται ως εξής:

$$ME_{x_j} = F\left(\beta_0 + \beta_j\right) - F(\beta_0) \qquad (5.16)$$

Ένας συνεπής και ασυμπτωτικά φυσιολογικός εκτιμητής για το παραπάνω ΜΕ επιτυγχάνεται με την εισαγωγή εκτιμητών (όπως οι εκτιμητές μέγιστης πιθανότητας) $\hat{\beta}_0$ και $\hat{\beta}_j$ για τους συντελεστές $\hat{\beta}_0$ και $\hat{\beta}_j$ αντίστοιχα, οπότε ο τύπος (5.16), καταλήγει στον εξής:

$$\widehat{\mathrm{ME}}_{x_j} \equiv F\left(\hat{\beta}_0 + \hat{\beta}_j\right) - F\left(\hat{\beta}_0\right) \qquad (5.17)$$

Για να την εύρεση της ασυμπτωτικής διασποράς του εκτιμητή $\widehat{\mathrm{ME}}_{x_j}$ από τις ασυμπτωτικές διασπορές και συν διασπορές των εκτιμημένων συντελεστών $\hat{\beta}_0$ και $\hat{\beta}_j$ θα χρησιμοποιηθεί η μέθοδος Δέλτα (Greene, 2000). Όπως έχει προαναφερθεί, η συγκεκριμένη μέθοδος δηλώνει την ασυμπτωτική ισοδυναμία μεταξύ της κατανομής της $\widehat{\mathrm{ME}}_{x_j}$ και της γραμμικής επέκτασης της μη γραμμικής συνάρτησης των $\hat{\beta}_0$ και $\hat{\beta}_j$ γύρω από τις πραγματικές τιμές των $\hat{\beta}_0$ και $\hat{\beta}_j$.

Οπότε, ισχύει:

$$\widehat{\mathrm{ME}}_{x_j} \sim ME_{x_j} + \left(f\left(\hat{\beta}_0 + \hat{\beta}_j\right) - f\left(\hat{\beta}_0\right)\right)\left(\hat{\beta}_0 - \beta_0\right) + f\left(\beta_0 + \beta_j\right)\left(\hat{\beta}_j - \beta_j\right)$$

Το οποίο οδηγεί στην ακόλουθη ασυμπτωτική διασπορά:

$$\widehat{Var}\left(\widehat{\mathrm{ME}}_{x_j}\right) = \widehat{Var}\left(\left(f(\beta_0+\beta_j) - f(\beta_0)\right)(\hat{\beta}_0-\beta_0) + f(\beta_0+\beta_j)(\hat{\beta}_j-\beta_j)\right) =$$

$$= \left[f(\beta_0+\beta_j) - f(\beta_0)\right]^2\sigma^2_{\hat{\beta}_0} + f(\beta_0+\beta_j)^2\sigma^2_{\hat{\beta}_j} + 2\left[f(\beta_0+\beta_j) - f(\beta_0)\right]f(\beta_0+\beta_j)\sigma_{\hat{\beta}_j\hat{\beta}_0}$$

Συνεπώς, ο υπολογισμός της ασυμπτωτικής διασποράς του ΜΕ μιας συγκεκριμένης διακριτής επεξηγηματικής μεταβλητής, απλοποιείται αρκετά και είναι εύκολο να πραγματοποιηθεί.

5.4.3 Σχόλια

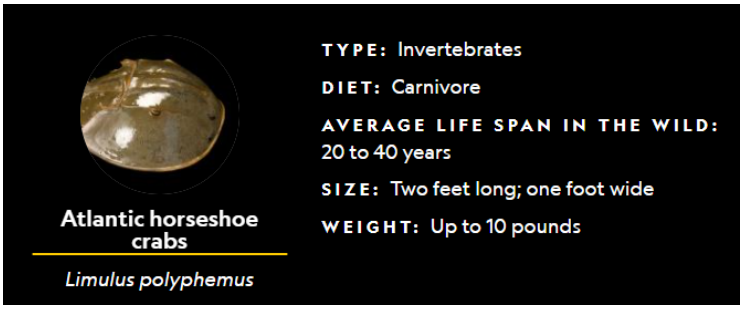
Αν και από τους παραπάνω τύπους που προέκυψαν δεν φαίνεται να προκύπτει μια δραματική απλοποίηση για τον υπολογισμό των τυπικών σφαλμάτων των ΜΕ, επισημαίνεται ό,τι χωρίς τις τυποποιήσεις των επεξηγηματικών μεταβλητών, αυτός ο υπολογισμός θα μπορούσε ενδεχομένως να περιλαμβάνει όλες τις καταχωρήσεις $m\,x\,m$ στον εκτιμώμενο πίνακας συν διασπορών, όπου m είναι ο αριθμός των εκτιμώμενων συντελεστών του μοντέλου. Ως εκ τούτου ύστερα από την απλοποίηση οι παραπάνω υπολογισμοί μπορούν να γίνουν και χειρωνακτικά από τον εκάστοτε αναλυτή.

Τέλος, η παραπάνω η εργασία των Anderson & Newell (2003), είχε πολύ σημαντική συνεισφορά για τον εκάστοτε αναλυτή εκείνης της εποχής που ήθελε να υπολογίσει τα ΜΕ (καθώς και τα τυπικά τους σφάλματα) και το μόνο που διέθετε από τα στατιστικά πακέτα ήταν οι εκτιμήσεις των συντελεστών του μοντέλου, ωστόσο τα στατιστικά πακέτα έχουν αναβαθμιστεί και βελτιωθεί πάρα πολύ από το 2003 και είναι πλέον ικανά με μία εντολή να δώσουν τα ΜΕ με τα τυπικά τους σφάλματα για οποιοδήποτε μοντέλο.

Κεφάλαιο 6

6.1 Σημαντικότητα των πεταλοειδών καβουριών

Το πεταλοειδές καβούρι **(**Limulus polyphemus**)** είναι αρθρόποδο που συσχετίζεται περισσότερο με αράχνες και σκορπιούς παρά με τα καβούρια. Συχνότερα συναντώνται στον Κόλπο του Μεξικού και στις ακτές της Βόρειας Αμερικής από την πλευρά του Ατλαντικού. Κάθε άνοιξη εκατοντάδες χιλιάδες καβούρια, καθοδηγούμενα από την πανσέληνο, ανεβαίνουν στις παραλίες στα μέσα του Ατλαντικού των ΗΠΑ για να γεννήσουν τα αυγά τους. Τα καβούρια αυτά είναι ευρέος γνωστά για το μπλε χρώμα του αίματός τους, το οποίο έχει φαρμακευτικές ιδιότητες. Για αυτό τον λόγο, όλες οι φαρμακευτικές εταιρείες σε όλο τον κόσμο βασίζονται σε αυτά τα καβούρια και είναι ένας κρίσιμος πόρος για την ασφάλεια των ανθρώπινων φαρμάκων (CARRIE, 2020).



An Atlantic horseshoe crab lies on the beach in Stone Harbor, New Jersey, not far from Delaware Bay.
PHOTOGRAPH BY JOEL SARTORE, NAT GEO IMAGE COLLECTION

Πιο συγκεκριμένα, το γαλάζιο αίμα αυτών των ζώων παρέχει τη μόνη γνωστή φυσική πηγή limulus amebocyte lysate, μια ουσία που ανιχνεύει την ενδοτοξίνη. Η ενδοτοξίνη είναι ένας τύπος βακτηριακής τοξίνης, η οποία ακόμη και εάν εισέλθει σε μικρές ποσότητες σε εμβόλια, ενέσιμα φάρμακα ή άλλα αποστειρωμένα φάρμακα, έχει μεγάλη πιθανότητα να είναι μοιραία για την ανθρώπινη ζωή. Κάθε χρόνο οι φαρμακευτικές εταιρείες συγκεντρώνουν μισό εκατομμύριο καβούρια, συλλέγουν το αίμα τους και τα επιστρέφουν στον ωκεανό, πολλά από τα οποία όμως δεν αντέχουν και πεθαίνουν. Αυτή η πρακτική, σε συνδυασμό με την υπερβολική συγκομιδή των καβουριών για αλιευτικό δόλωμα, προκάλεσε μείωση των ειδών στην περιοχή την τελευταία δεκαετία (CARRIE, 2020).

Horseshoe crabs are bled at the Charles River Laboratory in Charleston, South Carolina.
PHOTOGRAPH BY TIMOTHY FADEK, CORBIS/GETTY

Το 1990, οι βιολόγοι υπολόγισαν 1.24 εκατομμύρια καβούρια που γεννήθηκαν στον Κόλπο του Delaware, ένα βασικό σημείο ωοτοκίας και κορυφαίο σημείο συλλογής για τις εταιρείες. Μέχρι το 2002, ο αριθμός αυτός είχε μειωθεί σε 333.500. Τα τελευταία χρόνια, οι αριθμοί των καβουριών ωοτοκίας του Delaware Bay έχουν κυμανθεί περίπου στο ίδιο ποσό, με την έρευνα του 2019 να εκτιμάται σε περίπου 335.211, ενώ λόγω του Covid-19 δεν κατέστη δυνατή η μέτρηση τους για το 2020. Η σύλληψη καβουριών και η συλλογή του αίματος τους είναι χρονοβόρα και η τιμή του υπολογίζεται περίπου με 15.000$ το λίτρο. Εξαιτίας της μείωσης του πληθυσμού τους, οι επιστήμονες, το 2016 κατέφυγαν στην δημιουργία μίας νέας συνθετικής εναλλακτικής λύσης, η οποία ένας ανά συνδυασμένος παράγοντας C (rFC). Αυτός εγκρίθηκε ως εναλλακτική λύση στην Ευρώπη και αρκετές αμερικάνικες φαρμακευτικές εταιρείες άρχισαν να το χρησιμοποιούν. Ωστόσο, στης 1η Ιουνίου 2020, η Αμερικανική Φαρμακοποιία, η οποία θέτει τα επιστημονικά πρότυπα για τα φάρμακα και άλλα προϊόντα στις ΗΠΑ, αρνήθηκε να τοποθετήσει το rFC σε ίση βάση με το προϊόν των καβουριών, ισχυριζόμενη ό,τι η ασφάλειά του εξακολουθεί να είναι μη αποδεδειγμένη (CARRIE, 2020).

Στις μέρες του Covid-19, τα πεταλοειδές καβούρια, διαδραματίζουν σημαντικό ρόλο, προκείμενου να επιτευχθεί η ορθή αξιολόγηση των επιπέδων ενδοτοξίνης, για να διαπιστωθεί εάν τα νέα εμβόλια Covid-19 είναι ασφαλή. Κατά συνέπεια συμπεραίνει κανείς ό,τι για την παρασκευή μαζικού πλήθους εμβολίων θα χρειαστεί να συλλεχθούν μεγάλο ποσοστό αυτού του είδους καβουριών, το οποίο θα οδηγήσει και άλλο στην μείωση του πληθυσμού τους και κατ' επέκταση στην αποδυνάμωση του οικοσυστήματος.

6.2 Περιγραφή των δεδομένων

Από την προηγούμενη παράγραφο, έγινε κατανοητό πόσο ωφέλημα για την ανθρωπότητα είναι τα πεταλοειδή καβούρια και γενικότερα για το οικοσύστημα. Για αυτό τον λόγο είναι πολύ σημαντικό να πραγματοποιηθούν μελέτες με σκοπό να βοηθήσουν στην διατήρηση, αλλά και στην αύξηση του πληθυσμού τους. Πρωτοπόρος σε αυτές τις μελέτες υπήρξε η Brockmann, η οποία το 1996 σύλλεξε ορισμένες μετρήσεις με βάση τα θηλυκά καβούρια (βλ. Πίνακα 6.1), προκείμενου να μελετήσει τους παράγοντες που επηρεάζουν το πλήθος των αρσενικών καβουριών (τα οποία ονομάζονται δορυφόροι), που έχει κάθε θηλυκό στην φωλιά του (Brockmann, 1996). Τα δεδομένα του Πίνακα 6.1 αποτελούνται από πέντε μεταβλητές για 173 μετρήσεις, οι οποίες παρουσιάζονται λεπτομερώς εν συνεχεία:

1. **Satellites**: Είναι μία συνεχής μεταβλητή, η οποία εκφράζει τον αριθμό των αρσενικών (δορυφόρων) καβουριών που είχε στην φωλιά του το εκάστοτε θηλυκό καβούρι.

Σημειώνεται:

- Η συγκεκριμένη μεταβλητή στο dataset της R θα αναφέρεται ως "psi" και θα χρησιμοποιηθεί ως η μεταβλητή απόκρισης, για την παρασκευή το μοντέλων Poisson και ZIP.
- Αντιθέτως, κατά την παρασκευή των μοντέλων logit, probit & cloglog, θα χρειαστεί η μεταβλητή psi(Satellites) από συνεχής να μετασχηματιστεί σε δίτιμη. Δηλαδή:

$$\begin{cases} y = 0, & \text{όταν } Satellites = 0 \\ y = 1, & \text{όταν } Satellites \geq 1 \end{cases}$$

2. **Width**: Είναι μία συνεχής μεταβλητή, η οποία εκφράζει το πλάτος του κάθε θηλυκού καβουριού, μετρημένο σε cm

3. **Weight**: Είναι μία συνεχής μεταβλητή, η οποία εκφράζει το βάρους του κάθε θηλυκού καβουριού, μετρημένο σε kg

4. **Color**: Είναι μία διακριτή μεταβλητή, η οποία εκφράζει το χρώμα του εκάστοτε θηλυκού καβουριού, για την οποία ισχύει:
   - c1= 1, όταν Color= light
   - c2= 2, όταν Color= medium
   - c3= 3, όταν Color= dark
   - c4= 4, όταν Color= darker

5. **Spine**: Είναι μία διακριτή μεταβλητή, η οποία εκφράζει την κατάσταση της σπονδυλικής στήλης-ράχης του κάθε θηλυκού καβουριού, για την οποία ισχύει:
   - s1= 1, όταν Spine= good
   - s2= 2, όταν Spine= middle
   - s3= 3, όταν Spine= bad

Πίνακα 6.1: Αριθμός των δορυφόρων ανάλογα το χρώμα του θηλυκού (Color-C), την κατάσταση της σπονδυλικής τους στήλες (Spine-S), το πλάτος (Width-W) και το Βάρος (Weight-Wt).

C	S	W	Wt	Sa	C	S	W	Wt	Sa	C	S	W	Wt	Sa	C	S	W	Wt	Sa
2	3	28.3	3.05	8	3	3	22.5	1.55	0	1	1	26.0	2.30	9	3	3	24.8	2.10	0
3	3	26.0	2.60	4	2	3	23.8	2.10	0	3	2	24.7	1.90	0	2	1	23.7	1.95	0
3	3	25.6	2.15	0	3	3	24.3	2.15	0	2	3	25.8	2.65	0	2	3	28.2	3.05	11
4	2	21.0	1.85	0	2	1	26.0	2.30	14	1	1	27.1	2.95	8	2	3	25.2	2.00	1
2	3	29.0	3.00	1	4	3	24.7	2.20	0	2	3	27.4	2.70	5	2	2	23.2	1.95	4
1	2	25.0	2.30	3	2	1	22.5	1.60	1	3	3	26.7	2.60	2	4	3	25.8	2.00	3
4	3	26.2	1.30	0	2	3	28.7	3.15	3	2	1	26.8	2.70	5	4	3	27.5	2.60	0
2	3	24.9	2.10	0	1	1	29.3	3.20	4	1	3	25.8	2.60	0	2	2	25.7	2.00	0
2	1	25.7	2.00	8	2	1	26.7	2.70	5	4	3	23.7	1.85	0	2	3	26.8	2.65	0
2	3	27.5	3.15	6	4	3	23.4	1.90	0	2	3	27.9	2.80	6	3	3	27.5	3.10	3
1	1	26.1	2.80	5	1	1	27.7	2.50	6	2	1	30.0	3.30	5	3	1	28.5	3.25	9
3	3	28.9	2.80	4	2	3	28.2	2.60	6	2	3	25.0	2.10	4	2	3	28.5	3.00	3
2	1	30.3	3.60	3	4	3	24.7	2.10	5	2	3	27.7	2.90	5	1	1	27.4	2.70	6
2	3	22.9	1.60	4	2	1	25.7	2.00	5	2	3	28.3	3.00	15	2	3	27.2	2.70	3
3	3	26.2	2.30	3	2	1	27.8	2.75	0	4	3	25.5	2.25	0	3	3	27.1	2.55	0
3	3	24.5	2.05	5	3	1	27.0	2.45	3	2	3	26.0	2.15	5	2	3	28.0	2.80	1
2	3	30.0	3.05	8	2	3	29.0	3.20	10	2	3	26.2	2.40	0	2	1	26.5	1.30	0
2	3	26.2	2.40	3	3	3	25.6	2.80	7	3	3	23.0	1.65	1	3	3	23.0	1.80	0
2	3	25.4	2.25	6	3	3	24.2	1.90	0	2	2	22.9	1.60	0	3	2	26.0	2.20	3
2	3	25.4	2.25	4	3	3	25.7	1.20	0	2	3	25.1	2.10	5	3	2	24.5	2.25	0
4	3	27.5	2.90	0	3	3	23.1	1.65	0	3	1	25.9	2.55	4	2	3	25.8	2.30	0
4	3	27.0	2.25	3	2	3	28.5	3.05	0	4	1	25.5	2.75	0	4	3	23.5	1.90	0
2	2	24.0	1.70	0	2	1	29.7	3.85	5	2	1	26.8	2.55	0	4	3	26.7	2.45	0
2	1	28.7	3.20	0	3	3	23.1	1.55	0	2	1	29.0	2.80	1	3	3	25.5	2.25	0

3	3	26.5	1.97	1	3	3	24.5	2.20	1	3	3	28.5	3.00	1	2	3	28.2	2.87	1
2	3	24.5	1.60	1	2	3	27.5	2.55	1	2	2	24.7	2.55	4	2	1	25.2	2.00	1
3	3	27.3	2.90	1	2	3	26.3	2.40	1	2	3	29.0	3.10	1	2	3	25.3	1.90	2
2	3	26.5	2.30	4	2	3	27.8	3.25	3	2	3	27.0	2.50	6	3	3	25.7	2.10	0
2	3	25.0	2.10	2	2	3	31.9	3.33	2	4	3	23.7	1.80	0	4	3	29.3	3.23	12
3	3	22.0	1.40	0	2	3	25.0	2.40	5	3	3	27.0	2.50	6	3	3	23.8	1.80	6
1	1	30.2	3.28	2	3	3	26.2	2.22	0	2	3	24.2	1.65	2	2	3	27.4	2.90	3
2	2	25.4	2.30	0	3	3	28.4	3.20	3	4	3	22.5	1.47	4	2	3	26.2	2.02	2
2	1	24.9	2.30	6	1	2	24.5	1.95	6	2	3	25.1	1.80	0	2	1	28.0	2.90	4
4	3	25.8	2.25	10	2	3	27.9	3.05	7	2	3	24.9	2.20	0	2	1	28.4	3.10	5
3	3	27.2	2.40	5	2	2	25.0	2.25	6	2	3	27.5	2.63	6	2	1	33.5	5.20	7
2	3	30.5	3.32	3	3	3	29.0	2.92	3	2	1	24.3	2.00	0	2	3	25.8	2.40	0
4	3	25.0	2.10	8	2	1	31.7	3.73	4	2	3	29.5	3.02	4	3	3	24.0	1.90	10
2	3	30.0	3.00	9	2	3	27.6	2.85	4	2	3	26.2	2.30	0	2	1	23.1	2.00	0
2	1	22.9	1.60	0	4	3	24.5	1.90	0	2	3	24.7	1.95	4	2	3	28.3	3.20	0
2	3	23.9	1.85	2	3	3	23.8	1.80	0	3	2	29.8	3.50	4	2	3	26.5	2.35	4
2	3	26.0	2.28	3	2	3	28.2	3.05	8	4	3	25.7	2.15	0	2	3	26.5	2.75	7
2	3	25.8	2.20	0	3	3	24.1	1.80	0	3	3	26.2	2.17	2	3	3	26.1	2.75	3
3	3	29.0	3.28	4	1	1	28.0	2.62	0	4	3	27.0	2.63	0	2	2	24.5	2.00	0
1	1	26.5	2.35	0															

Data provided by Dr. Jane Brockmann, Zoology Department, University of Florida; study described in Ethology, 102: 1–21, 1996.

6.2.1 Μεθοδολογία

Με βάση την συγκεκριμένη βάση δεδομένων και με εργαλείο την θεωρεία των ΜΕ, όπως έχει παρουσιαστεί στα προηγούμενα κεφάλαια, θα ερμηνεύσουμε με την σειρά μας του παράγοντες που επηρεάζουν το πλήθος των δορυφόρων, που έχει κάθε θηλυκό στην φωλιά του. Αυτό θα επιτευχθεί κατασκευάζοντας 5 γενικευμένα γραμμικά μοντέλα, 3 για την περίπτωση της δίτιμης μεταβλητής απόκρισης y και 2 για την περίπτωση της μεταβλητής απόκρισης psi, για αριθμημένα δεδομένα. Σημειώνεται ό,τι όλοι οι στατιστική έλεγχοι θα πραγματοποιηθούν σε επίπεδο σημαντικότητας 5%.

Στην συνέχεια για κάθε μοντέλο που θα προκύψει θα υπολογιστούν τα ΜΕ, όπως και τα τυπικά τους σφάλματα, με βάση τις προσεγγίσεις MEM και AME. Αυτό θα πραγματοποιηθεί με την βοήθεια της γλώσσας προγραμματισμού R, με τους έξης τρόπους:

1) Για τον υπολογισμό των ΜΕ (και τυπικών σφαλμάτων) με βάση την προσέγγιση MEM, θα χρησιμοποιηθεί το πακέτο εντολών της R ‘mfx’. Το οποίο με τις κατάλληλες εντολές μπορεί να μας “δώσει” τα ΜΕ με βάση την προσέγγιση MEM, καθώς και τα τυπικά σφάλματα με βάση τη μέθοδο Δέλτα. Το μειονέκτημα της όμως προκύπτει στο γεγονός ό,τι είναι προγραμματισμένη να υπολογίζει τα ΜΕ για συγκεκριμένα μοντέλα (logit, probit & Poisson) (A. Fernihough, 2019). Κατά συνέπεια για τα μοντέλα cloglog και ZIP, θα αναπτυχθεί κώδικας με σκοπό τον υπολογισμό των συγκεκριμένων ποσοτήτων ενδιαφέροντας.

2) Ενώ, για τον υπολογισμό των ΜΕ (και τυπικών σφαλμάτων) με βάση την προσέγγιση AME, θα χρησιμοποιηθεί το πακέτο εντολών της R ‘margins’. Το οποίο με τις κατάλληλες εντολές μπορεί να μας “δώσει” τα ΜΕ με βάση την προσέγγιση AME, καθώς και τα τυπικά σφάλματα με βάση τη μέθοδο Δέλτα. Το μειονέκτημα αυτής της

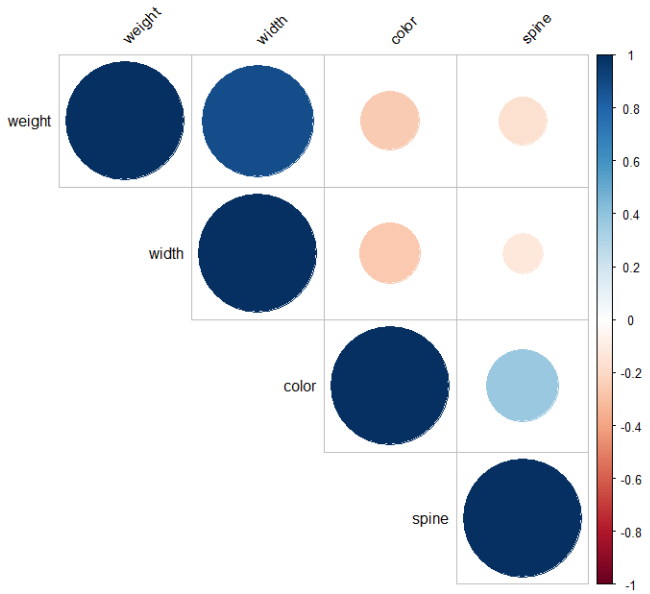
εντολής είναι ό,τι δεν έχει προγραμματιστεί σωστά για την περίπτωση υπολογισμού των ΜΕ και των τυπικών τους σφαλμάτων για το μοντέλο ZIP (J. Leeper, 2021). Συνεπώς θα αναπτυχθεί και πάλι κώδικας με σκοπό των υπολογισμό των ποσοτήτων ενδιαφέροντος με βάση την θεωρία.

Σημειώνεται, ό,τι παρόλο που στις περισσότερες περιπτώσεις, οι παραπάνω απλές εντολές μπορούν να "δώσουν" άμεσα τα ΜΕ και τα τυπικά τους σφάλματα, έχει κατασκευαστεί αναλυτικός κώδικας (βλ. Παράρτημα) που τις επαληθεύει ή χρησιμεύει στην εύρεση των ποσοτήτων ενδιαφέροντος, όταν δεν καθίσταται δυνατή από τα παραπάνω πακέτα εντολών.

Τέλος, από τα 3 πρώτα μοντέλα (logit, probit & cloglog), θα επιλεχθεί εκείνο που προσαρμόζεται καλύτερο με βάση το κριτήριο προσαρμογής AIC και στην συνέχεια θα ερμηνευθούν τα ΜΕ του. Αντίστοιχα, παρόμοια διαδικασία θα ακολουθηθεί και για τα άλλα 2 μοντέλα (Poisson & ZIP).

Έλεγχος πολυσυγγραμμικότητας

Ελέγχοντας την πιθανή ύπαρξη πολυσυγγραμμικότητας μεταξύ των επεξηγηματικών μεταβλητών, προκύπτει ό,τι η επεξηγηματική μεταβλητή Width και Weight, έχουν υψηλή συσχέτιση, συγκεκριμένα 88%, όποτε προκειμένου να αποφευχθούν λανθασμένα στατιστικά αποτελέσματα, δεν χρειάζονται να βρίσκονται και οι δύο στο μοντέλο πρόβλεψης, αφού άλλωστε η μία περιέχει σημαντικό ποσοστό πληροφορίας της άλλης. Συνεπώς κρατάμε την μεταβλητή Width.



6.3 Προσαρμογή γενικευμένων γραμμικών μοντέλων για δίτιμη μεταβλητή απόκρισης

Για την προσαρμογή των μοντέλων logit, probit & cloglog, η μεταβλητή απόκρισης των μοντέλων που θα χρησιμοποιηθεί θα είναι η μεταβλητή y, η πιθανότητα της οποίας, $\pi = P(y = 1)$, εκφράζει την πιθανότητα να υπάρχει τουλάχιστον ένας δορυφόρος στην φωλιά του θηλυκού.

6.3.1 Μοντέλο logit

Συνεχίζοντας θα προσαρμόσουμε ένα μοντέλο logit, το οποίο θα αποτελείται από την συνεχής μεταβλητή απόκρισης Width και τους παράγοντες Color και Spine.

```
Call:
glm(formula = y ~ width + factor(spine) + factor(color), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1206  -0.9724   0.5076   0.8750   2.1158

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -11.09953    2.97706  -3.728 0.000193 ***
width            0.45624    0.10779   4.233 2.31e-05 ***
factor(spine)2  -0.05782    0.70308  -0.082 0.934453
factor(spine)3   0.37703    0.50191   0.751 0.452540
factor(color)2  -0.14340    0.77838  -0.184 0.853830
factor(color)3  -0.52405    0.84685  -0.619 0.536030
factor(color)4  -1.66833    0.93285  -1.788 0.073706 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 186.61  on 166  degrees of freedom
AIC: 200.61

Number of Fisher Scoring iterations: 4
```

Από το p-values για τον έλεγχο Wald, υποψιαζόμαστε ό,τι η μεταβλητή Spine φαίνεται να μην είναι στατιστικά σημαντική για το μοντέλο μας. Αυτό θα το επιβεβαιώσουμε τρέχοντας ένα καινούργιο μοντέλο χωρίς την Spine και μέσω του ελέγχου γενικευμένου λόγου πιθανοφανειών θα συγκινούμε τα δύο μοντέλα.

```
Analysis of Deviance Table

Model 1: y ~ width + factor(color)
Model 2: y ~ width + factor(spine) + factor(color)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       168     187.46
2       166     186.61  2  0.84513   0.6554
```

Από τον παραπάνω έλεγχο παρατηρείται ό,τι το μοντέλο με τις μεταβλητές Width και Color είναι προτιμότερο.

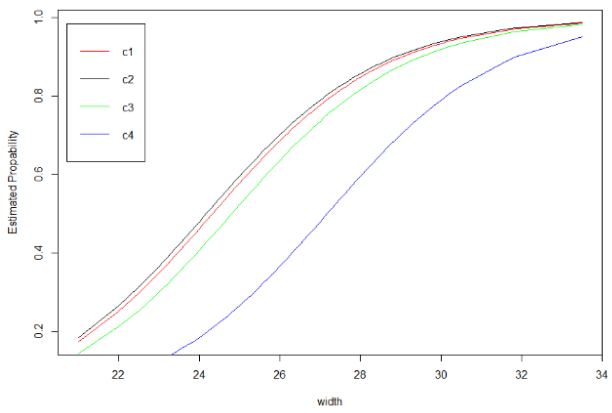
```
Call:
glm(formula = y ~ width + factor(color), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1124  -0.9848   0.5243   0.8513   2.1413

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -11.38519    2.87346  -3.962 7.43e-05 ***
width            0.46796    0.10554   4.434 9.26e-06 ***
factor(color)2   0.07242    0.73989   0.098    0.922
factor(color)3  -0.22380    0.77708  -0.288    0.773
factor(color)4  -1.32992    0.85252  -1.560    0.119
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.46  on 168  degrees of freedom
AIC: 197.46

Number of Fisher Scoring iterations: 4
```

Συνεχίζοντας από το νέο πλέον μοντέλο βλέπουμε ό,τι και η μεταβλητή Color ενδέχεται να μην στατιστικά σημαντική. Οπότε ακολουθώντας πάλι την προηγούμενη διαδικασία θα συγκρίνουμε το παραπάνω μοντέλο με ένα καινούργιο μοντέλο που αποτελείται μόνο από την width.

```
Analysis of Deviance Table

Model 1: y ~ width
Model 2: y ~ width + factor(color)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       171     194.45
2       168     187.46  3   6.9956  0.07204 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ενώ το p-value του ελέγχου του γενικευμένου λόγου πιθανοφανειών, μας ενημερώνει ό,τι καλύτερο είναι το μοντέλο που έχει μόνο την Width, η τιμή του είναι αρκετά κοντά στο 5%, οπότε η επιλογή του καλύτερου μοντέλου χρήζετε περαιτέρω ανάλυσης.

Σχεδιάζοντας την συνάρτηση π συναρτήσει της Width για τις 4 κατηγορίες της Color, όπως φαίνεται και στο παρακάτω διάγραμμα, η κατηγορία c4 της Color διαφέρει από τις άλλες 3.



Οπότε, θα κατασκευάσουμε ένα καινούργιο μοντέλο, το οποίο θα αποτελείται από την μεταβλητή Width και από την κατηγορία c4 της Color και στην συνέχεια θα το συγκρίνουμε με το μοντέλο που αποτελείται από την μεταβλητή Width.

```
Analysis of Deviance Table

Model 1: y ~ width
Model 2: y ~ width + factor(c4)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       171     194.45
2       170     187.96  1   6.4948  0.01082 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Συνεπώς, καταλήγουμε ό,τι το καλύτερο μοντέλο είναι αυτό το οποίο αποτελείται από την μεταβλητή Width και την κατηγορία c4.

```
Call:
glm(formula = y ~ width + factor(c4), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0821  -0.9932   0.5274   0.8606   2.1553

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.6790     2.6925  -4.338 1.44e-05 ***
width         0.4782     0.1041   4.592 4.39e-06 ***
factor(c4)1  -1.3005     0.5259  -2.473   0.0134 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.96  on 170  degrees of freedom
AIC: 193.96

Number of Fisher Scoring iterations: 4
```

Δηλαδή, το μοντέλο:

$$logit(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 Width + \hat{\beta}_2 c4$$

$$\Leftrightarrow$$

$$logit(\hat{\pi}) = -11.6790 + 0.4782\, Width - 1.3005 c4$$

ΜΕ και τυπικά σφάλματα των επεξηγηματικών μεταβλητών του μοντέλου logit

Στην συνέχεια υπολογίζονται τα ΜΕ και τα τυπικά τους σφάλματα για τις επεξηγηματικές μεταβλητές Width και c4, με βάση τις προσεγγίσεις MEM και AME.

Προσέγγιση MEM

```
  Variables     MEM     SD         z            p  Lower   Upper
1     Width  0.1049 0.0219  4.789954 1.668193e-06  0.062  0.1478
2        c4 -0.3098 0.1241 -2.496374 1.254703e-02 -0.553 -0.0666
```

Προσέγγιση AME

```
  Variables     AME     SD         z             p   Lower   Upper
1     Width  0.0875 0.0147  5.952381 2.642694e-09  0.0587  0.1163
2        c4 -0.2614 0.1057 -2.473037 1.339703e-02 -0.4686 -0.0542
```

Όπως παρατηρείται από τα παραπάνω αποτελέσματα, τα ΜΕ είναι στατιστικά σημαντικά. Επιπλέον, είναι φανερό ό,τι ανάλογα την προσέγγιση που χρησιμοποιείται, υπάρχουν διαφορές μεταξύ τους. Αυτό όπως έχει αναφερθεί και στο κεφάλαιο 3, ενδεχομένως να οφείλεται στο γεγονός ό,τι το μέγεθος τους δείγματος είναι αρκετά μικρό.

6.3.2 Μοντέλο probit

Για την εύρεση του μοντέλο probit που προσαρμόζεται καλύτερα στα δεδομένα, ακολουθείται ίδια διαδικασία όπως και με το μοντέλο logit, καταλήγοντας ό,τι το καλύτερο μοντέλο είναι αυτό που αποτελείται από τις μεταβλητές Width και c4.

```
Call:
glm(formula = y ~ width + factor(c4), family = binomial(link = probit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1042  -1.0126   0.5195   0.8674   2.1693

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.98838    1.54195  -4.532 5.84e-06 ***
width        0.28637    0.05924   4.834 1.34e-06 ***
factor(c4)1 -0.76494    0.31341  -2.441   0.0147 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.72  on 170  degrees of freedom
AIC: 193.72

Number of Fisher Scoring iterations: 5
```

Δηλαδή, το μοντέλο:

$$probit(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 Width + \hat{\beta}_2 c4$$

$$\Leftrightarrow$$

$$probit(\hat{\pi}) = -6.98838 + 0.28637\, Width - 0.76494 c4$$

ΜΕ και τυπικά σφάλματα των επεξηγηματικών μεταβλητών του μοντέλου probit

Στην συνέχεια υπολογίζονται τα ΜΕ και τα τυπικά τους σφάλματα για τις επεξηγηματικές μεταβλητές Width και c4, με βάση τις προσεγγίσεις MEM και AME.

Προσέγγιση MEM

```
  Variables     MEM     SD         z            p   Lower   Upper
1     Width  0.1035 0.0209  4.952153 7.339682e-07  0.0625  0.1445
2        c4 -0.2943 0.1203 -2.446384 1.442972e-02 -0.5301 -0.0585
```

Προσέγγιση AME

```
  Variables     MEM     SD         z            p   Lower   Upper
1     Width  0.0878 0.0144  6.097222 1.079274e-09  0.0596  0.1160
2        c4 -0.2553 0.1056 -2.417614 1.562266e-02 -0.4623 -0.0483
```

Όμοια και σε αυτή την περίπτωση, τα ΜΕ είναι στατιστικά σημαντικά. Επιπλέον, διαφέρουν μεταξύ τους αναλόγως την προσέγγιση.

6.3.3 Μοντέλο cloglog

Όμοια με τα προηγούμενα δύο μοντέλα, το καλύτερο μοντέλο cloglog είναι αυτό που αποτελείται από τις μεταβλητές Width και c4.

```
Call:
glm(formula = y ~ width + factor(c4), family = binomial(link = cloglog))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1944  -0.9571   0.4839   0.8919   1.9790

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.62635    1.61150  -4.732 2.22e-06 ***
width        0.29634    0.06072   4.881 1.06e-06 ***
factor(c4)1 -0.92458    0.40168  -2.302   0.0213 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 186.45  on 170  degrees of freedom
AIC: 192.45

Number of Fisher Scoring iterations: 5
```

Δηλαδή, το μοντέλο:

$$cloglog(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 Width + \hat{\beta}_2 c4$$

$$\Leftrightarrow$$

$$cloglog(\hat{\pi}) = -7.62635 + 0.29634 Width - 0.92458 c4$$

ΜΕ και τυπικά σφάλματα των επεξηγηματικών μεταβλητών του μοντέλου cloglog

Στην συνέχεια υπολογίζονται τα ΜΕ και τα τυπικά τους σφάλματα για τις επεξηγηματικές μεταβλητές Width και c4, με βάση τις προσεγγίσεις MEM και AME.

Προσέγγιση MEM

```
  Variables     MEM     SD         z            p   Lower   Upper
1     Width  0.1089 0.0224  4.861607 1.164365e-06  0.0650  0.1528
2        c4 -0.3190 0.1203 -2.651704 8.008670e-03 -0.5548 -0.0832
```

Προσέγγιση AME

```
  Variables     AME     SD         z            p   Lower   Upper
1     Width  0.0869 0.0133  6.533835 6.410675e-11  0.0608  0.1130
2        c4 -0.2792 0.1131 -2.468612 1.356383e-02 -0.5009 -0.0575
```

Όμοια και σε αυτή την περίπτωση, τα ΜΕ είναι στατιστικά σημαντικά. Επιπλέον, διαφέρουν μεταξύ τους αναλόγως την προσέγγιση.

6.3.4 Εύρεση καλύτερου μοντέλου και ερμηνεία των ΜΕ των επεξηγηματικών μεταβλητών

Προκειμένου να επιλεχθεί ποιο από τα παραπάνω τρία μοντέλα προσαρμόζεται καλύτερα στα δεδομένα, θα χρησιμοποιηθεί το κριτήριο προσαρμογής AIC.

```
  AIC.logit. AIC.probit. AIC.cloglog.
1   193.9579    193.7202     192.4524
```

Σύμφωνα με τα αποτελέσματα, καλύτερο μοντέλο επιλέγεται το μοντέλο cloglog

Δηλαδή:

$$cloglog(\hat{\pi}) = -7.62635 + 0.29634Width - 0.92458c4$$

Ερμηνείες των ΜΕ βάση της προσέγγισης MEM του μοντέλου cloglog

- **Width**: Η οριακής της επίδραση ισούται με 0.1089 (0.0224), η ερμηνεία της οποίας σημαίνει ό,τι η διαφορά της πιθανότητας να βρίσκεται τουλάχιστον ένας δορυφόρος στην φωλιά του θηλυκού καβουριού, όταν το θηλυκό καβούρι έχει το μέσο παρατηρούμενο χρώμα της κατηγορίας c4, δηλαδή 0.127, με πλάτος 26.3+ δ cm (όπου δ μία πολύ μικρή μεταβολή) και της πιθανότητας όταν το θηλυκό καβούρι έχει το μέσο παρατηρούμενο πλάτος (26.3 cm), εκτιμάται με 0.1089*δ. Για παράδειγμα όταν το θηλυκό έχει το μέσο χρώμα της κατηγορίας c4 και πλάτος 26.8 cm, η πιθανότητα στην φωλιά του να υπάρχει τουλάχιστον ένα αρσενικό υπολογίζεται περίπου με

0.1089*(26.8-26.3)= 0.05445 μεγαλύτερη από την πιθανότητα το θηλυκό καβούρι να έχει το μέσο παρατηρούμενο πλάτος (26.3 cm).

Σημειώνεται, ό,τι το μέσο παρατηρούμενο χρώμα της κατηγορίας c4, που ισούται με 0.127, δεν έχει κάποια ερμηνεία ως προς το ποια κατηγορία χρώματος ανήκει, αλλά υπολογίζεται προκείμενου να υπολογιστούν τα ΜΕ χρησιμοποιώντας την προσέγγιση MEM. Αυτό σημαίνει ό,τι, η δήλωση "όταν το θηλυκό καβούρι έχει το μέσο χρώμα της κατηγορίας c4", είναι ανακριβής. Αυτός είναι και ο κύριος λόγος, όπως διατυπώθηκε και στο κεφάλαιο 3, όπου η προσέγγιση MEM ενδέχεται να είναι υποδεέστερη της AME.

- **c4**: Η οριακή της επίδραση ισούται με -0.3190 (0.1203). Αυτό σημαίνει ό,τι για ένα θηλυκό καβούρι το οποίο έχει το μέσο παρατηρούμενο πλάτος (26.3 cm), η πιθανότητα να εμφανιστεί τουλάχιστον ένας δορυφόρος στην φωλιά του είναι μικρότερη κατά 0.3190 όταν έχει χρώμα darker από όταν έχει χρώμα light, medium ή dark.

Από τα παραπάνω αποτελέσματα μπορεί να συμπεράνει κανείς ό,τι τα αρσενικά καβούρια έχουν μία τάση να προτιμάνε τα ποιο μεγαλόσωμα και ανοιχτόχρωμα θηλυκά καβούρια.

Ερμηνείες των ΜΕ βάση την προσέγγισης AME του μοντέλου cloglog

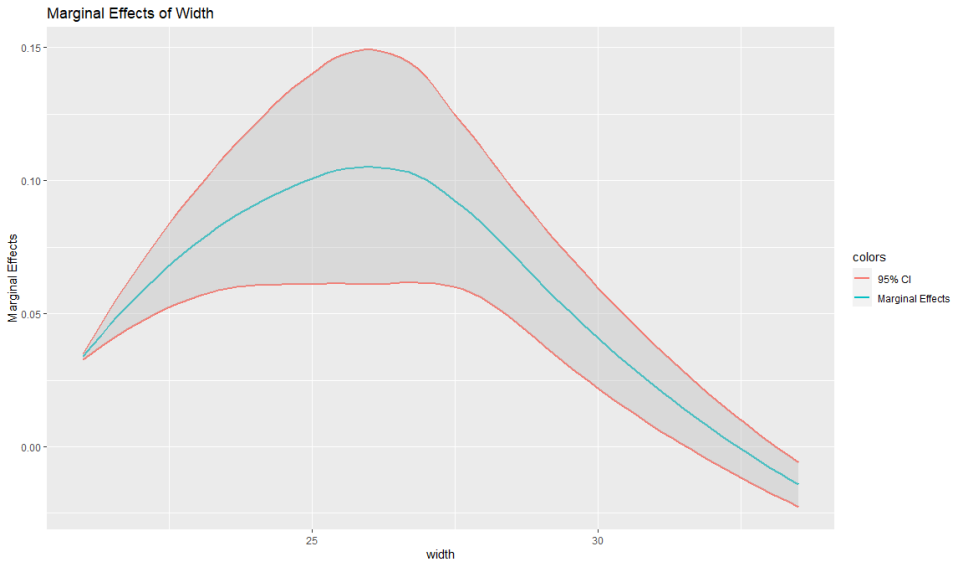
- **Width**: Η μέση οριακής της επίδραση ισούται με 0.0869 (0.013). Για την εύρεση της, υπολογίστηκαν τα ΜΕ για κάθε παρατήρηση του μοντέλου και στην συνέχεια υπολογίστηκε η μέση τιμή αυτών. Αυτό σημαίνει ό,τι όταν το πλάτος του θηλυκού, για κάθε παρατήρηση αυξηθεί κατά δ, δηλαδή $Width_i$+δ, τότε η πιθανότητα να υπάρχει τουλάχιστον ένας δορυφόρος στην φωλιά του, ανεξαρτήτως αν έχει χρώμα darker ή όχι, εκτιμάται περίπου κατά μέσο όρο με 0.0869*δ μεγαλύτερη από όταν το θηλυκό είχε το αρχικό πλάτος για κάθε παρατήρηση $Width_i$. Για παράδειγμα, αν το πλάτος κάθε παρατήρησης αυξανότανε κατά 1 cm, δηλαδή $Width_i$+1, τότε η πιθανότητα να υπάρχει τουλάχιστον ένας δορυφόρος στην φωλιά του θηλυκού είναι κατά μέσο όρο 0.0869*1= 0.0869 μεγαλύτερη από όταν το θηλυκό είχε το αρχικό πλάτος του, $Width_i$, για κάθε παρατήρηση.

- **c4**: Η μέση οριακή της επίδραση ισούται με -0.2792 (0.1131). Αυτό σημαίνει ό,τι η πιθανότητα να υπάρχει τουλάχιστον ένα αρσενικό στην φωλιά ενός θηλυκού το οποίο έχει χρώμα darker εκτιμάται κατά μέσο όρο 0.2792 μικρότερη από όταν το θηλυκό καβούρι έχει χώμα light, medium ή dark. Αυτό αποκρύπτει πολλές παραλλαγές σε ατομικό επίπεδο (π.χ. οι χρωματικές διαφορές ποικίλουν ανάλογα το πλάτος), ωστόσο εξακολουθεί να μας δίνει μία γενική ιδέα για το πόσο μεγάλες είναι οι διαφορές.

Γραφική απεικόνιση των ΜΕ και των 95% διαστημάτων εμπιστοσύνης

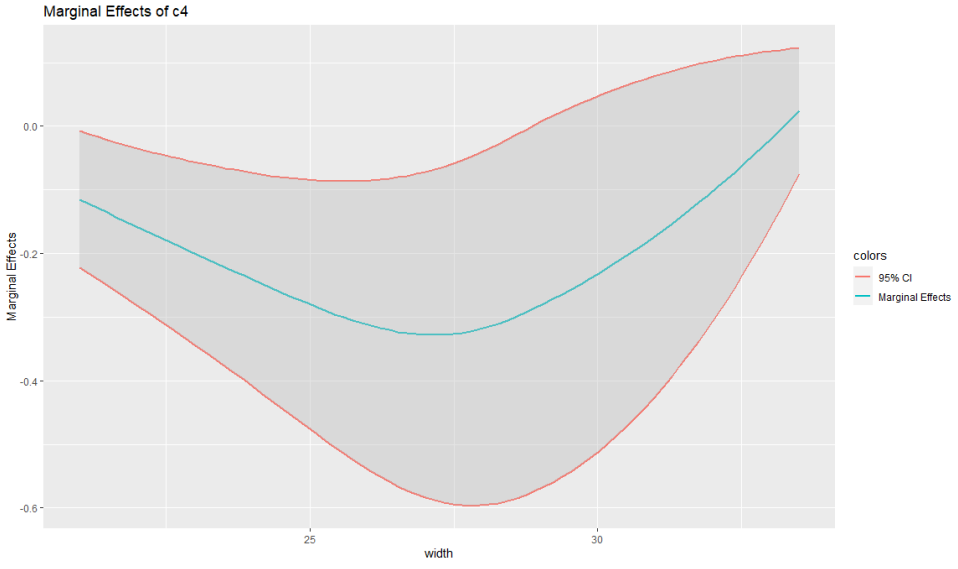
Επειδή όπως αναφέρθηκε, υπάρχουν διαφορές στα ΜΕ αναλόγως την προσέγγιση που χρησιμοποιήθηκε. Καλύτερη αίσθηση ως προς το πως κυμαίνονται τα ΜΕ μίας μεταβλητής προκύπτει από την κατασκευή του διαγράμματος που απεικονίζει τα ΜΕ και των 95% διαστημάτων εμπιστοσύνης τους. Με αυτό τον τρόπο έχουμε μία καλύτερη εικόνα ως προς την κατανομή των ΜΕ.

ΜΕ της επεξηγηματικής μεταβλητής Width



Από το παραπάνω σχήμα, προκύπτει ό,τι όσο μεγαλύτερο είναι το πλάτος του θηλυκού καβουριού, τόσο ό,τι η οριακή επίδραση της επεξηγηματικής μεταβλητής Width μικραίνει. Αντιθέτως η οριακή επίδραση της Width κορυφώνονται, όταν το θηλυκό καβούρι έχει πλάτος κοντά στη μέση παρατηρούμενη τιμή του.

ΜΕ της επεξηγηματικής μεταβλητής c4



Αντίθετα με προηγουμένως, παρατηρείται ό,τι όσο μεγαλύτερο ή μικρότερο είναι το πλάτος του θηλυκού καβουριού, τόσο ό,τι η οριακή επίδραση της επεξηγηματικής μεταβλητής c4 αυξάνεται, ενώ όσο το καβούρι έχει πλάτος κοντά στημέση παρατηρούμενη τιμή του τόσο μικραίνει η οριακή επίδραση της c4.

6.4 Προσαρμογή γενικευμένων γραμμικών μοντέλων με μεταβλητή απόκρισης απαριθμήσεων

Για την προσαρμογή των μοντέλων Poisson & ZIP, η μεταβλητή απόκρισης των μοντέλων που θα χρησιμοποιηθεί θα είναι η μεταβλητή psi, η οποία δηλώνει το πλήθος των παρατηρούμενων δορυφόρων που υπάρχουν στην φωλιά του κάθε θηλυκού.

6.4.1 Μοντέλο Poisson

Συνεχίζοντας θα προσαρμόσουμε ένα μοντέλο Poisson, το οποίο θα αποτελείται από την συνεχής μεταβλητή απόκρισης Width και τους παράγοντες Color και Spine.

```
Call:
glm(formula = psi ~ width + factor(spine) + factor(color), family = Poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0591  -1.9468  -0.4864   0.9620   4.7608

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.54385    0.62426  -4.075 4.60e-05 ***
width           0.14596    0.02189   6.669 2.58e-11 ***
factor(spine)2 -0.13879    0.21269  -0.653   0.5141
factor(spine)3  0.02363    0.11729   0.201   0.8403
factor(color)2 -0.22158    0.16789  -1.320   0.1869
factor(color)3 -0.46036    0.19554  -2.354   0.0186 *
factor(color)4 -0.48544    0.22824  -2.127   0.0334 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 558.63  on 166  degrees of freedom
AIC: 927.93

Number of Fisher Scoring iterations: 6
```

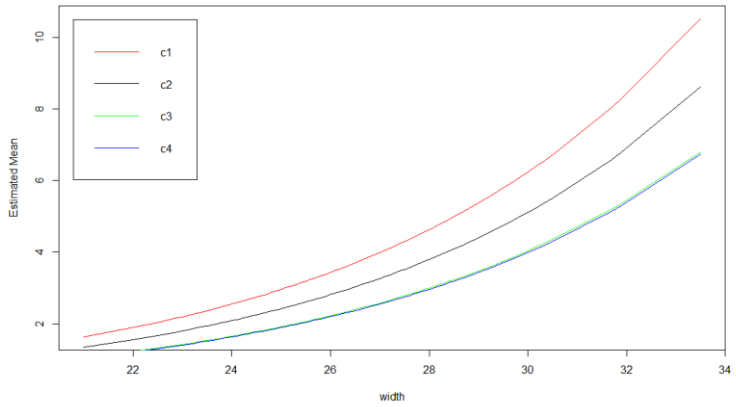
Από το p-values για τον έλεγχο Wald, υποψιαζόμαστε ό,τι η μεταβλητή Spine φαίνεται να μην είναι στατιστικά σημαντική για το μοντέλο μας. Αυτό θα επιβεβαιωθεί τρέχοντας ένα καινούργιο μοντέλο χωρίς την Spine και μέσω του ελέγχου γενικευμένου λόγου πιθανοφανειών θα συγκινούμε τα δύο μοντέλα.

```
Analysis of Deviance Table

Model 1: psi ~ width + factor(color)
Model 2: psi ~ width + factor(spine) + factor(color)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       168     559.34
2       166     558.63  2   0.7153   0.6993
```

Από τον παραπάνω έλεγχο παρατηρείται ό,τι το μοντέλο με τις μεταβλητές Width και Color είναι προτιμότερο.

```
Call:
glm(formula = psi ~ width + factor(color), family = Poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0415  -1.9581  -0.5575   0.9830   4.7523

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.65004    0.58802  -4.507 6.58e-06 ***
width           0.14934    0.02084   7.166 7.73e-13 ***
factor(color)2 -0.19969    0.15364  -1.300   0.1937
factor(color)3 -0.43636    0.17636  -2.474   0.0133 *
factor(color)4 -0.44736    0.20912  -2.139   0.0324 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 559.34  on 168  degrees of freedom
AIC: 924.64

Number of Fisher Scoring iterations: 6
```

Συνεχίζοντας από το παραπάνω αποτέλεσμα βλέπουμε ό,τι οι κατηγορίες c3 και c4, ενδέχεται να μην στατιστικά σημαντικές. Οι διαφορά των δύο κατηγοριών σε σχέση με τις άλλες κατηγορίες μπορεί αν το παρατηρήσει κανείς και από το παρακάτω γράφημα της μ συναρτήσει της Width για των 4 κατηγοριών της Color.



Οπότε, θα κατασκευάσουμε ένα καινούργιο μοντέλο, το οποίο θα αποτελείται από την μεταβλητή Width και από τις κατηγορίες c3 και c4 της Color και στην συνέχεια θα το συγκρίνουμε με το μοντέλο που αποτελείται από την μεταβλητή Width και την Color.

```
Analysis of Deviance Table

Model 1: psi ~ width + factor(c3) + factor(c4)
Model 2: psi ~ width + factor(color)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       169     560.95
2       168     559.34  1    1.609   0.2046
```

Συνεπώς, καταλήγουμε ό,τι το καλύτερο μοντέλο είναι αυτό το οποίο αποτελείται από την μεταβλητή Width και τις κατηγορίες c3 και c4.

```
Call:
glm(formula = psi ~ width + factor(c3) + factor(c4), family = Poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9359  -2.0007  -0.4107   1.0257   4.6929

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.81808    0.57177  -4.929 8.28e-07 ***
width        0.14909    0.02075   7.184 6.77e-13 ***
factor(c3)1 -0.26176    0.11625  -2.252   0.0243 *
factor(c4)1 -0.27289    0.16176  -1.687   0.0916 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 560.95  on 169  degrees of freedom
AIC: 924.25

Number of Fisher Scoring iterations: 6
```

Δηλαδή, το μοντέλο:

$$\hat{\mu} = exp^{\hat{\beta}_0+\hat{\beta}_1 Width+\hat{\beta}_2 c2+\hat{\beta}_3 c3}$$

$$\Leftrightarrow$$

$$\hat{\mu} = -2.818 + 0.1490\, Width - 0.2614c3 + 0.27289c4$$

ME και τυπικά σφάλματα των επεξηγηματικών μεταβλητών του μοντέλου Poisson

Στην συνέχεια υπολογίζονται τα ME και τα τυπικά τους σφάλματα για τις επεξηγηματικές μεταβλητές Width και c3 & c4, με βάση τις προσεγγίσεις MEM και AME.

Προσέγγιση MEM

```
  Variables     MEM     SD         z            p   lower   upper
1     Width  0.4059 0.0544  7.461397 8.560983e-14  0.2993  0.5125
2        c3 -0.6702 0.2787 -2.404736 1.618414e-02 -1.2164 -0.1240
3        c4 -0.6732 0.3596 -1.872080 6.119552e-02 -1.3780  0.0316
```

Προσέγγιση AME

```
  Variables     AME     SD         z            p   lower   upper
1     Width  0.4352 0.0636  6.842767 7.767778e-12  0.3105  0.5599
2        c3 -0.7113 0.2943 -2.416922 1.565239e-02 -1.2881 -0.1345
3        c4 -0.7166 0.3813 -1.879360 6.019534e-02 -1.4639  0.0307
```

Οι οριακές επιδράσεις, βλέπουμε ό,τι διαφέρουν από προσέγγιση σε προσέγγιση, γεγονός που προκύπτει από το μικρός μέγεθος της βάσης δεδομένων. Επιπλέον, παρατηρούμε ό,τι ενώ η δίτιμη μεταβλητή c4, συμβάλει στην καλή προσαρμογή του μοντέλου, ωστόσο η οριακή της επίδραση δεν είναι στατιστικά σημαντική και για τις δύο προσεγγίσεις.

6.4.2 Μοντέλο ZIP

Με παρόμοια διαδικασία όπως περιεγράφηκε προηγουμένως, καταλήγουμε ό,τι καλύτερο μοντέλο είναι αυτό που αποτελείται από την συνεχή επεξηγηματική μεταβλητή Width και την διακριτή Color.

```
Call:
zeroinfl(formula = psi ~ width + factor(color), dist = "Poisson")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.6960 -0.8139 -0.2839  0.6694  4.2873

Count model coefficients (Poisson with log link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.58082    0.61899   0.938   0.3481
width          0.04091    0.02215   1.847   0.0648 .
factor(color)2 -0.19489    0.15613  -1.248   0.2119
factor(color)3 -0.36806    0.17900  -2.056   0.0398 *
factor(color)4  0.22245    0.21037   1.057   0.2903

Zero-inflation model coefficients (binomial with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    11.4009     2.9532   3.860 0.000113 ***
width          -0.4694     0.1086  -4.321 1.55e-05 ***
factor(color)2  -0.1035     0.7543  -0.137 0.890839
factor(color)3   0.1793     0.7935   0.226 0.821222
factor(color)4   1.3486     0.8623   1.564 0.117841
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 16
Log-likelihood: -355.7 on 10 Df
```

Δηλαδή, το μοντέλο:

1η Συνάρτηση

$$log\ \hat{\lambda} = \hat{\beta}_0 + \hat{\beta}_1 Width + \hat{\beta}_2 c2 + \hat{\beta}_3 c3 + \hat{\beta}_4 c4$$

$$\Leftrightarrow$$

$$\hat{\lambda} = e^{0.58082 + 0.4091\, Width - 0.19489 c2 - 0.36806 c3 + 0.22245 c4}$$

2η Συνάρτηση

$$logit(\hat{\pi}) = \hat{\gamma}_0 + \hat{\gamma}_1 Width + \hat{\gamma}_2 c2 + \hat{\gamma}_3 c3 + \hat{\gamma}_4 c4$$

$$\Leftrightarrow$$

$$logit(\hat{\pi}) = -11.4009 - 0.4694\,Width - 0.1035c2 + 0.1793c3 + 1.3486c4$$

Από τις οποίες εκτιμάται η μέση τιμή:

$$\hat{\mu} = (1 - \hat{\pi})\hat{\lambda}$$

$$\Leftrightarrow$$

$$\hat{\mu} = \frac{e^{0.58082 + 0.4091\,Width - 0.19489c2 - 0.36806c3 + 0.22245c4}}{1 + e^{-11.4009 - 0.4694\,Width - 0.1035c2 + 0.1793c3 + 1.3486c4}}$$

ΜΕ και τυπικά σφάλματα των επεξηγηματικών μεταβλητών του μοντέλου ZIP

Προσέγγιση MEM

	Variables	MEM	SD	z	p	lower	upper
1	Width	0.5697	0.1159	4.9154443	8.858131e-07	0.3425	0.7969
2	c2	-0.4960	0.8690	-0.5707710	5.681549e-01	-2.1992	1.2072
3	c3	-1.1702	0.9520	-1.2292017	2.189962e-01	-3.0361	0.6957
4	c4	-0.9390	1.1625	-0.8077419	4.192392e-01	-3.2175	1.3395

Προσέγγιση AME

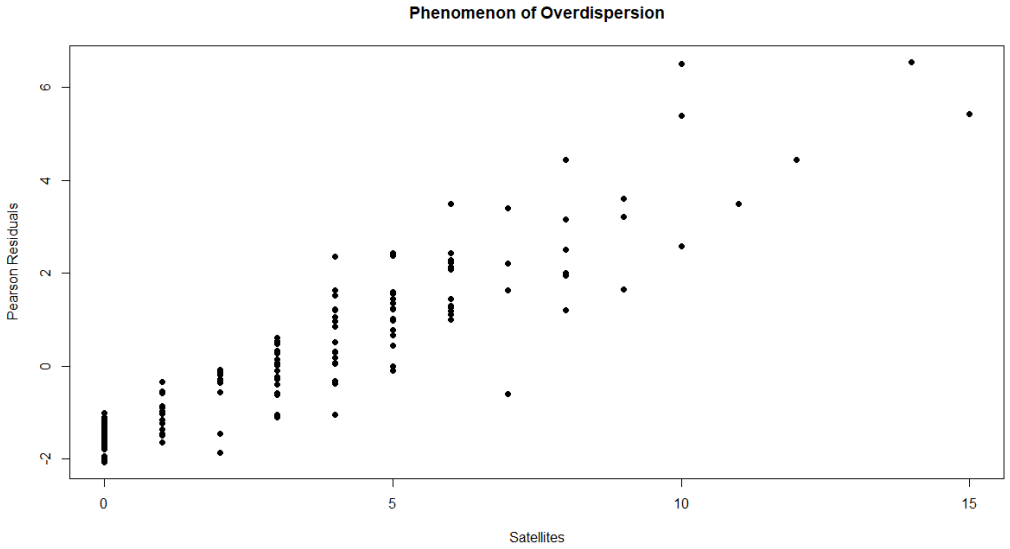
	Variables	AME	SD	z	p	lower	upper
1	Width	0.4908	0.0894	5.4899329	4.020865e-08	0.3156	0.6660
2	c2	-0.5023	0.7746	-0.6484637	5.166851e-01	-2.0205	1.0159
3	c3	-1.0997	0.6853	-1.6046987	1.085601e-01	-2.4429	0.2435
4	c4	-0.6857	1.0507	-0.6526125	5.140061e-01	-2.7450	1.3736

Παρατηρούμε ό,τι ενώ οι δίτιμες μεταβλητές c2, c3 & c4, χρειάζονται για να προσαρμόζεται όσο πιο καλά γίνεται το μοντέλο, ωστόσο η οριακή τους επίδραση δεν είναι στατιστικά σημαντική και για τις δύο προσεγγίσεις.

6.4.3 Εύρεση καλύτερου μοντέλου και ερμηνεία των ΜΕ των επεξηγηματικών μεταβλητών

Το μοντέλο Poisson, που προσαρμοστικέ προηγουμένως, υπόκειται στο φαινόμενο της υπερσκέδασης, γεγονός που επιβεβαιώνεται και από το παρακάτω διάγραμμα διασπορών και των καταλοίπων pearson. Συνεπώς, εφόσον δεν πληρείται η συνθήκη ισοδυναμίας της μέση τιμής και της διασποράς, το μοντέλο Poisson απορρίπτεται. Οπότε, για την ανάλυση των ΜΕ, θα συνεχίσουμε με το μοντέλο ZIP, το οποίο άλλωστε κρίνεται καλύτερο και με βάση το μέτρο προσαρμογής AIC.

	AIC.pois.	AIC.ZIP.
1	924.2514	731.4976



Ερμηνείες των ΜΕ βάση της προσέγγισης MEM του μοντέλο ZIP

- **Width**: Η οριακής της επίδραση ισούται με 0.5697 (0.1159), η ερμηνεία της οποίας σημαίνει ό,τι η διαφορά του μέσου πλήθους των δορυφόρος που βρίσκονται στην φωλιά του θηλυκού καβουριού, όταν το θηλυκό καβούρι έχει το μέσο παρατηρούμενο χρώμα της κατηγορίας c2 (0.549), c3 (0.254) & c4 (0.127), με πλάτος 26.3+ δ cm και του μέσου αριθμού των δορυφόρων όταν το θηλυκό καβούρι έχει το μέσο παρατηρούμενο πλάτος (26.3 cm), εκτιμάται με 0.5697*δ. Για παράδειγμα όταν το θηλυκό έχει το μέσο παρατηρούμενο χρώμα της κατηγορίας c2, c3, c4 και πλάτος 27 cm, ο μέσος αριθμός των δορυφόρων που υπάρχουν στην φωλιά του θηλυκού εκτιμάται με 0.5697*(27-26.3)= 0.39879 μεγαλύτερος από τον μέσο αριθμό των δορυφόρων όταν το θηλυκό καβούρι έχει το μέσο παρατηρούμενο πλάτος (26.3 cm).

Σημειώνεται, ό,τι τα ΜΕ των επιμέρους δίτιμων μεταβλητών c2,c3 & c4 προέκυψαν μην στατιστικά σημαντικά, κατά συνέπεια δεν έχουν νόημα οι ερμηνείες τους.

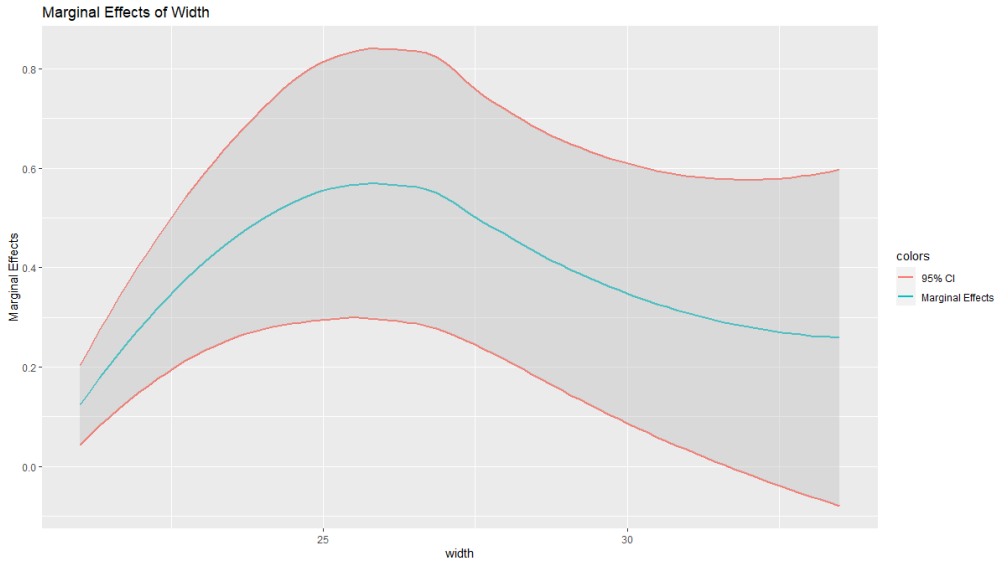
Ερμηνείες των ΜΕ βάση την προσέγγισης AME του μοντέλου ZIP

- **Width**: Η μέση οριακής της επίδραση ισούται με 0.4908 (0.0894). Αυτό σημαίνει ό,τι όταν το πλάτος του θηλυκού, για κάθε παρατήρηση αυξηθεί κατά δ, δηλαδή $Width_i$+δ, τότε ο μέσος αριθμός αρσενικών καβουριών που βρίσκονται στην φωλιά ενός θηλυκού, ανεξαρτήτως τι χρώμα έχει το θηλυκό, εκτιμάται περίπου κατά μέσο όρο με 0.4908*δ μεγαλύτερος από όταν το θηλυκό είχε το αρχικό πλάτος για κάθε παρατήρηση $Width_i$. Για παράδειγμα, αν το πλάτος κάθε παρατήρησης αυξανότανε με 1 cm, δηλαδή $Width_i$+1, τότε ο μέσος αριθμός των δορυφόρων μέσα στην φωλιά του θηλυκού είναι

κατά μέσο όρο 0.4908 *1= 0.4908 μεγαλύτερος από όταν το θηλυκό είχε το αρχικό πλάτος του, $Width_i$, για κάθε παρατήρηση.

Από τα παραπάνω αποτελέσματα, συμπεράνει κανείς ό,τι τα αρσενικά καβούρια έχουν μία τάση να προτιμάνε τα ποιο μεγαλόσωμα θηλυκά καβούρια.

Γραφική απεικόνιση των ΜΕ και των 95% διαστημάτων εμπιστοσύνης της μεταβλητής Width



Από το παραπάνω σχήμα, προκύπτει ό,τι για μικρές και μεγάλες τιμές του πλάτους του θηλυκού καβουριού, η οριακή επίδραση της επεξηγηματικής μεταβλητής Width λαμβάνει μικρές τιμές, ενώ για τιμές του πλάτους κοντά στην μέση του τιμή, το ΜΕ της Width λαμβάνει τις μεγαλύτερες δυνατές τιμές.

Παραρτήματα

Παράρτημα 1

Γενικευμένα Γραμμικά Μοντέλα με Δίτιμες Μεταβλητές Απόκρισης

```
# Correlations Plot
# Multicollinearity
library("corrplot"); res <- cor(mydata[3:6]); library(corrplot);
corrplot(res, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)
      ######---------- Survey for the Best Logit Model ----------######
# Model M2 will all variables
logit2<-glm(y~width+factor(spine)+factor(color),family = binomial);
# Model M3 will all variables without spine
logit3<-glm(y~width+factor(color),family = binomial);
# M3 vs M2 -> Best M3
anova(logit3,logit2,test="Chisq");
# Model M4 with width
logit4<-glm(y~width,family = binomial);
anova(logit4,logit3,test="Chisq")
# Plot for predicted probability p based on colors levels
b<-coef(logit3);
c1<-exp(b[1]+b[2]*sort(width))/(1+exp(b[1]+b[2]*sort(width)));
c2<-exp(b[1]+b[2]*sort(width)+b[3])/(1+exp(b[1]+b[2]*sort(width)+b[3]));
c3<-exp(b[1]+b[2]*sort(width)+b[4])/(1+exp(b[1]+b[2]*sort(width)+b[4]));
c4<-exp(b[1]+b[2]*sort(width)+b[5])/(1+exp(b[1]+b[2]*sort(width)+b[5]));
plot(sort(width),c1,type = "l",col="red",xlab = "width",ylab = "Estimated Propability")
lines(sort(width),c2);lines(sort(width),c3,col="green");lines(sort(width),c4,col="blue")
# Model M with width & c4
c4<-1*(color==4)
```

```
logit<-glm(y~width+factor(c4),family = binomial);
# M4 vs M -> Best M
anova(logit4,logit,test="Chisq");
```

####------ MEM & SD-----####

###----Width (continuous)---###

```
#MEM1
b<-matrix(coef(logit),nrow=3);
xt<-cbind(1,mean(width),mean(c4));  xtb<-xt%*%b;
logit_MEM1<-round(exp(xtb)/(1+exp(xtb))^2*b[2],4);
#Sd1
L<-as.numeric(exp(xtb)/(1+exp(xtb))); Lt<-as.numeric(L*(1-L)); I<-diag(3);
x<-t(xt); xbt<-x%*%t(b);
DME1<-Lt*(I+(1-2*L)*xbt); DMET1<-t(DME1);
# Coef's vcov matrix
V<-vcov(logit);
##Standard errors
VME1<-DMET1%*%V%*%DME1;
logit_SD1<-round(sqrt(VME1)[2,2],4);
# 95% Confidence Interval
logit_DE1<-round(logit_MEM1+c(-1,1)*qnorm(0.975,0,1)*logit_SD1,4);
```

###---- c4 (binary)---###

```
#MEM2
 L1<-exp(b[1]+b[2]*mean(width)+b[3])/(1+exp(b[1]+b[2]*mean(width)+b[3]));
L0<-exp(b[1]+b[2]*mean(width))/(1+exp(b[1]+b[2]*mean(width)));
logit_MEM2<-round(L1-L0,4);
#Sd2
DME2<-(L1/(1+exp(b[1]+b[2]*mean(width)+b[3])))*matrix(c(1,mean(width),1))-
(L0/(1+exp(b[1]+b[2]*mean(width))))*matrix(c(1,mean(width),0));
DMET2<-t(DME2);
```

```
logit_SD2<-round(sqrt(DMET2%*%V%*%DME2),4);
# 95% Confidence Interval
logit_DE2<-round(logit_MEM2+c(-1,1)*qnorm(0.975,0,1)*logit_SD2,4);
####------ Sum-----####
Variables<-c("Width","c4");
logit_MEM<-c(logit_MEM1,logit_MEM2);
logit_SD<-c(logit_SD1,logit_SD2);
logit_z<-c(logit_MEM1/logit_SD1,logit_MEM2/logit_SD2);
DE1<-c(logit_DE1[1],logit_DE2[1]);
DE2<-c(logit_DE1[2],logit_DE2[2]);
p<-c(2*pnorm(logit_MEM1/logit_SD1,lower.tail =
F),2*pnorm(logit_MEM2/logit_SD2,lower.tail = T));
data.frame(Variables,MEM=logit_MEM,SD=logit_SD,z=logit_z,p,lower=DE1,upper=DE2)
```

####------ AME & SD-----####

###----Width (continuous)---###

```
#AME1
Xt<-cbind(1,width,c4);
AME1<-0
for (i in 1:length(width)) {
  AME1<-AME1+((exp(Xt[i,]%*%b)/(1+exp(Xt[i,]%*%b))^2)*b[2])/173
}
logit_AME1<-round(AME1,4);

#Sd1
Xt%*%b;
L<-exp(Xt%*%b)/(1+exp(Xt%*%b)); Lt<-L*(1-L); I<-diag(3); X<-t(Xt);
dme1<-list();
D<-matrix(0,nrow=3,ncol=3);
for (i in 1:length(width)) {
  dme1[i]<-list(Lt[i]*(I+(1-2*L[i])*(X[,i]%*%t(b))));
```

```
  D<-D+matrix(unlist(dme1[i]),ncol = 3)
}
DME1<-D/173; DMET1<-t(DME1);
logit_Sd1<-round(sqrt(DMET1%*%V%*%DME1)[2,2],4);
```

 # 95% Confidence Interval

```
logit_de1<-round(logit_AME1+c(-1,1)*qnorm(0.975,0,1)*logit_Sd1,4);
```

###---- c4 (binary)---###

#AME2

```
l1<-exp(b[1]+b[2]*width+b[3])/(1+exp(b[1]+b[2]*width+b[3]));
l0<-exp(b[1]+b[2]*width)/(1+exp(b[1]+b[2]*width));
AME2<-0
for (i in 1:length(width)) {
  AME2<-AME2+(l1[i]-l0[i])/length(width)
}
logit_AME2<-round(AME2,4);
```

#Sd2

```
a1<-l1/(1+exp(b[1]+b[2]*width+b[3])); a0<-l0/(1+exp(b[1]+b[2]*width));
dme2<-list();
D<-matrix(0,nrow=3);
for (i in 1:length(width)) {
  dme2[i]<-list(a1[i]*matrix(c(1,width[i],1))-a0[i]*matrix(c(1,width[i],0)));
  D<-D+matrix(unlist(dme2[i]),nrow = 3)
}
DME2<-D/173; DMET2<-t(DME2);
logit_Sd2<-round(sqrt(DMET2%*%V%*%DME2),4);
```

 # 95% Confidence Interval

```
logit_de2<-round(logit_AME2+c(-1,1)*qnorm(0.975,0,1)*logit_Sd2,4);
```

####------ Sum-----####

```
Variables<-c("Width","c4");
```

```
logit_AME<-c(logit_AME1,logit_AME2);
logit_Sd<-c(logit_Sd1,logit_Sd2);
logit_z<-c(logit_AME1/logit_Sd1,logit_AME2/logit_Sd2);
DE1<-c(logit_de1[1],logit_de2[1]);
DE2<-c(logit_de1[2],logit_de2[2]);
p<-c(2*pnorm(logit_AME1/logit_Sd1,lower.tail =
F),2*pnorm(logit_AME2/logit_Sd2,lower.tail = T));
data.frame(Variables,AME=logit_AME,SD=logit_Sd,z=logit_z,p,lower=DE1,upper=DE2)
```

###----- Probit Model -----###

```
c4<-1*(color==4);
probit<-glm(y~width+factor(c4),family = binomial(link=probit));summary(probit)
```

####------ MEM & SD-----####

###----Width (continuous)---###

#MEM1

```
b<-matrix(coef(probit),nrow=3);
xt<-cbind(1,mean(width),mean(c4)); xtb<-xt%*%b; xtb<-as.numeric(xtb);
fi<-dnorm(xtb,0,1);
probit_MEM1<-round(fi*b[2],4);
```

#Sd1

```
I<-diag(3); x<-t(xt);  xbt<-x%*%t(b);
DME1<-fi*(I-xtb*xbt); DMET1<-t(DME1);
V<-vcov(probit);
VME1<-DMET1%*%V%*%DME1;
probit_SD1<-round(sqrt(VME1)[2,2],4);
```

 # 95% Confidence Interval

```
probit_DE1<-round(probit_MEM1+c(-1,1)*qnorm(0.975,0,1)*probit_SD1,4);probit_DE1
```

###---- c4 (binary)---###

#MEM2

```
F1<-pnorm(b[1]+b[2]*mean(width)+b[3],0,1);
F0<-pnorm(b[1]+b[2]*mean(width),0,1);
probit_MEM2<-round(F1-F0,4);
```

#Sd2

```
fi1<-dnorm(b[1]+b[2]*mean(width)+b[3],0,1); fi0<-dnorm(b[1]+b[2]*mean(width),0,1);
DME2<-fi1*matrix(c(1,mean(width),1))-fi0*matrix(c(1,mean(width),0));DMET2<-t(DME2);
probit_SD2<-round(sqrt(DMET2%*%V%*%DME2),4);
```

 # 95% Confidence Interval

```
probit_DE2<-round(probit_MEM2+c(-1,1)*qnorm(0.975,0,1)*probit_SD2,4);
```

####------ Sum-----####

```
Variables<-c("Width","c4");
probit_MEM<-c(probit_MEM1,probit_MEM2); probit_SD<-c(probit_SD1,probit_SD2);
probit_z<-c(probit_MEM1/probit_SD1,probit_MEM2/probit_SD2);
DE1<-c(probit_DE1[1],probit_DE2[1]); DE2<-c(probit_DE1[2],probit_DE2[2])
p<-c(2*pnorm(probit_MEM1/probit_SD1,lower.tail =
F),2*pnorm(probit_MEM2/probit_SD2,lower.tail = T))
data.frame(Variables,MEM=probit_MEM,SD=probit_SD,z=probit_z,p,lower=DE1,upper=D
E2)
```

####------ AME & SD-----####

###----Width (continuous)---###

#AME1

```
Xt<-cbind(1,width,c4); fi<-dnorm(Xt%*%b,0,1);
probit_AME1<-round(sum((fi*b[2])/length(width)),4);
```

#Sd1

```
I<-diag(3); X<-t(Xt);
dme1<-list();
D<-matrix(0,nrow=3,ncol=3);
for (i in 1:length(width)) {
  dme1[i]<-list(fi[i]*(I-(Xt%*%b)[i]*(X[,i]%*%t(b))));
  D<-D+matrix(unlist(dme1[i]),ncol = 3);
}
DME1<-D/173; DMET1<-t(DME1);DMET1
probit_Sd1<-round(sqrt(DMET1%*%V%*%DME1)[2,2],4);
```

 # 95% Confidence Interval

```
probit_de1<-round(probit_AME1+c(-1,1)*qnorm(0.975,0,1)*probit_Sd1,4);
```

###---- c4 (binary)---###

#AME2

```
FI1<-pnorm(b[1]+b[2]*width+b[3],0,1); FI0<-pnorm(b[1]+b[2]*width,0,1);
probit_AME2<-round(sum((FI1-FI0)/length(width)),4);
```

#Sd2

```
f1<-dnorm(b[1]+b[2]*width+b[3],0,1); f0<-dnorm(b[1]+b[2]*width,0,1);
dme2<-list();
D<-matrix(0,nrow=3);
for (i in 1:length(width)) {
  dme2[i]<-list(f1[i]*matrix(c(1,width[i],1))-f0[i]*matrix(c(1,width[i],0)));
  D<-D+matrix(unlist(dme2[i]),nrow = 3)
}
DME2<-D/173; DMET2<-t(DME2);
probit_Sd2<-round(sqrt(DMET2%*%V%*%DME2),4);
```

 # 95% Confidence Interval

```
probit_de2<-round(probit_AME2+c(-1,1)*qnorm(0.975,0,1)*probit_Sd2,4);
```

####------ Sum-----####

```
Variables<-c("Width","c4");
probit_AME<-c(probit_AME1,probit_AME2); probit_Sd<-c(probit_Sd1,probit_Sd2);
probit_z<-c(probit_AME1/probit_Sd1,probit_AME2/probit_Sd2);
DE1<-c(probit_de1[1],probit_de2[1]); DE2<-c(probit_de1[2],probit_de2[2])
p<-c(2*pnorm(probit_AME1/probit_Sd1,lower.tail =
F),2*pnorm(probit_AME2/probit_Sd2,lower.tail = T))
data.frame(Variables,AME=probit_AME,SD=probit_Sd,z=probit_z,p,lower=DE1,upper=DE2
)
```

###----- Cloglog Model -----###

```
c4<-1*(color==4);
cloglog<-glm(y~width+factor(c4),family = binomial(link=cloglog));summary(cloglog)
```

####------ MEM & SD-----####

###----Width (continuous)---###

#MEM1

```
b<-matrix(coef(cloglog),nrow=3);
xt<-cbind(1,mean(width),mean(c4)); xtb<-xt%*%b; xtb<-as.numeric(xtb);
w<-exp(xtb-exp(xtb));
cloglog_MEM1<-round(w*b[2],4);
```

#Sd1

```
I<-diag(3);
W<-1-exp(-exp(xtb));
x<-t(xt); xbt<-x%*%t(b);
DME1<-exp(xtb)*(1-W)*(I+(1-exp(xtb))*xbt); DMET1<-t(DME1);
V<-vcov(cloglog);
VME1<-DMET1%*%V%*%DME1;
```

```
cloglog_SD1<-round(sqrt(VME1)[2,2],4);
```

 # 95% Confidence Interval

```
cloglog_DE1<-round(cloglog_MEM1+c(-
1,1)*qnorm(0.975,0,1)*cloglog_SD1,4);cloglog_DE1
```

###---- c4 (binary)---###

#MEM2

```
W1<-1-exp(-exp(b[1]+b[2]*mean(width)+b[3])); W0<-1-exp(-exp(b[1]+b[2]*mean(width)))
cloglog_MEM2<-round(W1-W0,4);
```

#Sd2

```
w1<-exp((b[1]+b[2]*mean(width)+b[3])-exp(b[1]+b[2]*mean(width)+b[3]));
w0<-exp((b[1]+b[2]*mean(width))-exp(b[1]+b[2]*mean(width)));
DME2<-w1*matrix(c(1,mean(width),1))-w0*matrix(c(1,mean(width),0));
DMET2<-t(DME2);
cloglog_SD2<-round(sqrt(DMET2%*%V%*%DME2),4);
```

 # 95% Confidence Interval

```
cloglog_DE2<-round(cloglog_MEM2+c(-1,1)*qnorm(0.975,0,1)*cloglog_SD2,4);
```

####------ Sum-----####

```
Variables<-c("Width","c4")
cloglog_MEM<-c(cloglog_MEM1,cloglog_MEM2);
cloglog_SD<-c(cloglog_SD1,cloglog_SD2);
cloglog_z<-c(cloglog_MEM1/cloglog_SD1,cloglog_MEM2/cloglog_SD2);
DE1<-c(cloglog_DE1[1],cloglog_DE2[1]); DE2<-c(cloglog_DE1[2],cloglog_DE2[2])
p<-c(2*pnorm(cloglog_MEM1/cloglog_SD1,lower.tail =
F),2*pnorm(cloglog_MEM2/cloglog_SD2,lower.tail = T))
data.frame(Variables,MEM=cloglog_MEM,SD=cloglog_SD,z=cloglog_z,p,lower=DE1,upper
=DE2)
```

####------ AME & SD-----####

###----Width (continuous)---###

#AME1

```
Xt<-cbind(1,width,c4); Xtb<-Xt%*%b; w<-exp(Xtb-exp(Xtb));
cloglog_AME1<-round(sum((w*b[2])/length(width)),4);
```

#Sd1

```
I<-diag(3); X<-t(Xt); W<-1-exp(-exp(Xt%*%b));
dme1<-list();
D<-matrix(0,nrow=3,ncol=3);
for (i in 1:length(width)) {
  dme1[i]<-list(exp((Xt%*%b)[i])*(1-W[i])*(I+(1-exp((Xt%*%b)[i]))*(X[,i]%*%t(b))));
  D<-D+matrix(unlist(dme1[i]),ncol = 3);
}
DME1<-D/length(width); DMET1<-t(DME1);
cloglog_Sd1<-round(sqrt(DMET1%*%V%*%DME1)[2,2],4);
```

 # 95% Confidence Interval

```
cloglog_de1<-round(cloglog_AME1+c(-1,1)*qnorm(0.975,0,1)*cloglog_Sd1,4);
```

###---- c4 (binary)---###

#AME2

```
W1<-1-exp(-exp(b[1]+b[2]*width+b[3])); W0<-1-exp(-exp((b[1]+b[2]*width)));
cloglog_AME2<-round(sum((W1-W0)/length(width)),4);
```

#Sd2

```
w1<-exp((b[1]+b[2]*width+b[3])-exp(b[1]+b[2]*width+b[3]));
w0<-exp((b[1]+b[2]*width)-exp(b[1]+b[2]*width));
dme2<-list();
D<-matrix(0,nrow=3);
for (i in 1:length(width)) {
  dme2[i]<-list(w1[i]*matrix(c(1,width[i],1))-w0[i]*matrix(c(1,width[i],0)));
```

```
  D<-D+matrix(unlist(dme2[i]),nrow = 3)
}
DME2<-D/173; DMET2<-t(DME2);
cloglog_Sd2<-round(sqrt(DMET2%*%V%*%DME2),4);
# 95% Confidence Interval
cloglog_de2<-round(cloglog_AME2+c(-1,1)*qnorm(0.975,0,1)*cloglog_Sd2,4);

                    ####------ Sum-----####
cloglog_AME<-c(cloglog_AME1,cloglog_AME2);
cloglog_Sd<-c(cloglog_Sd1,cloglog_Sd2);
cloglog_z<-c(cloglog_AME1/cloglog_Sd1,cloglog_AME2/cloglog_Sd2);
DE1<-c(cloglog_de1[1],cloglog_de2[1]); DE2<-c(cloglog_de1[2],cloglog_de2[2])
p<-c(2*pnorm(cloglog_AME1/cloglog_Sd1,lower.tail =
F),2*pnorm(cloglog_AME2/cloglog_Sd2,lower.tail = T))
data.frame(Variables,AME=cloglog_AME,SD=cloglog_Sd,z=cloglog_z,p,lower=DE1,upper=
DE2)

                    #---BEST BINARY MODEL -> Cloglog
data.frame(AIC(logit),AIC(probit),AIC(cloglog))
```

Παράρτημα 2

Γενικευμένα Γραμμικά Μοντέλα με Μεταβλητές Απόκρισης Απαριθμήσεων

```
      ######---------- Survey for the Best Poisson Model ----------######
# Model M2 will all variables
pois2<-glm(psi~width+factor(spine)+factor(color),family = Poisson);
# Model M3 will all variables without Spine
pois3<-glm(psi~width+factor(color),family = Poisson);
#M3 vs M2 -> Best M3
```

```
anova(pois3,pois2,test="Chisq")
# Model M4 with width
pois4<-glm(psi~width,family = Poisson);
anova(pois4,pois3,test="Chisq")
b<-coef(pois3);
c1<-exp(b[1]+b[2]*sort(width));
c2<-exp(b[1]+b[2]*sort(width)+b[3]);
c3<-exp(b[1]+b[2]*sort(width)+b[4]);
c4<-exp(b[1]+b[2]*sort(width)+b[5]);
plot(sort(width),c1,type = "l",col="red",xlab = "width",ylab = "Estimated Propability")
lines(sort(width),c2); lines(sort(width),c3,col="green"); lines(sort(width),c4,col="blue")
# Model M with width, c3 & c4
c4<-1*(color==4); c3<-1*(color==3);
pois<-glm(psi~width+factor(c3)+factor(c4),family = Poisson);
#  M3 vs M -> Best M
anova(pois,pois3,test="Chisq")
```

####------ MEM & SD-----####

###----Width (continuous)---###

```
#MEM1
b<-matrix(coef(pois),nrow=4);
xt<-cbind(1,mean(width),mean(c3),mean(c4)); xtb<-xt%*%b; xtb<-as.numeric(xtb);
pois_MEM1<-round(exp(xtb)*b[2],4);
#Sd1
I<-diag(4); x<-t(xt);  xbt<-x%*%t(b);
DME1<-exp(xtb)*(I+xbt); DMET1<-t(DME1);
V<-vcov(pois);
VME1<-DMET1%*%V%*%DME1;
pois_SD1<-round(sqrt(VME1)[2,2],4);
# 95% Confidence Interval
pois_DE1<-round(pois_MEM1+c(-1,1)*qnorm(0.975,0,1)*pois_SD1,4);
```

###----c3 (binary)---###

#MEM2

```
m1<-exp(b[1]+b[2]*mean(width)+b[3]+b[4]*mean(c4));

m0<-exp(b[1]+b[2]*mean(width)+b[4]*mean(c4));

pois_MEM2<-round(m1-m0,4);
```

#Sd2

```
m1<-exp(b[1]+b[2]*mean(width)+b[3]+b[4]*mean(c4));

m0<-exp(b[1]+b[2]*mean(width)+b[4]*mean(c4));

DME2<-m1*matrix(c(1,mean(width),1,mean(c4)))-
m0*matrix(c(1,mean(width),0,mean(c4))); DMET2<-t(DME2);

pois_SD2<-round(sqrt(DMET2%*%V%*%DME2),4);
```

 # 95% Confidence Interval

```
pois_DE2<-round(pois_MEM2+c(-1,1)*qnorm(0.975,0,1)*pois_SD2,4);
```

###----c4 (binary)---###

#MEM3

```
m1<-exp(b[1]+b[2]*mean(width)+b[3]*mean(c3)+b[4]);

m0<-exp(b[1]+b[2]*mean(width)+b[3]*mean(c3));

pois_MEM3<-round(m1-m0,4);
```

 # Sd3

```
DME2<-m1*matrix(c(1,mean(width),mean(c3),1))-
m0*matrix(c(1,mean(width),mean(c3),0)); DMET2<-t(DME2);

pois_SD3<-round(sqrt(DMET2%*%V%*%DME2),4);
```

 # 95% Confidence Interval

```
pois_DE3<-round(pois_MEM3+c(-1,1)*qnorm(0.975,0,1)*pois_SD3,4);
```

####------ Sum-----####

```
Variables<-c("Width","c3","c4"); pois_MEM<-c(pois_MEM1,pois_MEM2,pois_MEM3);

pois_SD<-c(pois_SD1,pois_SD2,pois_SD3);

pois_z<-c(pois_MEM1/pois_SD1,pois_MEM2/pois_SD2,pois_MEM3/pois_SD3);
```

```
DE1<-c(pois_DE1[1],pois_DE2[1],pois_DE3[1]);
DE2<-c(pois_DE1[2],pois_DE2[2],pois_DE3[2]);
p<-c(2*pnorm(pois_MEM1/pois_SD1,lower.tail =
F),2*pnorm(pois_MEM2/pois_SD2,lower.tail =
T),2*pnorm(pois_MEM3/pois_SD3,lower.tail = T))
data.frame(Variables,MEM=pois_MEM,SD=pois_SD,z=pois_z,p,lower=DE1,upper=DE2)
```

####------ AME & SD-----####

###----Width (continuous)---###

#AME1

```
Xt<-cbind(1,width,c3,c4); Xtb<-Xt%*%b; m<-exp(Xtb);
pois_AME1<-round(sum((m*b[2])/length(width)),4);
```

#Sd1

```
I<-diag(4); X<-t(Xt);
dme1<-list(); D<-matrix(0,nrow=4,ncol=4);
for (i in 1:length(width)) {
  dme1[i]<-list(exp((Xt%*%b)[i])*(I+(X[,i]%*%t(b))));
  D<-D+matrix(unlist(dme1[i]),ncol = 4);
}
DME1<-D/length(width); DMET1<-t(DME1);
pois_Sd1<-round(sqrt(DMET1%*%V%*%DME1)[2,2],4);
```

 # 95% Confidence Interval

```
pois_de1<-round(pois_AME1+c(-1,1)*qnorm(0.975,0,1)*pois_Sd1,4);
```

###----c3 (binary)---###

#AME2

```
m1<-exp(b[1]+b[2]*width+b[3]+b[4]*c4); m0<-exp(b[1]+b[2]*width+b[4]*c4);
pois_AME2<-round(sum((m1-m0)/length(width)),4);
```

#Sd2

```
dme2<-list(); D<-matrix(0,nrow=4);
for (i in 1:length(width)) {
```

```
  dme2[i]<-list(m1[i]*matrix(c(1,width[i],1,c4[i]))-m0[i]*matrix(c(1,width[i],0,c4[i])));
  D<-D+matrix(unlist(dme2[i]),nrow = 4,ncol = 1)
}
DME2<-D/173; DMET2<-t(DME2);
pois_Sd2<-round(sqrt(DMET2%*%V%*%DME2),4);
```

 # 95% Confidence Interval

```
pois_de2<-round(pois_AME2+c(-1,1)*qnorm(0.975,0,1)*pois_Sd2,4);
```

###----c4 (binary)---###

#AME3

```
m1<-exp(b[1]+b[2]*width+b[3]*c3+b[4]); m0<-exp(b[1]+b[2]*width+b[3]*c3);
pois_AME3<-round(sum((m1-m0)/length(width)),4);
```

#Sd2

```
dme2<-list(); D<-matrix(0,nrow=4);
for (i in 1:length(width)) {
  dme2[i]<-list(m1[i]*matrix(c(1,width[i],c3[i],1))-m0[i]*matrix(c(1,width[i],c3[i],0)));
  D<-D+matrix(unlist(dme2[i]),nrow = 4,ncol = 1)
}
DME2<-D/173; DMET2<-t(DME2);
pois_Sd3<-round(sqrt(DMET2%*%V%*%DME2),4);
```

 # 95% Confidence Interval

```
pois_de3<-round(pois_AME3+c(-1,1)*qnorm(0.975,0,1)*pois_Sd3,4);
```

###------ Sum-----####

```
Variables<-c("Width","c3","c4");
pois_AME<-c(pois_AME1,pois_AME2,pois_AME3);
pois_Sd<-c(pois_Sd1,pois_Sd2,pois_Sd3);
pois_z<-c(pois_AME1/pois_Sd1,pois_AME2/pois_Sd2,pois_AME3/pois_Sd3);
DE1<-c(pois_de1[1],pois_de2[1],pois_de3[1]);
DE2<-c(pois_de1[2],pois_de2[2],pois_de3[2]);
```

```
p<-c(2*pnorm(pois_AME1/pois_Sd1,lower.tail =
F),2*pnorm(pois_AME2/pois_Sd2,lower.tail = T),2*pnorm(pois_AME3/pois_Sd3,lower.tail
= T))

data.frame(Variables,AME=pois_AME,SD=pois_Sd,z=pois_z,p,lower=DE1,upper=DE2)
```

###----- ZIP Model -----###

```
library(pscl)

ZIP<-zeroinfl(psi~width+factor(color),dist="Poisson")
```

####------ MEM & SD-----####

###----Width (continuous)---###

#MEM1

```
c2<-1*(color==2);c3<-1*(color==3);c4<-1*(color==4);

b<-matrix(coef(ZIP)[1:5]);

b0<-b[1];b1<-b[2];b2<-b[3];b3<-b[4];b4<-b[5];

g<-matrix(coef(ZIP)[6:10]);

g0<-g[1];g1<-g[2];g2<-g[3];g3<-g[4];g4<-g[5];


xt<-cbind(1,mean(width),mean(c2),mean(c3),mean(c4)); xtb<-xt%*%b;xtb<-as.numeric(xtb);

xtg<-xt%*%g; xtg<-as.numeric(xtg);

ZIP_MEM1<-round((exp(xtb+xtg)*(b1-g1)+exp(xtb)*b1)/(1+exp(xtg))^2,4);
```

#Sd1

```
x<-matrix(c(1,mean(width),mean(c2),mean(c3),mean(c4))); xbt<-x%*%t(b);

m<-exp(xtb)/(1+exp(xtg)); p<-exp(xtg)/(1+exp(xtg));

A<-m/(1+exp(xtg)); B<--(m*p)/(1+exp(xtg));

xbtgtn<-x%*%(t(b)-t(g)); xbtgtp<-x%*%(t(b)+t(g))

I<-diag(5);

D1<-A*((I+xbt)+exp(xtg)*(I+xbtgtn)); D2<-B*(I+xbtgtp+exp(xtg)*(I+xbtgtn));

D<-rbind(D1,D2); DT<-t(D);

V<-vcov(ZIP);

Total<-DT%*%V%*%D;

SD1<-round(sqrt(Total),4);
```

```
ZIP_SD1<-SD1[2,2];
```

 # 95% Confidence Interval

```
ZIP_DE1<-round(ZIP_MEM1+c(-1,1)*qnorm(0.975,0,1)*ZIP_SD1,4);
```

###----c2 (binary)---###

#MEM2

```
a1<-exp(xtb-b2*mean(c2)+b2)/(1+exp(xtg-g2*mean(c2)+g2));

a0<-exp(xtb-b2*mean(c2))/(1+exp(xtg-g2*mean(c2)));

ZIP_MEM2<-round(a1-a0,4);
```

#Sd2

```
m1<-exp(xtb-b2*mean(c2)+b2)/(1+exp(xtg-g2*mean(c2)+g2));

m0<-exp(xtb-b2*mean(c2))/(1+exp(xtg-g2*mean(c2)));

p1<-exp(xtg-g2*mean(c2)+g2)/(1+exp(xtg-g2*mean(c2)+g2));

p0<-exp(xtg-g2*mean(c2))/(1+exp(xtg-g2*mean(c2)));

D1<-m1*matrix(c(1,mean(width),1,mean(c3),mean(c4)))-
m0*matrix(c(1,mean(width),0,mean(c3),mean(c4)));

D2<-(-
m1*p1)*matrix(c(1,mean(width),1,mean(c3),mean(c4)))+m0*p0*matrix(c(1,mean(width),0,
mean(c3),mean(c4)))

D<-rbind(D1,D2); DT<-t(D);

To<-DT%*%V%*%D;

ZIP_SD2<-round(sqrt(To),4);
```

 # 95% Confidence Interval

```
ZIP_DE2<-round(ZIP_MEM2+c(-1,1)*qnorm(0.975,0,1)*ZIP_SD2,4);
```

###----c3 (binary)---###

#MEM3

```
a1<-exp(xtb-b3*mean(c3)+b3)/(1+exp(xtg-g3*mean(c3)+g3));

a0<-exp(xtb-b3*mean(c3))/(1+exp(xtg-g3*mean(c3)));

ZIP_MEM3<-round(a1-a0,4);
```

#Sd3

```
m1<-exp(xtb-b3*mean(c3)+b3)/(1+exp(xtg-g3*mean(c3)+g3));

m0<-exp(xtb-b3*mean(c3))/(1+exp(xtg-g3*mean(c3)));

p1<-exp(xtg-g3*mean(c3)+g3)/(1+exp(xtg-g3*mean(c3)+g3));

p0<-exp(xtg-g3*mean(c3))/(1+exp(xtg-g3*mean(c3)));

D1<-m1*matrix(c(1,mean(width),mean(c2),1,mean(c4)))-
m0*matrix(c(1,mean(width),mean(c2),0,mean(c4)));

D2<-(-
m1*p1)*matrix(c(1,mean(width),mean(c2),1,mean(c4)))+m0*p0*matrix(c(1,mean(width),me
an(c3),0,mean(c4)))

D<-rbind(D1,D2); DT<-t(D);

To<-DT%*%V%*%D;

ZIP_SD3<-round(sqrt(To),4);
```

 # 95% Confidence Interval

```
ZIP_DE3<-round(ZIP_MEM3+c(-1,1)*qnorm(0.975,0,1)*ZIP_SD3,4);
```

###----c4 (binary)---###

#MEM4

```
a1<-exp(xtb-b4*mean(c4)+b4)/(1+exp(xtg-g4*mean(c4)+g4));

a0<-exp(xtb-b4*mean(c4))/(1+exp(xtg-g4*mean(c4)));

ZIP_MEM4<-round(a1-a0,4);
```

#Sd4

```
m1<-exp(xtb-b4*mean(c4)+b4)/(1+exp(xtg-g4*mean(c4)+g4));

m0<-exp(xtb-b4*mean(c4))/(1+exp(xtg-g4*mean(c4)));

p1<-exp(xtg-g4*mean(c4)+g4)/(1+exp(xtg-g4*mean(c4)+g4));

p0<-exp(xtg-g4*mean(c4))/(1+exp(xtg-g4*mean(c4)));

D1<-m1*matrix(c(1,mean(width),mean(c2),mean(c3),1))-
m0*matrix(c(1,mean(width),mean(c2),mean(c3),0));

D2<-(-
m1*p1)*matrix(c(1,mean(width),mean(c2),mean(c3),1))+m0*p0*matrix(c(1,mean(width),me
an(c2),mean(c3),0))

D<-rbind(D1,D2); DT<-t(D);

To<-DT%*%V%*%D;

ZIP_SD4<-round(sqrt(To),4);
```

 # 95% Confidence Interval

```
ZIP_DE4<-round(ZIP_MEM4+c(-1,1)*qnorm(0.975,0,1)*ZIP_SD4,4);
```

###------ Sum-----####

```
Variables<-c("Width","c2","c3","c4")

ZIP_MEM<-c(ZIP_MEM1,ZIP_MEM2,ZIP_MEM3,ZIP_MEM4);

ZIP_SD<-c(ZIP_SD1,ZIP_SD2,ZIP_SD3,ZIP_SD4);

ZIP_z<-
c(ZIP_MEM1/ZIP_SD1,ZIP_MEM2/ZIP_SD2,ZIP_MEM3/ZIP_SD3,ZIP_MEM4/ZIP_SD4)
;ZIP_z

DE1<-c(ZIP_DE1[1],ZIP_DE2[1],ZIP_DE3[1],ZIP_DE4[1]);

DE2<-c(ZIP_DE1[2],ZIP_DE2[2],ZIP_DE3[2],ZIP_DE4[2]);

p<-c(2*pnorm(ZIP_MEM1/ZIP_SD1,lower.tail =
F),2*pnorm(ZIP_MEM2/ZIP_SD2,lower.tail = T),2*pnorm(ZIP_MEM3/ZIP_SD3,lower.tail
= T),2*pnorm(ZIP_MEM4/ZIP_SD4,lower.tail = T))

data.frame(Variables,MEM=ZIP_MEM,SD=ZIP_SD,z=ZIP_z,p,lower=DE1,upper=DE2)
```

####------ AME & SD-----####

###----Width (continuous)---###

#AME1

```
xt<-cbind(1,width,c2,c3,c4); xtb<-xt%*%b; xtg<-xt%*%g;

ZIP_AME1<-round(sum(((exp(xtb+xtg)*(b1-
g1)+exp(xtb)*b1)/(1+exp(xtg))^2)/length(width)),4);
```

#Sd1

```
I<-diag(5); m<-exp(xtb)/(1+exp(xtg)); p<-exp(xtg)/(1+exp(xtg));

A<-m/(1+exp(xtg)); B<--(m*p)/(1+exp(xtg)); X<-t(xt);

dme1<-list(); dme2<-list();

D1<-matrix(0,nrow=5,ncol=5); D2<-matrix(0,nrow=5,ncol=5);

for (i in 1:length(width)) {

  dme1[i]<-list(A[i]*((I+X[,i]%*%t(b))+exp(xtg[i])*(I+X[,i]%*%(t(b)-t(g)))));

  dme2[i]<-list(B[i]*(I+X[,i]%*%(t(b)+t(g))+exp(xtg[i])*(I+X[,i]%*%(t(b)-t(g)))));
```

```
  D1<-D1+matrix(unlist(dme1[i]),ncol = 5);  D2<-D2+matrix(unlist(dme2[i]),ncol = 5);
}
D<-rbind(D1,D2)/length(width); DT<-t(D);
V<-vcov(ZIP);
Total<-DT%*%V%*%D;
Sd1<-round(sqrt(Total),4);
ZIP_Sd1<-Sd1[2,2];
```

 # 95% Confidence Interval

```
ZIP_de1<-round(ZIP_AME1+c(-1,1)*qnorm(0.975,0,1)*ZIP_Sd1,4);
```

###----c2 (binary)---###

#AME2

```
a1<-exp(b0+b1*width+b2+b3*c3+b4*c4)/(1+exp(g0+g1*width+g2+g3*c3+g4*c4));
a2<-exp(b0+b1*width+b3*c3+b4*c4)/(1+exp(g0+g1*width+g3*c3+g4*c4))
ZIP_AME2<-round(sum((a1-a2)/173),4);
```

 # Sd2

```
m1<-exp(b0+b1*width+b2+b3*c3+b4*c4)/(1+exp(g0+g1*width+g2+g3*c3+g4*c4));
m0<-exp(b0+b1*width+b3*c3+b4*c4)/(1+exp(g0+g1*width+g3*c3+g4*c4));
p1<-exp(g0+g1*width+g2+g3*c3+g4*c4)/(1+exp(g0+g1*width+g2+g3*c3+g4*c4));
p0<-exp(g0+g1*width+g3*c3+g4*c4)/(1+exp(g0+g1*width+g3*c3+g4*c4));
dme1<-list(); dme2<-list();
D1<-matrix(0,nrow=5); D2<-matrix(0,nrow=5);
for (i in 1:length(width)) {
  dme1[i]<-list(m1[i]*matrix(c(1,width[i],1,c3[i],c4[i]))-
m0[i]*matrix(c(1,width[i],0,c3[i],c4[i])));
  dme2[i]<-list(-
(m1[i]*p1[i])*matrix(c(1,width[i],1,c3[i],c4[i]))+(m0[i]*p0[i])*matrix(c(1,width[i],0,c3[i],c4[
i])));
  D1<-D1+matrix(unlist(dme1[i]),nrow = 5);  D2<-D2+matrix(unlist(dme2[i]),nrow = 5);
}
D<-rbind(D1,D2)/length(width); DT<-t(D);
V<-vcov(ZIP);
```

```
Total<-DT%*%V%*%D;Total
ZIP_Sd2<-round(sqrt(Total),4);
```

 # 95% Confidence Interval

```
ZIP_de2<-round(ZIP_AME2+c(-1,1)*qnorm(0.975,0,1)*ZIP_Sd2,4);
```

###----c3 (binary)---###

#AME3

```
a1<-exp(b0+b1*width+b2*c2+b4*c4+b3)/(1+exp(g0+g1*width+g3+g2*c2+g4*c4))
a2<-exp(b0+b1*width+b2*c2+b4*c4)/(1+exp(g0+g1*width+g2*c2+g4*c4))
ZIP_AME3<-round(sum((a1-a2)/173),4);
```

#Sd3

```
m1<-exp(b0+b1*width+b3+b2*c2+b4*c4)/(1+exp(g0+g1*width+g3+g2*c2+g4*c4));
m0<-exp(b0+b1*width+b2*c2+b4*c4)/(1+exp(g0+g1*width+g2*c2+g4*c4));
p1<-exp(g0+g1*width+g3+g2*c2+g4*c4)/(1+exp(g0+g1*width+g3+g2*c2+g4*c4));
p0<-exp(g0+g1*width+g2*c2+g4*c4)/(1+exp(g0+g1*width+g2*c2+g4*c4));
dme1<-list(); dme2<-list();
D1<-matrix(0,nrow=5); D2<-matrix(0,nrow=5);
for (i in 1:length(width)) {
  dme1[i]<-list(m1[i]*matrix(c(1,width[i],c2[i],1,c4[i]))-
m0[i]*matrix(c(1,width[i],c2[i],0,c4[i])));
  dme2[i]<-list(-
(m1[i]*p1[i])*matrix(c(1,width[i],c2[i],1,c4[i]))+(m0[i]*p0[i])*matrix(c(1,width[i],c2[i],0,c4[
i])));
  D1<-D1+matrix(unlist(dme1[i]),nrow = 5);  D2<-D2+matrix(unlist(dme2[i]),nrow = 5);
}
D<-rbind(D1,D2)/length(width); DT<-t(D);
V<-vcov(ZIP);
Total<-DT%*%V%*%D;
ZIP_Sd3<-round(sqrt(Total),4);
```

 # 95% Confidence Interval

```
ZIP_de3<-round(ZIP_AME3+c(-1,1)*qnorm(0.975,0,1)*ZIP_Sd3,4);
```

###----c4 (binary)---###

#AME4

```
a1<-exp(b0+b1*width+b2*c2+b3*c3+b4)/(1+exp(g0+g1*width+g4+g2*c2+g3*c3));
a2<-exp(b0+b1*width+b2*c2+b3*c3)/(1+exp(g0+g1*width+g2*c2+g3*c3));
ZIP_AME4<-round(sum((a1-a2)/173),4);
```

#Sd4

```
m1<-exp(b0+b1*width+b4+b2*c2+b3*c3)/(1+exp(g0+g1*width+g4+g3*c3+g2*c2));
m0<-exp(b0+b1*width+b2*c2+b3*c3)/(1+exp(g0+g1*width+g2*c2+g3*c3));
p1<-exp(g0+g1*width+g4+g2*c2+g3*c3)/(1+exp(g0+g1*width+g4+g2*c2+g3*c3));
p0<-exp(g0+g1*width+g2*c2+g3*c3)/(1+exp(g0+g1*width+g2*c2+g3*c3));
dme1<-list(); dme2<-list();
D1<-matrix(0,nrow=5); D2<-matrix(0,nrow=5);
for (i in 1:length(width)) {
  dme1[i]<-list(m1[i]*matrix(c(1,width[i],c2[i],c3[i],1))-
m0[i]*matrix(c(1,width[i],c2[i],c3[i],0)));
  dme2[i]<-list(-
(m1[i]*p1[i])*matrix(c(1,width[i],c2[i],c3[i],1))+(m0[i]*p0[i])*matrix(c(1,width[i],c2[i],c3[i]
,0)));
  D1<-D1+matrix(unlist(dme1[i]),nrow = 5); D2<-D2+matrix(unlist(dme2[i]),nrow = 5);
}
D<-rbind(D1,D2)/length(width); DT<-t(D);
V<-vcov(ZIP);
Total<-DT%*%V%*%D;
ZIP_Sd4<-round(sqrt(Total),4);
```

 # 95% Confidence Interval

```
ZIP_de4<-round(ZIP_AME4+c(-1,1)*qnorm(0.975,0,1)*ZIP_Sd4,4);
```

###------ Sum-----####

```
Variables<-c("Width","c2","c3","c4")
ZIP_AME<-c(ZIP_AME1,ZIP_AME2,ZIP_AME3,ZIP_AME4);ZIP_AME
ZIP_Sd<-c(ZIP_Sd1,ZIP_Sd2,ZIP_Sd3,ZIP_Sd4);ZIP_Sd
```

```
ZIP_z<-
c(ZIP_AME1/ZIP_Sd1,ZIP_AME2/ZIP_Sd2,ZIP_AME3/ZIP_Sd3,ZIP_AME4/ZIP_Sd4);ZI
P_z

DE1<-c(ZIP_de1[1],ZIP_de2[1],ZIP_de3[1],ZIP_de4[1]);DE1

DE2<-c(ZIP_de1[2],ZIP_de2[2],ZIP_de3[2],ZIP_de4[2]);DE2

p<-c(2*pnorm(ZIP_AME1/ZIP_Sd1,lower.tail = F),2*pnorm(ZIP_AME2/ZIP_Sd2,lower.tail
= T),2*pnorm(ZIP_AME3/ZIP_Sd3,lower.tail = T),2*pnorm(ZIP_AME4/ZIP_Sd4,lower.tail
= T))

data.frame(Variables,AME=ZIP_AME,SD=ZIP_Sd,z=ZIP_z,p,lower=DE1,upper=DE2)
```

#---BEST COUNT DATA MODEL -> ZIP

```
data.frame(AIC(pois),AIC(ZIP))
```

Παράρτημα 3

Γραφικές απεικόνισεις των ME και των 95% διαστημάτων εμπιστοσύνης

###----- Cloglog Model -----###

#---- Marginal Effects Width ---#

```
c4<-1*(color==4); y<-1*(psi>0)

cloglog<-glm(y~width+factor(c4),family = binomial(link=cloglog))

library(margins)

A<-marginal_effects(cloglog); b<-coef(cloglog); V<-vcov(cloglog);
```

 # Marginal Effects

```
A$dydx_width
```

 # Standard Errors

```
Xt<-cbind(1,width,c4); Xtb<-Xt%*%b;

w<-exp(Xtb-exp(Xtb)); I<-diag(3); X<-t(Xt); W<-1-exp(-exp(Xt%*%b));

dme1<-list(); dme2<-list();

SD<-list();

test<-matrix(0,nrow=173);test

for (i in 1:length(width)) {
```

```
 dme1[i]<-list(exp((Xt%*%b)[i])*(1-W[i])*(I+(1-exp((Xt%*%b)[i]))*(X[,i]%*%t(b))));
 dme2[i]<-list(t(matrix(dme1[[i]],ncol=3)));
 SD[i]<-list(sqrt(matrix(dme2[[i]],ncol=3)%*%V%*%matrix(dme1[[i]],ncol=3)));
 test[i]<-SD[[i]][2,2]
}
```

 # 95% Confidence Interval

```
lower<-A$dydx_width-1*qnorm(0.975,0,1)*test;
upper<-A$dydx_width+1*qnorm(0.975,0,1)*test;
```

Width Marginal effects Plot with the 95% CI

```
D<-data.frame(width,A$dydx_width,lower,upper);
library(ggplot2); library(dplyr)
D<-data.frame(width,A$dydx_width,lower,upper);
g1<-ggplot(D,aes(x=width))+stat_smooth(aes(y=A.dydx_width,color="Marginal
Effects"),se=F)+labs(title="Marginal Effects of Width",y="Marginal Effects")+
 stat_smooth(aes(y=lower,color="95% CI"),se=F)+stat_smooth(aes(y=upper,color="95%
CI"),se=F)+scale_color_discrete(name="colors")
gg1<-ggplot_build(g1)
df2<-data.frame(x=gg1$data[[1]]$x,ymin=gg1$data[[2]]$y,ymax=gg1$data[[3]]$y)
g1+
 geom_ribbon(data=df2,aes(x=x,ymin=ymin,ymax=ymax),fill="grey",alpha=0.4)
```

#---- Marginal Effects c4 ---#

 # Marginal Effects

```
A$dydx_c41
```

 # Standard Errors

```
w1<-exp((b[1]+b[2]*width+b[3])-exp(b[1]+b[2]*width+b[3]));
w0<-exp((b[1]+b[2]*width)-exp(b[1]+b[2]*width));
DEM1<-list(); DEM2<-list();
sd<-matrix(0,nrow=173);
for (i in 1:length(width)) {
```

```
  DME1[i]<-list(w1[i]*matrix(c(1,width[i],1))-w0[i]*matrix(c(1,width[i],0)));
  DME2[i]<-list(t(w1[i]*matrix(c(1,width[i],1))-w0[i]*matrix(c(1,width[i],0))));
  sd[i]<-sqrt(DME2[[i]]%*%V%*%DME1[[i]]);
}
```

 # 95% Confidence Interval

```
Lower<-A$dydx_c41-1*qnorm(0.975,0,1)*sd;
Upper<-A$dydx_c41+1*qnorm(0.975,0,1)*sd;
```

c4 Marginal effects Plot with the 95% CI

```
D1<-data.frame(width,A$dydx_c41,Lower,Upper);D1
library(ggplot2); library(dplyr)
g2<-ggplot(D1,aes(x=width))+stat_smooth(aes(y=A.dydx_c41,color="Marginal Effects"),se=F)+labs(title="Marginal Effects of c4",y="Marginal Effects")+
  stat_smooth(aes(y=Lower,color="95% CI"),se=F)+stat_smooth(aes(y=Upper,color="95% CI"),se=F)+scale_color_discrete(name="colors")
gg2<-ggplot_build(g2)
df2<-data.frame(x=gg2$data[[1]]$x,ymin=gg2$data[[2]]$y,ymax=gg2$data[[3]]$y)
g2+
  geom_ribbon(data=df2,aes(x=x,ymin=ymin,ymax=ymax),fill="grey",alpha=0.4)
```

###----- ZIPModel -----###

#---- Marginal Effects Width ---#

```
library(pscl)
ZIP<-zeroinfl(psi~width+factor(color),dist="Poisson")
c2<-1*(color==2);c3<-1*(color==3);c4<-1*(color==4);
b<-matrix(coef(ZIP)[1:5]);
b0<-b[1];b1<-b[2];b2<-b[3];b3<-b[4];b4<-b[5];
g<-matrix(coef(ZIP)[6:10]); g0<-g[1];g1<-g[2];g2<-g[3];g3<-g[4];g4<-g[5];
V<-vcov(ZIP);
```

 # Marginal Effects

```
library(margins);
ME<-marginal_effects(ZIP)[,1];
```

 # Standard Errors

```
I<-diag(5); xt<-cbind(1,width,c2,c3,c4);  xtb<-xt%*%b; xtg<-xt%*%g;
m<-exp(xtb)/(1+exp(xtg)); p<-exp(xtg)/(1+exp(xtg));
A<-m/(1+exp(xtg)); B<--(m*p)/(1+exp(xtg)); X<-t(xt);
dm1<-list(); dm2<-list();
D<-list(); DT<-list();
SD<-list(); t<-matrix(0,nrow=173);
for (i in 1:length(width)){
  dm1[i]<-list(A[i]*((I+X[,i]%*%t(b))+exp(xtg[i])*(I+X[,i]%*%(t(b)-t(g)))));
  dm2[i]<-list(B[i]*(I+X[,i]%*%(t(b)+t(g))+exp(xtg[i])*(I+X[,i]%*%(t(b)-t(g)))));
  D[i]<-list(rbind(dm1[[i]],dm2[[i]]));
  DT[i]<-list(t(rbind(dm1[[i]],dm2[[i]])));
  SD[i]<-list(sqrt(DT[[i]]%*%V%*%D[[i]]));
  t[i]<-SD[[i]][2,2]
}
```

 # 95% Confidence Interval

```
l<-ME-1*qnorm(0.975,0,1)*t;
u<-ME+1*qnorm(0.975,0,1)*t;
```

Width Marginal effects Plot with the 95% CI

```
W<-data.frame(width,ME,l,u);
library(ggplot2); library(dplyr)
g1<-ggplot(W,aes(x=width))+stat_smooth(aes(y=ME,color="Marginal
Effects"),se=F)+labs(title="Marginal Effects of Width",y="Marginal Effects")+
  stat_smooth(aes(y=l,color="95% CI"),se=F)+stat_smooth(aes(y=u,color="95%
CI"),se=F)+scale_color_discrete(name="colors")
gg1<-ggplot_build(g1)
df2<-data.frame(x=gg1$data[[1]]$x,ymin=gg1$data[[2]]$y,ymax=gg1$data[[3]]$y)
g1+
```

```
geom_ribbon(data=df2,aes(x=x,ymin=ymin,ymax=ymax),fill="grey",alpha=0.4)
```

Βιβλιογραφία

Ελληνική Βιβλιογραφία

Γ. Ηλιόπουλος, Πειραιάς 2019. *Γενικευμένα γραμμικά μοντέλα*, Πανεπιστημιακές Σημειώσεις.

Ξενόγλωσση Βιβλιογραφία

Alan Fernihough (2019). mfx: *Marginal Effects, Odds Ratios and Incidence Rate Ratios for GLMs*. R package version 1.2-2. https://CRAN.R-project.org/package=mfx

Aldrich, John H., and Forrest D. Nelson. (1984). *Linear Probability, Logit, and Probit Models*. Newbury Park, CA: Sage.

Anderson, Soren T. & Newell, Richard G., (2003). *Simplified Marginal Effects in Discrete Choice Models*. Discussion Papers 10631, Resources for the Future.

Bartus, T. (2005). *Estimation of marginal effects using margeff*. Stata Journal, 5(3), 309–329. doi:10.1177/1536867x0500500303

Bockarjova, M. and M. Hazans. (2000). *Marginal effects distributions in logit models of labour markets*. Paper presented at the seminar "Labor Markets, Work, and Welfare During the Transition and Integration Processes", Vilnius, Lithuania

Breen, R. J., Karlson, K. B., & Holm, A. (2018). *Interpreting and understanding logits, probits, and other non-linear probability models*. Annual Review of Sociology, 44, 39–54.

Brockmann, H.. (2010). *Satellite Male Groups in Horseshoe Crabs, Limulus polyphemus*. Ethology. 102. 1 - 21. doi:10.1111/j.1439-0310.1996.tb01099.x.

Cameron, A., & Trivedi, P. (1998). *Regression analysis of count data*. New York: Cambridge University Press.

Cameron, A., & Trivedi, P. (2005). *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511811241

Carlevaro, Fabrizio & Senegas, Marc-Alexandre. (2006). *Simplified marginal effects in discrete choice models*: A correction. Economics Letters. 92. 44-46. 10.1016/j,econlet.2006.01.012.

Carrie, A. (2020). *Horseshoe crab blood is key to making a COVID-19 vaccine—but the ecosystem may suffer*. NATIONAL GEOGRAPHIC.

https://www.nationalgeographic.com/animals/2020/07/covid-vaccine-needs-horseshoe-crab-blood/

Coca, M. (2019). Week 13 : *Interpreting Model Results : Marginal Effects and the margins Command*. Health Services Research Methods I.

Cox, D.R. and Snell, E.J. (1989). *Analysis of Binary Data*. 2nd Edition. London: Chapman and Hall/CRC.

Dowd, B. E., Greene, W. H., & Norton, E. C. (2014). *Computation of standard errors*. Health Services Research, 49(2), 731–750. doi:10.1111/1475-6773.12122

Efron, B. (1979). *Bootstrap Methods: Another Look at the Jackknife*. Annals of Statistics 7 (1): 1–26

Fernihough, A. (1989). *Marginal Effects for Generalized Linear Models: The mfx Package for R*. (2003). Retrieved from https://cran.r-project.org/web/packages/mfx/vignettes/mfxarticle.pdf

Gelman, A., & Pardoe, I. (2007). 2. *Average Predictive Comparisons for Models with Nonlinearity, Interactions, and Variance Components*. Sociological Methodology, 37(1), 23–51. doi:10.1111/j.1467-9531.2007.00181.

Gelman, A., and Jennifer, H. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Gould, W. & Pitblado, J. & Poi, B. (2005). *Maximum Likelihood Estimation with Stata*.

Greene, W. H. (1997). *Econometric Analysis*. 3rd Edition. Upper Saddle River, NJ: Prentice Hall.

Greene, W. H. (2003). *Econometric Analysis*. 5th Edition. Upper Saddle River, NJ: Prentice Hall.

Greene, W. H. (2012). *Econometric Analysis*. 7th Edition. Upper Saddle River, NJ: Prentice Hall.

Greene, W. H. (2000). *Econometric Analysis*. 4th edition. Upper Saddle River, NJ: Prentice Hall.

Hanmer, M. J., & Ozan Kalkan, K. (2013). *Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models*. American Journal of Political Science, 57(1), 263–277. doi:10.1111/j.1540-5907.2012.00602.x

Hannachi, A. (2006). *Exploratory Data Analysis with MATLAB*. In Journal of the Royal Statistical Society: Series A (Statistics in Society) (Vol. 169). doi:10.1111/j.1467-985x.2006.00414_10.x

Hanushek, Eric A., and John E. Jackson. (1977). *Statistical Methods for Social Scientists*. New York: Academic Press.

Hole, A. R. (2008). *Editorial Identification of Treatment Effects*. 1131(2007), 1127–1131. doi:10.1002/hec

Hosmer, D.W. and Lemeshow, S. (2000). *Applied logistic regression*. 2nd Edition. New York: John Wiley & Sons, Inc.. doi:10.1002/0471722146

King, Gary, and Langche Zeng. (2006). *The Dangers of Extreme Counterfactuals*. Political Analysis 14(2): 131–59.

King, Gary. (1998). *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Ann Arbor: University of Michigan Press.

Lambert, Diane. (1992). *Zero-Inflated Poisson Regression, With An Application to Defects in Manufacturing*. Technometrics. 34. 1-14. doi:10.1080/00401706.1992.10485228.

Leeper, T. J. (2018). *Interpreting Regression Results using Average Marginal Effects with R's margins*.Https://Cran.rProject.Org/Web/Packages/Margins/Vignettes/TechnicalDetails.Pdf, 32.Retrieved from https://rdrr.io/cran/margins/f/inst/doc/TechnicalDetails.pdf%0Ahttps://cran.r-project.org/web/packages/margins/vignettes/TechnicalDetails.pdf

Long, J Scott & Mustillo, Sarah. (2018). *Using Predictions and Marginal Effects to Compare Groups in Regression Models for Binary Outcomes*. Sociological Methods & Research. 004912411879937. doi:10.1177/0049124118799374.

Long, J. S., & Freese, J. (2014). *Regression models for categorical dependent variables using stata*. 3rd Edition. Stata Press.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Long, J.S. and Freese, J (2006). *Regression Models for Categorical Dependent Variables Using Stata*. 2nd Edition, Stata Press, Texas.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd Edition. London: Chapman and Hall.

Mize, T. (2019). *Best Practices for Estimating, Interpreting, and Presenting Nonlinear Interaction Effects*. Sociological Science. doi:6. 81-117. 10.15195/v6.a4.

Norton, E. C., & Dowd, B. E. (2018). *Log Odds and the Interpretation of Logit Models*. Health Services Research, 53(2), 859–878. doi:10.1111/1475-6773.12712

Onukwugha, E., Bergtold, J., & Jain, R. (2014). *A Primer on Marginal Effects—Part I: Theory and Formulae*. PharmacoEconomics, 33(1), 25–30. doi:10.1007/s40273-014-0210-6

Onukwugha, E., Bergtold, J., & Jain, R. (2015). *A Primer on Marginal Effects—Part II: Health Services Research Applications*. PharmacoEconomics, 33(2), 97–103. doi:10.1007/s40273-014-0224-0

Piegorsch, W. W. (1992). *Complementary log regression for generalized linear models*. American Statistician, 46(2), 94–99. doi:10.1080/00031305.1992.10475858

Richard Williams (2019). *Marginal Effects for Continuous Variables*. University of Notre Dame. https://www3.nd.edu/~rwilliam/

Signorino, C., & Tarar, A. (2006). *A Unified Theory and Test of Extended Immediate Deterrence*. American Journal of Political Science, 50(3), 586-605. Retrieved January 28, 2021, from http://www.jstor.org/stable/3694236

Thomas J. Leeper (2021). *margins: Marginal Effects for Model Objects*. R package version 0.3.26.

Tomz, Michael, Jason Wittenberg, and Gary King. (2001). *CLARIFY: Software for Interpreting and Presenting Statistical Results*. Version 2.0. Cambridge, MA: Harvard University. http: //gking.harvard.edu.

Wolfinger, Raymond E., and Steven J. Rosenstone. (1980). *Who Votes?* New Haven, CT: Yale University Press.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.