ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

ΣΤΗΝ ΑΝΑΛΟΓΙΣΤΙΚΗ ΕΠΙΣΤΗΜΗ ΚΑΙ ΔΙΟΙΚΗΤΙΚΗ ΚΙΝΔΥΝΟΥ

Τιμολόγηση ασφαλιστηρίων συμβολαίων πυρός με τη χρήση γενικευμένων γραμμικών μοντέλων

Χατζηκωτούλας Αθανάσιος-Εμμανουήλ

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 Αναλογιστική Επιστήμη και Διοικητική Κινδύνου

Πειραιάς

2022

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣ 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 Αναλογιστική Επιστήμη και Διοικητική Κινδύνου.

Τα μέλη της Επιτροπής ήταν:

- Αναπλ. Καθηγητής, Ευστάθιος Χατζηκωνσταντινίδης (Επιβλέπων)

- Καθηγητής, Μιλτιάδης Νεκτάριος

- Αναπλ. Καθηγητής, Πλάτων Τήνιος

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

UNIVERSITY OF PIREAUS

DEPARTMENT OF STATISTICS & INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN ACTUARIAL SCIENCE & RISK MANAGEMENT

Fire insurance pricing using generalized linear models

Chatzikotoulas Athanasios-Emmanouil

MSc Dissertation

submitted to the Department of Statistics & Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master in Actuarial Science and Risk Management

Piraeus

2022

Στους φίλους και την οικογένειά μου

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον κ. Ευστάθιο Χατζηκωνσταντινίδη, Αναπληρωτή καθηγητή του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης και διευθυντή του ΠΜΣ «Αναλογιστική Επιστήμη και Διοικητική Κινδύνου» για την αποδοχή επίβλεψης της διπλωματικής μου εργασίας καθώς και τον Αναπληρωτή Καθηγητή κ. Ψαρράκο Γεώργιο για την άμεση ανταπόκρισή του στην επίλυση των προβληματισμών μου και στην παροχή πολύτιμου υλικού κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

ΠΕΡΙΛΗΨΗ

Στη διπλωματική εργασία θα επιχειρηθεί η τιμολόγηση ασφαλιστηρίων συμβολαίων στον κλάδο Πυρός και συγκεκριμένα ασφαλιστήρια συμβόλαια έναντι του κινδύνου το σεισμού με την βοήθεια γενικευμένων γραμμικών μοντέλων και του στατιστικού πακέτου R.

Αρχικά θα γίνει μια θεωρητική ανάλυση του κλάδου Πυρός, της ασφαλιστικής κάλυψης των καταστροφικών κινδύνων καθώς και των τρόπων διαχείρισής τους. Στη συνέχεια θα παρουσιαστεί το μαθηματικό υπόβαθρο των γενικευμένων γραμμικών μοντέλων και τέλος με τη βοήθεια του στατιστικού πακέτου R θα γίνει υπολογισμός του ασφαλίστρου έναντι του κινδύνου του σεισμού χρησιμοποιώντας πραγματικά πρωτογενή στοιχεία παρεχόμενα από ελληνική ασφαλιστική εταιρεία.

`

ABSTRACT

At this thesis a pricing of insurance contracts in the In the fire line of business and more specifically of insurance contracts against the earthquake risk using generalized linear models and the statistical package R.

Initially a theoretical analysis of the fire line of business, the insurance coverage of catastrophic dangers as well as their management methods will be made. Subsequently the basic mathematical theory of generalized linear models will be presented and finally the earthquake risk premium will be calculated by using the statistical package R based on raw data provided by a Greek insurance company.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1: Ασφάλιση περιουσίας και καταστροφικοί κίνδυνοι

Στο παρόν θα παρουσιαστούν επιγραμματικά όλοι κλάδοι των ασφαλίσεων κατά Ζημιών σύμφωνα με την ελληνική νομοθεσία και το περιεχόμενο του καθενός. Έπειτα θα γίνει μια εκτενέστερη ανάλυση του «Κλάδου Περιουσίας» ,των ασφαλίσεων κατοικίας καθώς και των τρόπων τιμολόγησης των ασφαλιστηρίων συμβολαίων κατοικίας. Τέλος θα γίνει αναφορά στους καταστροφικούς κινδύνους (με ιδιαίτερη έμφαση σε εκείνους που αφορούν καιρικά φαινόμενα) καθώς και στους τρόπους μετριασμού των κινδύνων αυτών.

1.1 Κλάδοι ασφάλισης κατά Ζημιών

Η ιδιωτική ασφάλιση χωρίζεται σε δύο μεγάλες κατηγορίες. Τις ασφαλίσεις που έχουν να κάνουν σχέση με τη ζωή και την υγεία του ατόμου οι οποίες αποκαλούνται «Ασφαλίσεις Ζωής» και την κατηγορία ασφαλειών που αφορούν την περιουσία και την αστική ευθύνη και ονομάζονται «Γενικές Ασφαλίσεις» κατηγορία με την οποίες θα ασχοληθούμε.

Σύμφωνα με το Άρθρο 04 «Κλάδοι ασφαλίσεων κατά ζημιών (άρθρα 2, 197 και Παραρτήματα I και V της Οδηγίας 2009/138/ΕΚ)» , οι ασφαλίσεις κατά ζημιών ταξινομούνται στους ακόλουθους κλάδους ανάλογα με τον κίνδυνο που καλύπτεται από την ασφαλιστική σύμβαση:

α) Κλάδος 1 «Ατυχήματα» :

(συμπεριλαμβανομένων των εργατικών ατυχημάτων και των επαγγελματικών ασθενειών). Περιλαμβάνει τα εξής:

- κατ’ αποκοπή παροχές
- περιοδικές παροχές αποζημιώσεων
- συνδυασμούς των ανωτέρω
- μεταφερόμενα πρόσωπα.

β) Κλάδος 2 «Ασθένειες»

Περιλαμβάνει τα εξής:

- περιοδικές παροχές
- κατ’ αποκοπή παροχές
- συνδυασμούς των ανωτέρω.

γ) Κλάδος 3 «Χερσαία οχήματα» (εκτός σιδηροδρομικών):

Καλύπτει κάθε ζημία, την οποία υφίστανται αυτοκινούμενα και μη, χερσαία οχήματα εκτός των σιδηροδρομικών.

δ) Κλάδος 4 «Σιδηροδρομικά οχήματα»:

Καλύπτει κάθε ζημία, την οποία υφίστανται τα σιδηροδρομικά οχήματα.

ε) Κλάδος 5 «Αεροσκάφη»:

Καλύπτει κάθε ζημία την οποία υφίστανται αεροσκάφη.

στ) Κλάδος 6 Πλοία (θαλάσσια, λιμναία και ποτάμια σκάφη):

Καλύπτει κάθε ζημία, την οποία υφίστανται:

- ποτάμια σκάφη
- λιμναία σκάφη
- θαλάσσια σκάφη/ πλοία

ζ)Κλάδος 7 «Μεταφερόμενα εμπορεύματα»

(συμπεριλαμβανομένων των εμπορευμάτων, αποσκευών και κάθε άλλου αγαθού):

Καλύπτει κάθε ζημία την οποία υφίστανται τα μεταφερόμενα εμπορεύματα, περιλαμβανομένων αποσκευών και κάθε άλλου αγαθού, ανεξαρτήτως του μεταφορικού μέσου.

η) Κλάδος 8 «Πυρκαϊά και στοιχεία της φύσεως»:

Καλύπτει κάθε ζημία που υφίστανται αγαθά, εξαιρουμένων των αγαθών που περιλαμβάνονται στους κλάδους 3, 4, 5, 6 και 7 ανωτέρω, εφόσον προξενείται από:

- πυρκαϊά,
- έκρηξη,
- θύελλα,
- στοιχεία της φύσεως άλλα εκτός θυέλλης,
- πυρηνική ενέργεια και
- καθίζηση του εδάφους.

θ) Κλάδος 9 «Λοιπές ζημίες αγαθών»:

Καλύπτει κάθε ζημία που υφίστανται αγαθά, εξαιρουμένων των αγαθών που περιλαμβάνονται στους κλάδους 3, 4, 5, 6 και 7 ανωτέρω, εφόσον προξενήθηκε από χαλάζι ή παγετό, καθώς και από κάθε άλλο συμβάν, όπως επί παραδείγματι κλοπή, εκτός των συμβάντων που υπάγονται στον κλάδο 8.

ι) Κλάδος 10 «Αστική ευθύνη από χερσαία αυτοκίνητα οχήματα»:

Καλύπτει κάθε είδους αστική ευθύνη, που προκύπτει από τη χρήση χερσαίων αυτοκινήτων οχημάτων, συμπεριλαμβανομένης της ευθύνης του μεταφορέως.

ια) Κλάδος 11 «Αστική ευθύνη από αεροσκάφη»:

Καλύπτει κάθε είδους αστική ευθύνη, που προκύπτει από τη χρήση αεροσκαφών, συμπεριλαμβανομένης της ευθύνης του μεταφορέως.

ιβ) Κλάδος 12 «Αστική ευθύνη από θαλάσσια, λιμναία και ποτάμια σκάφη»:

Καλύπτει κάθε είδους αστική ευθύνη, που προκύπτει από τη χρήση θαλάσσιων, λιμναίων ή ποτάμιων σκαφών, συμπεριλαμβανομένης της ευθύνης του μεταφορέως.

ιγ) Κλάδος 13 «Γενική αστική ευθύνη»:

Καλύπτει κάθε είδους αστική ευθύνη που δεν εμπίπτει στους κλάδους 10 έως 12 ανωτέρω.

ιδ) Κλάδος 14 «Πιστώσεις»:

Στον ως άνω κλάδο, ο ασφαλιστής έναντι ασφαλίστρου καλύπτει τον ασφαλισμένο για ζημία την οποία αυτός πιθανόν να υποστεί ως αποτέλεσμα της αποτυχίας ενός ή περισσοτέρων οφειλετών του να εκπληρώσουν τις υποχρεώσεις τους προς αυτόν (ασφαλισμένο) και καλύπτει τα εξής:

- γενική αφερεγγυότητα,
- εξαγωγικές πιστώσεις (αφορά εξαγωγικές πιστώσεις που δεν γίνονται για λογαριασμό ή με την υποστήριξη του Κράτους),
- πωλήσεις με δόσεις,
- ενυπόθηκες πιστώσεις,
- αγροτικές πιστώσεις

ιε) Κλάδος 15 «Εγγυήσεις»:

Στον ως άνω κλάδο ο ασφαλιστής έναντι ασφαλίστρου εγγυάται για τον ασφαλισμένο την εκτέλεση από αυτόν των συμβατικών του υποχρεώσεων. Περιλαμβάνει:

· άμεσες εγγυήσεις

· έμμεσες εγγυήσεις

ιστ) Κλάδος 16 «Διάφορες χρηματικές απώλειες»:

Καλύπτει διάφορες χρηματικές απώλειες που προκαλούνται από κινδύνους, όπως:

- κίνδυνος απώλειας επαγγελματικής απασχόλησης,
- γενική ανεπάρκεια εισοδήματος,
- κακοκαιρία,
- απώλεια κερδών,
- τρέχοντα γενικά έξοδα,
- απρόβλεπτες εμπορικές δαπάνες,
- απώλεια / μείωση αγοραίας αξίας,
- απώλεια μισθωμάτων ή εισοδημάτων,
- έμμεσες εμπορικές απώλειες εκτός από αυτές που ήδη αναφέρθηκαν
- μη εμπορικές οικονομικές απώλειες,
- λοιπές οικονομικές απώλειες.

ιζ) Κλάδος 17 «Νομική προστασία»:

Περιλαμβάνει την ανάληψη δικαστικών εξόδων και την παροχή νομικής προστασίας.

ιη) Κλάδος 18 «Βοήθεια»:

Περιλαμβάνει την ανάληψη της υποχρέωσης άμεσης παροχής βοήθειας, στις περιπτώσεις και με τους όρους που προβλέπει σύμβαση, σε χρήμα ή σε είδος, έναντι προηγούμενης καταβολής ασφαλίστρου, προς πρόσωπα, που περιέρχονται σε δυσχερή θέση κατά τη διάρκεια μετακινήσεων ή απουσίας από την κατοικία ή από τον τόπο συνήθους διαμονής τους είτε υπό άλλες περιστάσεις ανεξάρτητα από μετακίνηση ή απουσία. Η σε είδος παροχή βοήθειας είναι δυνατόν να συνίσταται και στην χρησιμοποίηση του προσωπικού και του εξοπλισμού που ανήκουν σε αυτόν που παρέχει την βοήθεια. Δεν συνιστούν υπηρεσίες βοήθειας οι υπηρεσίες συντήρησης ή διατήρησης, η εξυπηρέτηση μετά την πώληση, ούτε η απλή υπόδειξη ή πρόβλεψη παροχής βοήθειας ως μεσολάβηση.

Στην παρούσα διπλωματική εργασία θα ασχοληθούμε με τον κλάδο 8 και συγκεκριμένα με την τιμολόγηση των ασφαλιστηρίων συμβολαίων κατά του φυσικού φαινομένου του σεισμού σε κατοικίες.

1.2 Κλάδος Περιουσίας (Πυρκαγιά και στοιχεία της φύσεως)

Ο κλάδος περιουσίας – γνωστός στο ευρύ κοινό ως κλάδος πυρός- είναι αυτός που καλύπτει την επιχείρηση ή την κατοικία από τυχαία και απρόβλεπτα γεγονότα όπως η πυρκαϊά, ο σεισμός, οι ζημιές από φυσικά φαινόμενα ή κλοπή αντικειμένων (τα χρήματα ασφαλίζονται από τον κλάδο κλοπής) κ.α. Η ασφάλιση μπορεί να επεκταθεί και στις οικονομικές συνέπειες από ζημία που οφείλεται σε καλυπτόμενο κίνδυνο όπως είναι η κάλυψη απώλειας κερδών/διακοπής εργασιών.

Με βάση τα στοιχεία της μελέτης που διεξήγαγε η Ένωση Ασφαλιστικών Εταιριών Ελλάδος σχετικά με την παραγωγή ασφαλίστρων την περίοδο Ιανουαρίου-Μαΐου 2021 45 εταιρίες δραστηριοποιήθηκαν στις ασφαλίσεις κατά Ζημιών συγκέντρωσαν το 94,6% της παραγωγής ασφαλίστρων στις ασφαλίσεις κατά Ζημιών και τα εισπραχθέντα ασφάλιστρα για τον κλάδο «Πυρκαϊά και στοιχεία της φύσεως» ανέρχονται στα 132.598.695,48 ευρώ ήτοι 14,6% των συνολικά εισπραχθέντων ασφαλίστρων γεγονός που τοποθετεί τον κλάδο αυτό στην 3η θέση πίσω από τους κλάδους των «Ασθενειών» και «Αστικής ευθύνης χερσαίων οχημάτων» . (Ε.Α.Ε.Ε., 2021)

1.2.1 Ασφάλεια Κατοικίας

Όπως αναφέρθηκε παραπάνω ο κλάδος της ασφάλεια κατοικίας (homeowners insurance) είναι ο τρίτος μεγαλύτερος κλάδος προσωπικής ασφάλισης για τους ασφαλιστές περιουσίας-ευθύνης μετά την ασφάλιση ασθένειας και την ασφάλιση αυτοκινήτου. Τα ασφαλιστήρια συμβόλαια κατοικίας καλύπτουν περισσότερους από ένα κινδύνους (multiple-line, multiple peril policies).

Επιγραμματικά αναφέρεται ότι οι καλύψεις που παρέχουν

είναι οι εξής :

- Κάλυψη της περιουσίας του ασφαλισμένου (first party property coverage)
- Κάλυψη ευθύνης έναντι τρίτων (third-party property coverage)
- Κάλυψη για ιατρικές δαπάνες τρίτων (third-party medical expenses coverage)

Η περιουσία καλύπτεται για ζημιές εξαιτίας πολλών κινδύνων (π.χ. πυρκαγιά, θύελλα, κλοπή). Σε ορισμένες περιπτώσεις το ασφαλιστήριο συμβόλαιο είναι κατανομαζομένων κινδύνων ,δηλαδή όλοι οι κίνδυνοι έναντι των οποίων έχει ασφαλιστεί ο πελάτης της ασφαλιστικής αναφέρονται ρητά στο συμβόλαιο προκειμένου να μειωθούν οι ασάφειες σχετικά με τους κινδύνους που δεν καλύπτουν (named peril policy).

Στα ασφαλιστήρια συμβόλαια κατονομαζόμενων κινδύνων υπάρχουν βέβαια και εξαιρέσεις στις παρεχόμενες καλύψεις. Παράδειγμα εξαιρέσεων αποτελούν οι ζημιές στην περιουσία που προέρχονται από πράξεις εκ προθέσεως, φυσιολογική φθορά, καπνό. Κάποιες εξαιρέσεις μειώνουν την έκθεση του ασφαλισμένου σε γεγονότα που βασίζονται αρκετά στις ενέργειες του ιδιοκτήτη της κατοικίας και επομένως είναι αντικείμενα ηθικού κινδύνου. Άλλες εξαιρέσεις περιορίζουν την κάλυψη για γεγονότα που έχουν υψηλή πιθανότητα εμφάνισης και μικρή αβεβαιότητα(π.χ. φυσιολογική φθορά) και επομένως θα ήταν αρκετά δαπανηρό να ασφαλιστούν λόγω του υψηλού κόστους διαχείρισης της ζημιάς. Επίσης ασφαλιστήρια συμβόλαια έχουν εξαιρέσεις οι οποίες μειώνουν την έκθεση στον κίνδυνο του ασφαλισμένου σε συ σχετιζόμενες ζημίες. Επί παραδείγματι ζημιές που προκλήθηκαν από αλλαγή νομοθεσίας, πλημμύρες, πυρηνικά ατυχήματα, σεισμό και πόλεμο εξαιρούνται. Οι καταναλωτές αν το επιθυμούν μπορούν να ασφαλιστούν έναντι κάποιων από αυτών των κινδύνων μέσω πρόσθετων πράξεων (π.χ., κάλυψη σεισμού) ή μέσω ειδικών κρατικών προγραμμάτων (π.χ., κάλυψη για πλημύρες). (Νεκτάριος, 2010)

Σε άλλες περιπτώσεις ,το ασφαλιστήριο συμβόλαιο καλύπτει όλους τους κινδύνους εκτός από αυτούς που ειδικά εξαιρούνται και είναι γνωστό ως «κατά παντός κινδύνου» (all risk policy).

1.2.2 Τιμολόγηση ασφαλιστηρίων συμβολαίων κατοικίας

Τα ασφάλιστρα των συμβολαίων διαμορφώνονται με βάση δύο βασικά στοιχεία. Το πρώτο είναι η ασφαλισμένη αξία της περιουσίας (ή αλλιώς αξία αντικατάστασης, δηλαδή το κόστος που καταβάλλει ο ασφαλιστής για μια αντίστοιχη κατασκευή ίδιων υλικών) και το άλλο είναι το ασφάλιστρο ανά ευρώ ασφαλισμένης αξίας.

Έστω μια μονοκατοικία με κόστος αντικατάστασης 150.000 € και ο ασφαλιστής κοστολογεί το ασφάλιστρο ίσο με 0,005 ανά 1.000 € κάλυψης το οποίο μας δίνει ένα ασφάλιστρο ίσο με 750 € .

Τα ασφαλιστήρια συμβόλαια αυτά συνήθως προσφέρονται με ένα όριο ίσο με το 100% της αξίας αντικατάστασης της κατοικίας. Οι αναλογιστές χρησιμοποιούν διάφορες μεθόδους για να εκτιμήσουν το κόστος της αντικατάστασης. Χρησιμοποιούν πίνακες που παρέχουν μια εκτίμηση για το κόστος αντικατάστασης για κάθε τ.μ. ,για κάθε συνδυασμούς παραγόντων όπως τύπος κατοικίας ,επιφάνεια ορόφου είδος κατασκευής και υλικά (π.χ. αν το σπίτι είναι χτισμένο με τσιμέντο ή με ξύλο) .Το κόστος αντικατάστασης για κάθε τετραγωνικό μέτρο πολλαπλασιάζεται με την επιφάνεια της κατοικίας, ώστε να βρεθεί το βασικό κόστος αντικατάστασης και το οποίο με τη σειρά του πολλαπλασιάζεται με ένα δείκτη τοποθεσίας (location modifier) και έτσι έχουμε την τελική αξία αντικατάστασης σύμφωνα με τη διακύμανση του κόστους κατασκευής σε διάφορες περιοχές. (Νεκτάριος, 2010)

Για να εξηγήσουμε τον υπολογισμό της αξίας αντικατάστασης αρκεί να φανταστούμε ένα ιδιοκτήτη σπιτιού στην περιοχή του Ελληνικού έκτασης 100 τ.μ.. Ο πίνακας της ασφαλιστικής εταιρίας για την αντικατάσταση κάθε τετραγωνικού μέτρου του σπιτιού ανέρχεται στα 1500 €. Συνεπώς το βασικό κόστος αντικατάστασης είναι 100*1500=150.000.Υποθέτουμε ότι ο δείκτης τοποθεσίας είναι 1,2 άρα το κόστος αποκατάστασης φτάνει τις 180.000 € (150.000*1,2). Αν και αυτός ο τρόπος υπολογισμού είναι αρκετά ακριβής ενέχει τον κίνδυνο σημαντικών λαθών για κατοικίες υψηλής αξίας ή χτισμένες με βάση τις ανάγκες του εκάστοτε ιδιοκτήτη. Συνεπώς, ορισμένοι ασφαλιστές έχουν εφαρμόσει πολυπλοκότερους τρόπους υπολογισμού της αξίας αντικατάστασης. (Νεκτάριος, 2010)

Τα ασφάλιστρα ανά ευρώ αξίας αντικατάστασης για την ασφάλισης κατοικίας ποικίλλουν βάσει ορισμένων παραγόντων. Η τοποθεσία της κατοικίας είναι ένας σημαντικός παράγοντας, διότι καθορίζει την έκθεση της κατοικίας σε κινδύνους που σχετίζονται με τις καιρικές συνθήκες (τυφώνες, σεισμούς, ανεμοστρόβιλους χαλαζοπτώσεις κτλ) μέχρι την εγκληματικότητα και την απόσταση από την πυροσβεστική υπηρεσία. (Νεκτάριος, 2010)

Τα ασφάλιστρα, επίσης ,εξαρτώνται από τα υλικά κατασκευής (π.χ. ξύλινα σπίτια και σπίτια κατασκευασμένα από τούβλο). Στις Η.Π.Α. για την περίπτωση της ασφάλειας κατά του σεισμού ένα σπίτι από τούβλα κοστίζει περίπου $3 με $15 δολάρια ανά $1.000 στις Βορειοδυτικές πολιτείες του Ειρηνικού. Στον αντίποδα στη Νέα Υόρκη το κόστος θα ήταν μεταξύ 60 και 90 σεντς. (Insurance Information Instutute, 2021)

1.3 Ασφάλιση φυσικών καταστροφών

Ως φυσικός κίνδυνος ορίζεται η αβεβαιότητα σχετικά με την επέλευση μιας ζημίας εξαιτίας ενός φυσικού αιτίου. Ένας φυσικός κίνδυνος, μπορεί να προκαλέσει μεγάλη συσσώρευση υλικών ζημιών και σωματικών βλαβών, σε εκτεταμένες γεωγραφικές περιοχές. Σε μια τέτοια περίπτωση, ένας φυσικός κίνδυνος εξελίσσεται σε φυσική καταστροφή. Η εκδήλωση των φυσικών καταστροφών οφείλεται σε γεωλογικούς παράγοντες (σεισμοί, ηφαίστεια, πλημμύρες κατολισθήσεις, τσουνάμι) ή σε ατμοσφαιρικούς παράγοντες (πυρκαγιές φυσικής προέλευσης ,καταιγίδες ,χιόνι, πάγος, ομίχλη). (Νεκτάριος, 2010)

Στατιστικά αποδεικνύεται ότι κατά τη διάρκεια της ζωής ενός ατόμου είναι βέβαιο ότι τουλάχιστον ένας φυσικός κίνδυνος θα επηρεάσει τη ζωή του. Η διεθνής εμπειρία χαρακτηρίζει «καταστροφικές» τις ζημιές που υπερβαίνουν το 2% του ΑΕΠ της χώρας που τις υφίσταται. Κατά την περίοδο 1985-1999 οι απώλειες των πλουσιότερων χωρών συνέπεια φυσικών καταστροφών ανήλθαν στο 2% επί του ΑΕΠ, ενώ στις φτωχές χώρες στο 13%. (Νεκτάριος, 2010)

Οι άνθρωποι δεν αναμένουν την επέλευση ενός φυσικού φαινομένου σε συγκεκριμένο χρόνο και συνεπώς δεν μπορούν να τα ελέγξουν. Παρ’ όλα αυτά, ο ανθρώπινος παράγοντας έχει επίδραση στο περιβάλλον: οι κλιματικές αλλαγές, η υπερθέρμανση του πλανήτη, οι εκτός προδιαγραφών κατασκευαστικές υποδομές και η συγκέντρωση μεγάλου μέρους του πληθυσμού σε περιοχές που είναι επιρρεπείς σε φυσικά φαινόμενα έχουν συντελέσει στην αύξηση της συχνότητας και της σφοδρότητας των φυσικών καταστροφών τα τελευταία χρόνια. Μεγάλες ζημιές που συνέβαιναν κάθε 20 με 30 χρόνια περίπου, τώρα εκδηλώνονται πιο συχνά και μερικές φορές μετράμε δύο τέτοιες ζημιές μέσα σε ένα έτος, σύμφωνα με τις εκτιμήσεις μεγάλων αντασφαλιστών. Μόνο για το έτος 2005 υπολογίζεται ότι το 99,7% επί του συνόλου των καταστροφικών ζημιών παγκοσμίως οφείλεται σε καιρικά φαινόμενα. (Νεκτάριος, 2010)

1.3.1 Ασφαλιστική κάλυψη Καταστροφικών Κινδύνων

Αυτή η ενότητα παρέχει μια επισκόπηση των εναλλακτικών προσεγγίσεων ασφάλισης των καταστροφικών κινδύνων ,περιλαμβάνοντας :

1) την κάλυψη του σεισμού
2) τα ασφαλιστικά προγράμματα για τις πλημμύρες
3) τα ασφαλιστικά προγράμματα για τις περιουσίες στα κέντρα των μεγαλουπόλεων

1.3.1.1 Κάλυψη σεισμού

Αν και η κάλυψη σεισμού (earthquake coverage) εξαιρείται από το βασικό ασφαλιστήριο κατοικίας, οι περισσότεροι ασφαλιστές προσφέρουν την συγκεκριμένη κάλυψη με μια πρόσθετη πράξη (ή ορισμένες φορές με ένα ξεχωριστό ασφαλιστήριο συμβόλαιο). Οι ζημιές συνήθως καλύπτονται πάνω από κάποιο αφαιρετέο ποσό το οποίο ορίζεται ως ένα ποσοστό επί της αξίας αντικατάστασης της περιουσίας. Αυτό σημαίνει ότι αν το κόστος επανακατασκευής του κτηρίου είναι 100.000€ τότε με ένα αφαιρετέο ποσό της τάξης του 2% ο ιδιοκτήτης καλείται να πληρώσει τα πρώτα 2000€ και οι ασφαλιστική εταιρία τα υπόλοιπα (Νεκτάριος, 2010)

Στις Η.Π.Α. σε πολιτείες όπως η Ουάσιγκτον, η Νεβάδα και η Γιούτα οι ασφαλιστές θέτων το ελάχιστο ποσοστό γύρω στο 10%. Συνήθως οι ασφαλισμένοι επιλέγουν υψηλά ποσοστά ιδίας κράτησης προκειμένου να πληρώνουν χαμηλότερα ασφάλιστρα. (Insurance Information Instutute, 2021)

Σε αυτό το σημείο χρήσιμο βέβαια θα ήταν να αναφέρουμε τις προϋποθέσεις παροχής της ασφάλισης που ισχύουν στην Ελλάδα, τις επιπρόσθετες καλύψεις αλλά και σε ποιες περιπτώσεις μια ασφαλιστική μπορεί να αρνηθεί την καταβολή αποζημίωσης :

Προϋποθέσεις ασφάλισης

- Το κτίσμα να είναι κατασκευασμένο μετά το 1960. Σε περίπτωση που το οίκημα έχει οικοδομηθεί πριν το 1960 οι ιδιοκτήτες του οικήματος θα πρέπει να προχωρήσουν μέσω πολιτικού μηχανικού στη σύνταξη έκθεσης τρωτότητας όπου μελετάτε η στατικότητα του κτιρίου και η αντοχή των υλικών κατασκευής.
- Το κτίσμα να διαθέτει όλες τις νόμιμες άδειες οικοδόμησης από τις δημόσιες αρχές
- Να μην έχει υποστεί ζημιές από προγενέστερο σεισμό

(Τόγιας, 2019)

Πρόσθετες / προαιρετικές καλύψεις στην ασφάλεια σπιτιού από σεισμό

Τα ασφαλιστήρια που καλύπτουν τον σεισμό μπορούν να περιλαμβάνουν και πολλές πρόσθετες καλύψεις, όπως:

- Δαπάνες καθαρισμού συντριμμιών
- Αμοιβές μηχανικών
- Δαπάνες έκδοσης αδειών
- Δαπάνες μεταφοράς και στέγασης οικοσκευής
- Δαπάνες στέγασης κατοίκων και χρηστών
- Κάλυψη παγίων εξόδων (πχ δάνειο)
- Απώλεια μισθωμάτων
- Δαπάνες θεμελίωσης

(Cosmote, n.d.)

Περιπτώσεις άρνησης καταβολής αποζημίωσης :

- Με την ύπαρξη προγενέστερων ζημιών στην οικοδομή οι οποίες δεν έχουν αποκατασταθεί.
- Με τη μη τήρηση του αντισεισμικού κανονισμού που ίσχυε κατά τη περίοδο έκδοσης της οικοδομικής αδείας και ανέγερσης του κτιρίου.
- Κατασκευαστικό ελάττωμα που ο ασφαλισμένος γνώριζε αλλά σκόπιμα απέκρυψε από την ασφαλιστική εταιρία.

(Τόγιας, 2019)

Στην Καλιφόρνια οι ασφαλιστές κατοικίας αναγκάζονται δια του νόμου να προσφέρουν ασφάλεια σεισμού μαζί με την ασφάλεια κατοικίας. Οι ιδιοκτήτες μπορούν να επιλέξουν να την αγοράσουν είτε από τον ίδια είτε άλλο ασφαλιστικό πάροχο είτε ακόμη και να την αρνηθούν εντελώς. Την 17η Ιανουαρίου 1994 μετά τον σεισμό στο Νόρθριντζ της Καλιφόρνια οι ασφαλισμένες ζημιές έφτασαν τα $15.3 δισεκατομμύρια δολάρια[1] ($ 28.1 δις σε ισοτιμία δολαρίου εν έτη 2021) ποσό το οποίο ξεπερνούσε τα εισπραχθέντα ασφάλιστρα των προηγούμενων 30 ετών για το συγκεκριμένο κίνδυνο.

Η αδυναμία των ασφαλιστών να αρνηθούν την κάλυψη αυτή μαζί με το ασφαλιστήριο κατοικίας οδήγησε ορισμένους από αυτούς είτε να σταματήσουν την πώληση ασφαλιστηρίων κατοικίας κάτι το οποίο μείωσε τη διαθεσιμότητα της κάλυψης από την ιδιωτική αγορά είτε αυξάνοντας τα αφαιρετέα ποσά από το 10% στο 15% και άνω. Το γεγονός αυτό είχε αντίθετο αποτέλεσμα από το επιθυμητό καθώς στα μέσα του 1996 πυροδότησε μια κρίση η οποία απείλησε την βιωσιμότητα της στεγαστικής αγοράς της Πολιτείας και καθυστέρησε την ανάκαμψή της από την ύφεση.

Το 1996 το νομοθετικό σώμα της Καλιφόρνια ίδρυσε την CEA (California Earthquake Authority) ως μια αρχή δημόσιας διαχείρισης και αλλά κυρίως ιδιωτικής χρηματοδότησης που ενεργούσε μόνο στην Πολιτεία. Η CEA είναι ο μεγαλύτερος πάροχος ασφάλισης κατοικίας κατά του σεισμού στις Η.Π.Α., με 1.113.964 συμβόλαια εν ισχύ στο τέλος του Φεβρουαρίου του 2020. Η CEA παρέχει τα δύο

[1]Για τον υπολογισμό των σημερινών τιμών χρησιμοποιήθηκαν τα site https://www.usinflationcalculator.com/ και https://www.in2013dollars.com/us/inflation/ τα οποία αμφότερα κατέληξαν στο ίδιο αποτέλεσμά

τρίτα των συμβολαίων για το συγκεκριμένο σκοπό , που πωλούνται στην πολιτεία και έχει ετήσιο εισόδημα από τα ασφάλιστρα που ανέρχεται στα $360 εκατομμύρια δολάρια. Το πρόγραμμα είναι αναλογιστικά «υγειές»(εκτιμήθηκε από ειδικούς ότι διαθέτει επαρκή κεφάλαια) με πάνω από 18 δισεκατομμύρια δολάρια διαθέσιμα για πληρωμή αποζημιώσεων από καταστροφικούς σεισμούς. (Insurance Information Institute , 2020)

1.3.1.2 Ασφάλιση πλημμύρας

Όπως αναφέρθηκε νωρίτερα, τα ασφαλιστήρια συμβόλαια κατοικίας εξαιρούν την κάλυψη ζημιών που προκαλούνται από πλημμύρες. Ο κίνδυνος της πλημμύρας, γενικά, εξετάζεται ως μη ασφαλίσιμος σε ιδιωτικό επίπεδο, λόγω της υψηλής συσχέτισης ζημιών, της ενδεχόμενης αντεπιλογής και του ηθικού κινδύνου που ενέχει. Η αδυναμία ασφάλισης του κινδύνου της πλημμύρας στην ιδιωτική ασφαλιστική αγορά οδήγησε τα ενδιαφερόμενα κράτη στην ανάπτυξη κρατικών ασφαλιστικών οργανισμών. (Νεκτάριος, 2010)

Ενδιαφέρον παρουσιάζει η περίπτωση πολλών Πολιτειών των ΗΠΑ οι οποίες κινδυνεύουν από πλημμύρες. Οι κάτοικοι των Πολιτειών αυτών μπορούν μέσω ιδιωτικών ασφαλιστικών εταιριών να εκμεταλλευτούν το Εθνικού Προγράμματος Ασφάλισης Πλημμύρας (National Flood Insurance Program ή NFIP) το διαχειρίζεται η Ομοσπονδιακή Υπηρεσία Διαχείρισης Έκτακτης Ανάγκης και προσφέρεται στο κοινό από ένα δίκτυο περίπου 60 ασφαλιστικών εταιρειών. Οι εταιρείες αυτές εξυπηρετούν αυτά τα ασφαλιστήρια συμβόλαια αλλά αυτή η οποία καλύπτει τον κίνδυνο είναι η Ομοσπονδιακή Κυβέρνηση.

Το NFIP παρέχει ασφάλιση πλημμύρας σε ιδιοκτήτες ακινήτων, ενοικιαστές και επιχειρήσεις και η κάλυψη αυτή τους βοηθά να ανακάμψουν γρηγορότερα όταν υποχωρούν οι πλημμύρες. Συνεργάζεται επίσης με κοινότητες στις οποίες απαιτείται η υιοθέτηση και επιβολή κανονισμών διαχείρισης πλημμυρικών περιοχών που συμβάλλουν στον μετριασμό των επιπτώσεων πλημμύρας.

Η ασφάλιση πλημμύρας είναι διαθέσιμη σε οποιονδήποτε ζει σε μία από τις 23.000 συμμετέχουσες κοινότητες NFIP. Τα σπίτια και οι επιχειρήσεις σε περιοχές πλημμυρών υψηλού κινδύνου με στεγαστικά δάνεια από κρατικούς δανειστές πρέπει να έχουν ασφάλιση πλημμύρας. (Insurance Information Institute, 2021)

Χαρακτηριστικό παράδειγμα αποτελεί ο τυφώνας Katrina ο οποίος από το 2005 μέχρι και σήμερα παραμένει το μεγαλύτερο μεμονωμένο ζημιογόνο γεγονός στην ιστορία της παγκόσμιας ασφαλιστικής αγοράς, προκαλώντας εκτιμώμενες ασφαλισμένες ζημιές της τάξης των $41.1 δισεκατομμυρίων[2] δολαρίων ($57.4 δις σε ισοτιμία δολαρίου εν έτη 2021) και 1.7 εκατομμύρια αναγγελθείσες απαιτήσεις σε 6 Πολιτείες.

Ο αριθμός αυτός δεν περιλαμβάνει τα $16.1 δισεκατομμύρια σε ζημιές από πλημμύρα τα οποία απορροφήθηκαν από το NFIP ή τα $2 με $3 δισεκατομμύρια δολάρια ασφαλισμένων ζημιών σε παράκτιες ενεργειακές εγκαταστάσεις. Οι ζημιές από τον τυφώνα Katrina κατέρριψαν το ρεκόρ του τυφώνα Andrew ο οποίος προκάλεσε $15.5 δισεκατομμύρια δολάρια σε ασφαλισμένες ζημιές με 790.000 αναγγελθείσες απαιτήσεις σε 3 Πολιτείες.(Hartwig & Wilkinson, 2010)

1.3.1.3 Προγράμματα υπολειπόμενων ασφαλιστικών αγορών

Ιδιαίτερη περίπτωση ασφάλισης αποτελούν τα κτήρια στα κέντρα των μεγαλουπόλεων λόγω των υψηλών αναμενόμενων ζημιών. Οι ζημιές αυτές οφείλονται πολλές φορές σε διαδηλώσεις κάτι το οποίο έχει οδηγήσει τις ασφαλιστικές εταιρίες στο να επιλέγουν την μη παροχή ασφάλισης ή στην κατασκευή προϊόντων με μεγάλα επασφάλιστρα ή αφαιρετέες απαλλαγές οι οποίες τα καθιστούν μη ελκυστικά για όσους ιδιοκτήτες αναζητούν ασφάλιση. Σε αυτές τις περιπτώσεις το κράτος πρέπει να παρεμβαίνει ώστε να προσφέρει ασφάλιση στους ανθρώπους αυτούς σε λογικές πάντοτε τιμές. (Νεκτάριος, 2010)

Ένα τέτοιο παράδειγμα σαν αυτό που περιγράφεται ανωτέρω είναι τα προγράμματα FAIR (Fair Access to Insurance Requirements) τα οποία δημιουργήθηκαν στα τέλη του 1960 στις Η.Π.Α. και λειτουργούν μέχρι και σήμερα.

[2] Για τον υπολογισμό των σημερινών τιμών χρησιμοποιήθηκαν τα site https://www.usinflationcalculator.com/ και https://www.in2013dollars.com/us/inflation/ τα οποία αμφότερα κατέληξαν στο ίδιο αποτέλεσμά

Τα προγράμματα FAIR δημιουργήθηκαν για να βοηθήσουν ανθρώπους με ιδιοκτησίες υψηλού κινδύνου να αποκτήσουν ασφάλιση. Σχεδόν κάθε Πολιτεία έχει τέτοια προγράμματα τα οποία είναι ουσιαστικά επιδοτούμενα από τους φορολογούμενους και από τους ιδιωτικούς ασφαλιστές κατοικίας. Αντί μια ασφαλιστική να αναλάβει το ρίσκο του κτηρίου πολλοί συμβαλλόμενοι μαζί επιμερίζονται το ρίσκο. Στην περίπτωση πραγματοποίησης ενός ζημιογόνου γεγονότος το κόστος της απαίτησης θα μοιραστεί σε όλες τις συμμετέχουσες εταιρίες οι οποίες καταβάλλουν εισφορές ανάλογα με το μερίδιο τους στον κλάδο της ασφάλισης περιουσίας. (Moon, 2021)

1.3.2 Μέθοδοι Διαχείρισης Φυσικών Καταστροφών

Οι φυσικές καταστροφές μπορεί να έχουν τεράστιες συνέπειες σε όλες τις κοινωνίες. Ένα γεγονός ανωτέρας βίας μπορεί να αλλάξει τη ζωή ενός ατόμου, την πορεία μιας επιχείρησης, τη δομή ενός ολόκληρου κράτους. Οι ειδικοί τονίζουν ότι οι φυσικές καταστροφές θα κάνουν ολοένα και πιο αισθητή την παρουσία τους λόγω των κλιματικών αλλαγών. Το ενδιαφέρον εστιάζεται στις διάφορες τεχνικές που είναι διαθέσιμες για την διαχείριση των φυσικών καταστροφών, σε μια προσπάθεια να αντιμετωπιστεί τουλάχιστον το οικονομικό βάρος το οποίο επωμίζονται όσοι βιώνουν τα αποτελέσματα τους. Οι μέθοδοι διαχείρισής των φυσικών καταστροφών είναι οι εξής:

1.3.2.1 Ίδια κράτηση-Μέτρα ελέγχου

Μετά την επέλευση του ζημιογόνου γεγονότος αυτοί οι οποίοι βιώνουν άμεσα τις συνέπειές του είναι οι ιδιοκτήτες των κατοικιών ή των επιχειρήσεων διότι ενέχεται ο κίνδυνος όχι μόνο απομείωσης της αξίας της εν λόγω ακίνητης περιουσίας τους αλλά ολοκληρωτικής καταστροφής της. Μια λύση για την αποφυγή ενός τέτοιου ενδεχομένου είναι η επιλογή αρχικής εγκατάστασης σε μια περιοχή στην οποία η συχνότητα και οι σφοδρότητα τέτοιων φυσικών φαινομένων δεν είναι μεγάλη. Έχοντας στο νου την ατομική ευθύνη του καθενός οι ιδιοκτήτες θα πρέπει να σέβονται τους πολεοδομικούς κανόνες και να βελτιώνουν συνεχώς τις συνθήκες του κινδύνου κάτι το οποίο οδηγεί αν όχι στον εκμηδενισμό τουλάχιστον στον μετριασμό ενός μεγάλου μέρους του κινδύνου. Πρέπει βέβαια να επισημάνουμε ότι η λεγόμενη ατομική ευθύνη δεν είναι ανεπτυγμένη στην πλειοψηφία των ανθρώπων (τουλάχιστον όσον αφορά το συγκεκριμένο θέμα) διότι δεν είναι πολλοί αυτοί που επιλέγουν την ασφαλιστική κάλυψη έναντι μιας φυσικής καταστροφής ή την λήψη προληπτικών μέτρων ασφαλείας. Κάτι τέτοιο οφείλεται κατά κύριο λόγο στην αμέλεια ,την έλλειψη κινήτρων και της λανθασμένης εκτίμησης της κατάστασης σε ότι έχει να κάνει σχέση με την πιθανότητα επέλευσης του κινδύνου.

1.3.2.2 Ασφάλιση

Σε ότι αφορά τον ασφαλιστικό κλάδο καταστροφικό για εκείνον θεωρείται το γεγονός κατά το οποίο οι εμπειρικές ζημιές εμφανίζουν μεγάλη απόκλιση από τις εκτιμώμενες. Μια ενδεχόμενη εκτεταμένη καταστροφή θα μπορούσε να οδηγήσει μέχρι και σε κατάρρευση του ασφαλιστικού μηχανισμού. Ένα βασικό πρόβλημα όσον αφορά την ασφάλιση των καταστροφικών κινδύνων είναι η μεγάλη συσχέτισή τους υπό την έννοια ότι ένα ζημιογόνο γεγονός προξενεί ανυπολόγιστες ζημίες σε ευρύτερες γεωγραφικές περιοχές. Ενδέχεται η περίπτωση τα συνολικά εισπραχθέντα ασφάλιστρα να μην επαρκούν για την κάλυψη των συνολικών καταβαλλόμενων αποζημιώσεων .

Σε ότι αφορά τους ασφαλιστές τα τελευταία χρόνια έχουν προχωρήσει στην ανάπτυξη μοντέλων που αφορούν τους καταστροφικούς κινδύνους. Τα μοντέλα αυτά βοηθούν στην βέλτιστη αντιμετώπιση των κινδύνων επιτρέποντας την εκτίμηση της συχνότητας, της έντασης και της γεωγραφικής θέσης ενός καταστροφικού γεγονότος που αναμένεται να εκδηλωθεί στο μέλλον. Για την πιο αποτελεσματική αντιμετώπισή τους οι ασφαλιστές πρέπει να γνωστοποιούν τις πληροφορίες αυτές σε κάθε άμεσα ενδιαφερόμενο αποσκοπώντας στην απόκτηση μιας καλύτερης εικόνας σχετικά με την έκθεσή τους σε καταστροφικούς κινδύνους. Συνεπώς οι ασφαλιστικές επιχειρήσεις σε συνεργασία με το κράτος θα πρέπει να δημιουργήσουν κίνητρα στα άτομα και στις επιχειρήσεις να ασφαλιστούν ώστε να περιοριστεί ο κίνδυνος. (Νεκτάριος, 2010)

Η Ένωσης Ασφαλιστικών Εταιρειών Ελλάδος σχετικά με τις Φυσικές καταστροφές και ταραχές στην Ελλάδα κατά την περίοδο 1993-2018, διεξήγαγε μελέτη της οποίας μέρος των αποτελεσμάτων παρουσιάζονται συνοπτικά στον επόμενο πίνακα:

Αιτία ζημιάς	**Πλήθος ζημιών**	**Ποσό απαίτησης (εκατ.€)**	**Μέση ζημιά (χιλ. €)**
Χιονοπτώσεις	646	2,4	3,7
Βροχοπτώσεις	10.412	128,6	12,4
Δασικές πυρκαγιές	1.542	45,8	12,4
Σεισμός	10.313	133,5	12,9
Ταραχές	1.193	48,5	40,7
Σύνολο	**24.106**	**358,8**	**14,9**

Πηγή: http://www1.eaee.gr/sites/default/files/oikmel_natcat_1993_2018_gr.pdf

Σύμφωνα με τα συγκεντρωτικά στοιχεία της Ένωσης , εκτός από τον μεγάλο σεισμό της Αθήνας το 1999, που επέφερε 9.480 ασφαλισμένες ζημιές, σημαντικές ζημιές προκάλεσαν οι βροχοπτώσεις τον Οκτώβριο του 2014 και του 2015, η μεγάλη πυρκαγιά στο Μάτι τον Αύγουστο του 2018, αλλά και οι βροχοπτώσεις τον Σεπτέμβριο του ίδιου χρόνου τόσο στην Αττική όσο και στην υπόλοιπη Ελλάδα. Η Αττική άλλωστε συγκεντρώνει το μεγαλύτερο πλήθος των ζημιών σε όλη την εξεταζόμενη στη σχετική έρευνα περίοδο. Αυτό γιατί στο Λεκανοπέδιο είναι συγκεντρωμένος ο μεγαλύτερος αριθμός των ασφαλισμένων κατοικιών ή επιχειρήσεων. (Τζώρτζη, 2019)

1.3.2.3 Αντασφάλιση

Οι ασφαλιστικές που επιλέγουν να ασφαλίζουν έναντι φυσικών καταστροφών δεν επιθυμούν να είναι εκτεθειμένοι στο 100% του κινδύνου και έτσι επιλέγουν διάφορες τεχνικές επιμερισμού του. Μια από αυτές τις τεχνικές είναι η αντασφάλιση. Στην πλειοψηφία τους οι ασφαλιστικές εταιρείες υποστηρίζονται από αντασφαλιστές καθώς τα κόστη, καθώς τα μεγέθη των αποζημιώσεων σε περίπτωση πραγματοποίησης του ζημιογόνου γεγονότος είναι τεράστια. Ανάλογα με το μέγεθος του ασφαλιστή και την κεφαλαιακή του επάρκεια θα εκχωρεί όλο και λιγότερο μέρος του κινδύνου στον αντασφαλιστή. Κάτι τέτοιο είναι λογικό καθώς το τίμημα της αντασφάλισης για τον ασφαλιστή είναι ποσοστά επί του καταβληθέντος ασφαλίστρου. Έτσι εταιρείες με μικρή κεφαλαιακή επάρκεια μπορεί να επιλέγουν να αντασφαλίζονται στο 100% του κινδύνου το οποίο τους μετατρέπει σε απλούς διαμεσολαβητές μεταξύ του πελάτη και της αντασφαλιστικής και το μόνο που αποκομίζουν σαν κέρδος είναι τα διάφορα έξοδα διαμεσολάβησης. Λόγω των υπέρογκων ποσών της ενδεχόμενης αποζημίωσης δεν υπάρχουν πολλοί στην αντασφαλιστική βιομηχανία που να προσφέρουν μια τέτοια επιλογή κάτι το οποίο αυξάνει τα απαιτούμενα ασφάλιστρα. Πιο συγκεκριμένα, υπάρχει υψηλή συσχέτιση μεταξύ των αυξημένων ασφαλίστρων σε μια περιοχή παρότι είναι στατιστικά λιγότερο επικίνδυνη από μια άλλη. Αυτό συμβαίνει γιατί τα συγκεντρωθέντα ασφάλιστρα από μια περιοχή που δεν έχει επηρεαστεί από μια φυσική καταστροφή χρησιμοποιούνται για την καταβολή αποζημιώσεων σε μια πληγείσα περιοχή παρότι μπορεί αυτές οι περιοχές να είναι γεωγραφικά αντίθετες. Αυτό συμβαίνει γιατί οι ασφαλιστές και αντασφαλιστές των δύο αυτών περιοχών ανήκουν στην ίδια κοινοπραξία και συνεπώς συμμετέχουν και οι δύο στην καταβολή αποζημιώσεων. (Νεκτάριος, 2010)

1.3.2.4 Ομόλογα Φυσικών Καταστροφών

Η στενότητα της αντασφαλιστικής αγοράς, το υψηλό κόστος της αντασφάλισης και η δυσκολία της εξισορρόπησης των αναμενόμενων ζημιών με μια πραγματικά μεγαλύτερη ζημιά είναι προβλήματα που ζητούν λύση.Το 1992, μετά από την έλευση του τυφώνα Andrew οποίος κατέστρεψε ολοκληρωτικά 63.500 κατοικίες, προκάλεσε ζημιές σε 124.000 κατοικίες, στέρησε τη ζωή σε 65 ανθρώπους και προκάλεσε συνολικές ζημίες που ξεπέρασαν τα 27 δισεκατομμύρια[3] δολάρια ($52,5 δις δολάρια εν έτη 2021) η ασφαλιστική και αντασφαλιστική αγορά έφτασαν στα όρια τους. Την περίοδο αυτή ξεκινάει η αναζήτηση εναλλακτικών τρόπων μεταφοράς των κινδύνων

[3] Για τον υπολογισμό των σημερινών τιμών χρησιμοποιήθηκαν τα site https://www.usinflationcalculator.com/ και https://www.in2013dollars.com/us/inflation/ τα οποία αμφότερα κατέληξαν στο ίδιο αποτέλεσμά

πέρα πάνω από την παραδοσιακή μέθοδο της αντασφάλισης. Το ενδιαφέρον στρέφεται προς τις αγορές χρήματος και κεφαλαίου και η αισιοδοξία είναι έκδηλη λόγω της εξειδίκευσης, του βάθους, της διαφάνειας και της ρευστότητας των διεθνών κεφαλαιαγορών. Οι ασφαλιστές προσφεύγουν στις χρηματοοικονομικές αγορές με την έκδοση αρχικά «Ασφαλιστικών παραγώγων» και αργότερα των «Ομολόγων καταστροφών». (Wikipedia, 2021)

Τα ασφαλιστικά παράγωγα αποσύρθηκαν από την αγορά καθώς οι επενδυτές θεώρησαν ότι οι ασφαλιστές διέθεταν το πλεονέκτημα της εσωτερικής πληροφόρησης καθώς επίσης και λόγω της χαμηλής τους ζήτησης. Παρόλα αυτά διέθεταν χαμηλό συναλλακτικό κόστος και σχετική ευελιξία καθώς ο αρχικός τους σκοπός ήταν η αντιστάθμιση του κινδύνου του ασφαλιστή. Σε δεύτερη φάση ήρθε η τιτλοποίηση των ασφαλιστικών κινδύνων (securitization of catastrophe risk) με την έκδοση από ασφαλιστικές και αντασφαλιστικές εταιρείες προϊόντων εξωχρηματιστηριακής αγοράς (over the counter) όπως τον ομολόγων φυσικών καταστροφών (catastrophe bonds). Το προϊόν αυτό ουσιαστικά υποχρεώνει τον αγοραστή του ομολόγου να διαγράψει από τον εκδότη όλο ή μέρος του επιτοκίου ή του αρχικού ποσού στην περίπτωση πραγματοποίησης του ζημιογόνου γεγονότος. Οι ασφαλιστικές και αντασφαλιστικές εταιρείες δεν αποσκοπούν μόνο στην συσσώρευση κεφαλαίων αλλά και στην διαφοροποίηση του επενδυτικού τους χαρτοφυλακίου (diversification) και στην μεγαλύτερη διασπορά κινδύνων.

Το πλεονέκτημα τέτοιων προϊόντων είναι η παροχή μακροχρόνιας προστασίας καθώς η διάρκειά τους είναι πολυετής σε σχέση με τις αντασφαλιστικές συμβάσεις οι οποίες είναι ετησίως ανανεούμενες και των οποίων οι όροι δεν είναι σταθεροί από χρόνο σε χρόνο καθώς οι συνθήκες στην αγορά μεταβάλλονται. Από την πλευρά των επενδυτών η επιλογή αυτών των ομολόγων είναι συμφέρουσα καθώς στην περίπτωση επέλευσης του κινδύνου δεν λαμβάνουν κεφάλαιο και τόκους, κάτι το οποίο ενέχει σημαντικό ρίσκο, από την άλλη όμως καρπώνονται υψηλές αποδόσεις στα χρεόγραφά τους των οποίων τα επιτόκια δύνανται να ξεπεράσουν κατά πολύ το επιτόκιο Libor.

(Νεκτάριος, 2010)

1.3.2.5 Κρατικοί Ασφαλιστικοί Οργανισμοί

Ανάλογα με το επίπεδο ανάπτυξης της κάθε χώρας λαμβάνονται και διαφορετικά μέτρα για την αντιμετώπιση των φυσικών καταστροφών. Οι πιο ανεπτυγμένες χώρες έχουν πρόσβαση σε ιδιωτικές πηγές χρηματοδότησης και στα οργανωμένα κρατικά ασφαλιστικά συστήματα τα οποία έχουν δημιουργηθεί ακριβώς για την αντιμετώπιση τέτοιων καταστάσεων. Οι υποδομές αυτές έχουν δημιουργηθεί εκ των προτέρων, σε αντίθεση με τις λιγότερο ανεπτυγμένες χώρες οι οποίες συνήθως δεν λαμβάνουν τέτοια προληπτικά μέτρα με αποτέλεσμα να έρχονται αντιμέτωπες με καταστροφικές συνέπειες.

Συνεπώς ο ρόλος του κράτους είναι να λαμβάνει προκαταβολικά μέτρα αντιμετώπισης τέτοιων κρίσεων ,να βελτιώνει πριν την έλευση του γεγονότος τις συνθήκες του κινδύνου και να δίνει κίνητρα στους πολίτες να προσαρμόζουν τη συμπεριφορά τους έτσι ώστε μετά την πραγματοποίηση του οι συνολικές ζημιές να είναι όσο το δυνατό μικρότερες.

Μέσα από τον έλεγχο των κατασκευαστικών υποδομών προκειμένου τα κτήρια να έχουν περισσότερη αντοχή, η επιβολή συγκεκριμένων κατασκευαστικών κανονισμών και η απαγόρευση δόμησης σε περιοχές που δεν είναι κατάλληλες είναι μερικά από τα μέτρα που ο κρατικός μηχανισμός πρέπει να λάβει . (Νεκτάριος, 2010)

Το κράτος θα πρέπει να συμμετέχει στην αντιμετώπιση των φυσικών καταστροφών είτε με το ρόλο του πρωτασφαλιστή είτε με το ρόλο του αντασφαλιστή. Στην πρώτη περίπτωση αναλαμβάνει την ανάληψη του του μεγαλύτερου μέρους του κινδύνου δίνοντας στην ιδιωτική ασφαλιστική αγορά υποστηρικτικό ρόλο. Στον αντίποδα όταν αυτό αναλαμβάνει αντασφαλιστικό ρόλο τότε παρέχει οικονομική ενίσχυση στην ασφαλιστική αγορά. Έμπρακτα παραδείγματα παρέμβασης του κράτους αποτελεί η CEA και το NFIP των οποίων η συνεισφορά έχει αναλυθεί ανωτέρω.

ΚΕΦΑΛΑΙΟ 2: Γενικευμένα Γραμμικά Μοντέλα

2.1 Εισαγωγή στα γραμμικά και γενικευμένα γραμμικά μοντέλα

Το 1972, οι Nelder και Wedderburn ανέπτυξαν μία θεωρία γενικευμένων γραμμικών μοντέλων (GLM) τα οποία ενοποίησαν μεγάλο μέρος της υπάρχουσας θεωρίας της γραμμικής μοντελοποίησης και διεύρυναν το πεδίο εφαρμογής της σε μια ευρεία κατηγορία κατανομών. Συγκεκριμένα, η θεωρία περιλαμβάνει την κατηγορία των μοντέλων logit και probit, τα οποία είναι χρήσιμα μοντέλα που αναπτύχθηκαν για το χειρισμό των διωνυμικών δεδομένων, καθώς και την παλινδρόμηση Poisson και τα γραμμικά μοντέλα Poisson για την ανάλυση πινάκων συνάφειας. Η εισαγωγή γενικευμένων γραμμικών μοντέλων είχε πολύ σημαντικό αντίκτυπο στην πρακτική μοντελοποίηση σε πολλούς κλάδους, ιδίως με την εφαρμογή διαδικασιών για τη δημιουργία μοντέλων στατιστικών λογισμικών όπως GLIM, Genstat, S-Plus και R, SAS, Stata, Systat και SPSS. Στην αναλογιστική επιστήμη η θεωρία έχει χρησιμοποιηθεί για να μοντελοποιήσει προβλήματα που αφορούν την τιμολόγηση του ασφαλίστρου, τη θνησιμότητα και την αποθεματοποίηση των ζημιών. (Boland, 2007)

Αυτό το κεφάλαιο παρέχει μία σύντομη εισαγωγή στα γενικευμένα γραμμικά μοντέλα.

Η μοντελοποίηση σχέσεων μεταξύ παρατηρήσεων (αποκρίσεων) και μεταβλητών αποτελεί την ουσία της περισσότερης στατιστικής έρευνας και ανάλυσης. Πώς επηρεάζεται η συχνότητα εμφάνισης καρκίνου του πνεύμονα από την ηλικία, το φύλο, το κάπνισμα, τη διατροφή και άλλες μεταβλητές; Ο αριθμός των αυτοκινήτων που κατέχει ένας οδηγός σχετίζονται με την ηλικία, την εκπαίδευση, το φύλο, τον τύπο του οχήματος, το μέγεθος του κινητήρα και την καθημερινή χρήση; Με ποιο τρόπο το μέγεθος μιας αξίωσης αποζημίωσης εργοδότη σχετίζεται με τα προσωπικά χαρακτηριστικά του εργαζομένου (ηλικία, φύλο, μισθός) και το εργασιακό περιβάλλον (πρότυπα ασφάλειας, ώρες εργασίας, προωθητικές προοπτικές); Η κατασκευή ερμηνεύσιμων μοντέλων για σύνδεση τέτοιων αποκρίσεων σε μεταβλητές μπορεί συχνά να δώσει πρόσθετη εικόνα για την πολυπλοκότητα της σχέσης που μπορεί συχνά να κρύβεται σε τεράστιο όγκο δεδομένων. Η φειδωλότητα είναι μία πτυχή της μοντελοποίησης που είναι ιδιαίτερα επιθυμητή. Συχνά ένα απλό μοντέλο με λιγότερες επεξηγηματικές μεταβλητές είναι πιο χρήσιμο (ή φειδωλό) από ένα περίπλοκο που στερείται διαίσθησης και είναι πιο δύσκολο να ερμηνευτεί. (Boland, 2007)

Στο κλασσικό γραμμικό μοντέλο, μια τυχαία απόκριση Y διαμορφώνεται μέσω μιας σχέσης του μοντέλου:

$$Y = \boldsymbol{x}^{\boldsymbol{T}}\mathrm{Q} + \varepsilon, \tag{2.1}$$

όπου β είναι ένα διάνυσμα άγνωστων παραμέτρων, το x είναι ένα διάνυσμα γνωστών επεξηγηματικών μεταβλητών και ε είναι το σφάλμα. Δύο θεμελιώδεις πτυχές ενός τέτοιου μοντέλου είναι η γραμμική εξάρτηση της απόκρισης από τις άγνωστες παραμέτρους και η δομή του πρόσθετου σφάλματος. Συχνά κάποιος υποθέτει ότι $E(\varepsilon) = 0$ και πως $Var(\varepsilon)$ είναι σταθερό, οπότε προκύπτει ότι $\mu = E(Y) = x^T\beta$. Στο κλασσικό γραμμικό μοντέλο ο μέσος όρος της Y είναι μία γραμμική συνάρτηση των επεξηγηματικών μεταβλητών. Ειδικά, αν το y_i είναι η i-οστή απόκριση που αντιστοιχεί στο διάνυσμα γνωστών επεξηγηματικών μεταβλητών (ή συμμεταβλητών) $x_i^T = (x_{i1}, \dots, x_{ip})$, τότε η προβλεπόμενη τιμή του μοντέλου για το y_i (χρησιμοποιώντας την έννοια του «καπέλου» για εκτίμηση) δίνεται από το μοντέλο:

$$\hat{y} = \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \equiv \hat{\eta}$$

Μια τέτοια γραμμική σχέση μεταξύ του μέσου μ_i του Y_i και του γραμμικού παράγοντα πρόβλεψης η_i είναι και μαθηματικά προσιτή και συχνά ερμηνεύσιμη, αλλά υπάρχουν πολλές καταστάσεις όπου μπορεί να υπάρχει μία τέτοια γραμμική σύνδεση ή η σύνδεση να είναι ακατάλληλη ή αδύνατη.

Εάν οι άγνωστες παράμετροι β επιτρέπεται να μεταβάλλονται ελεύθερα, τότε οι γραμμικοί παράγοντες πρόβλεψης η_i μπορούν θεωρητικά να λάβουν πραγματικές τιμές εκτός του εύρους της μεταβλητής απόκρισης που μας ενδιαφέρει. Για παράδειγμα, εάν το Y είναι πιθανότητα ή αναλογία, τότε η μέση τιμή του Y θα περιοριστεί σε πεπερασμένο διάστημα. (Boland, 2007)

Υπάρχουν δύο κάπως πιο φυσικές μέθοδοι για να ξεπεραστεί ένα τέτοιο πρόβλημα. Μια μέθοδος είναι να μετατρέψουμε το Y με μια συνάρτηση g και να χρησιμοποιήσουμε γραμμική μοντελοποίηση στο $g(Y)$. Με αυτήν την προσέγγιση, προτείνεται ότι ο μέσος όρος μιας απάντησης Y_i έχει τη μορφή:

$$E(g(Y_i)) = \sum_{j=1}^{p} \beta_j x_{ij}.$$

Για παράδειγμα, η συνάρτηση log έχει χρησιμοποιηθεί εκτενώς σε διάφορες εφαρμογές για τη μετατροπή των δεδομένων σε μια κατά προσέγγιση κανονική μεταβλητή (απόκρισης). Στην ασφάλιση ο λογαριθμικός μετασχηματισμός εφαρμόζεται συχνά σε δεδομένα υψών αποζημιώσεων, λόγω της εγγενούς φύσης των απαιτήσεων να είναι τόσο ασύμμετρες όσο και θετικές. Μια άλλη προσέγγιση στο πρόβλημα είναι η χρήση ενός μετασχηματισμού g για να περιγράψει τη σχέση μεταξύ του μέσου μ του Y, και των γραμμικών συνδυασμών των επεξηγηματικών μεταβλητών. Σε αυτή την περίπτωση, συνδέουμε το μέσο μ_i μιας παρατήρησης Y_i μέσω ενός λειτουργικού μετασχηματισμού g με τον οποίο προκύπτει ότι:

$$E(g(Y_i)) = g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}.$$

Αυτή η προσέγγιση είναι η ουσία της γενικευμένης γραμμικής μοντελοποίησης.

Η λογιστική παλινδρόμηση είναι ένα καλό παράδειγμα γενικευμένου γραμμικού μοντέλου το οποίο δεν είναι ένα γραμμικό μοντέλο με την κλασσική έννοια. Ας υποθέσουμε ότι μας ενδιαφέρει το ποσοστό π_x των ασφαλισμένων ανδρών οδηγών ηλικίας x που θα κάνουν αίτηση αποζημίωσης ατυχήματος το επόμενο έτος. Το Y_x είναι το δείγμα αναλογίας των ασφαλισμένων οδηγών ηλικίας x που κάνουν αίτηση, και είναι σαφές ότι η μέση τιμή $E(Y_x) = \pi_x$ περιορίζεται στο διάστημα [0, 1]. Για να μοντελοποιηθεί η αναλογία π_x ως γραμμική συνάρτηση της μορφής $\pi_x = \beta_0 + \beta_1 x$ για το σχετικό ηλικιακό πεδίο είναι συνήθως πολύ περιοριστική και ακατάλληλη. Δηλαδή, καθώς το x ποικίλλει, $\pi_x = \beta_0 + \beta_1 x$ συχνά λαμβάνει τιμές εκτός του διαστήματος [0,1] και επομένως μπορεί να μην είναι ερμηνευτούν ως ποσοστά (ή πιθανότητες). Το $logit$ μιας πιθανότητας ή ποσοστού π είναι ο λογάριθμος του λόγου πιθανότητας $\pi/(1-\pi)]$, δηλαδή $logit(\pi) = log[\pi/(1-\pi)]$, και αυτό αναφέρεται επίσης ως λογιστική συνάρτηση. Να σημειωθεί ότι η $logit(\pi)$ είναι μια συνεχώς αυξανόμενη παραγωγίσιμη συνάρτηση από (0, 1) στην πραγματική γραμμή αριθμών R. Σε ένα γραμμικό λογιστικό μοντέλο ($logit$ ή λογιστική παλινδρόμηση), ένα μοντέλο $logit(\pi)$ είναι ο γραμμικός συνδυασμός κατάλληλων επεξηγηματικών μεταβλητών. Στο παράδειγμά μας, η συνάρτηση σύνδεσης g μεταξύ της μέσης τιμής $E(Y_x) = \pi_x = \mu_x$ και της γραμμικής πρόβλεψης $\eta_x = \beta_0 + \beta_1 x$, μπορεί να δοθεί από το μοντέλο:

$$g(\pi_x) = \log it(\pi_x) = log\frac{\pi_x}{1-\pi_x} = \beta_0 + \beta_1 x = \eta_x,$$

$$\text{ή} \quad \pi_x = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)} = \frac{\exp(y_x)}{1+\exp(y_x)}$$

Οι πρώτες χρήσεις των γραμμικών λογιστικών μοντέλων ήταν σε πειράματα βιολογικής ανάλυσης και ανάλυσης των δεδομένων έρευνών, αλλά είναι πλέον ένα κοινό εργαλείο σε πολλούς κλάδους όπου τα αναλογικά δεδομένα παρουσιάζουν ενδιαφέρον. Για παράδειγμα, μπορεί να εξετάσουμε ένα μοντέλο $logit$ εάν μας ενδιαφέρει να μοντελοποιήσουμε πώς η πιθανότητα ενός ατόμου να αναπτύξει κρίσιμη ασθένεια (ή να έχει τροχαίο ατύχημα) σχετίζεται με την ηλικία, το φύλο και διάφορους άλλους παράγοντες υγείας ή κινδύνου. Σε μελέτες θνησιμότητας, το αρχικό ποσοστό θνησιμότητας q_x στην ηλικία x έχει διαμορφωθεί με βάση τη λογιστική παλινδρόμηση. (Boland, 2007)

Η κατανομή Poisson με μέση τιμή μ είναι μία βασική και σημαντική κατανομή για την καταμέτρηση των δεδομένων, αλλά επίσης έχει τη γνωστή ιδιότητα πως η μέση τιμή και η διακύμανση είναι ίσες. Επομένως, εάν ο μέσος όρος μιας μεταβλητής απόκρισης Poisson εξαρτάται γραμμικά από ένα διάνυσμα x γνωστών επεξηγηματικών μεταβλητών, τότε η διακύμανσή του είναι μη σταθερή. Συνήθως, κάποιος μπορεί να επιχειρήσει να μοντελοποιήσει αυτήν την κατάσταση μέσω της κλασσικής προσέγγισης της γραμμικής μοντελοποίησης είτε μετατρέποντας τη μεταβλητή απόκρισης είτε χρησιμοποιώντας μια προσέγγιση σταθμισμένων ελάχιστων τετραγώνων όπου τα βάρη υπολογίζονται (μέσω επαναλαμβανόμενων μεθόδων) από τα δεδομένα.

Η γενικευμένη προσέγγιση γραμμικού μοντέλου θα ξεκινούσε σημειώνοντας ότι εάν το Y είναι "Poisson" με μέσο μ, τότε αν και το μ είναι θετικό, μπορεί να μην είναι λογικό να εκφραστεί ως γραμμικός συνδυασμός μεταβλητών πρόβλεψης. Ωστόσο, μπορεί κανείς να χρησιμοποιήσει ένα (πολλαπλασιαστικό) λογαριθμικό μοντέλο για την κατανομή Poisson, όπου μοντελοποιούμε:

$$g(\mu) = log\mu = \eta = \boldsymbol{x}^{\boldsymbol{T}}\mathrm{Q}, \text{ ή ισοδύναμα, } \mu = \varepsilon^{y}$$

Εδώ η συνάρτηση που συνδέει τη μέση τιμή $E(Y) = \mu$ με τη γραμμική πρόβλεψη είναι η λογαριθμική συνάρτηση, η οποία διαφοροποιείται συνεχώς από (0, ∞) στη γραμμή των πραγματικών αριθμών R. Αυτό το παράδειγμα ενός γενικευμένου γραμμικού μοντέλου είναι επίσης γνωστό ως (γραμμική) παλινδρόμηση Poisson. Η παλινδρόμηση Poisson μπορεί να χρησιμοποιηθεί όταν θέλουμε να μελετήσουμε πώς ο αριθμός των απαιτήσεων που υποβάλλονται από ένα ασφαλισμένο άτομο εξαρτάται από διάφορες επεξηγηματικές μεταβλητές ή προβλέψεις ή πώς ο αριθμός των ατυχημάτων σε μια διασταύρωση εξαρτάται από τον καιρό, την κυκλοφορία, την ώρα της ημέρας ή της ημέρας της εβδομάδας. Στη μελέτη της θνησιμότητας, η ένταση θνησιμότητας μ_x μπορεί να μοντελοποιηθεί μέσω παλινδρόμησης Poisson. (Boland, 2007)

Η μέθοδος ανάλυσης $probit$ είναι ένα από τα πρώτα παραδείγματα γενικευμένου γραμμικού μοντέλου που δεν είναι γραμμικό με την κλασσική έννοια. Έχει τις ρίζες του στο έργο του Bliss το 1935 σε βιολογική έρευνα, όπου η θνησιμότητα διαμορφώνεται σε επίπεδο δόσης. Ας υποθέσουμε ότι μας ενδιαφέρει η αναλογία Y_x των ατόμων που επιβιώνουν από ένα φάρμακο ή μία τοξίνη που χορηγείται σε γνωστό επίπεδο δόσης x (το οποίο συχνά μετράται σε λογαριθμικές μονάδες) και

$\pi_x = E(Y_x)$ υποδηλώνει τη μέση τιμή του Y_x. Στο μοντέλο $probit$, χρησιμοποιείται το αντίστροφο Φ^{-1} της αθροιστικής συνάρτησης κατανομής της τυπικής κανονικής κατανομής που συνδέει τη μέση τιμή $\pi_x = E(Y_x) = \mu_x$ και το γραμμικό μοντέλο πρόβλεψης $\eta_x = \beta_0 + \beta_1 x$, όπου

$$g(\pi_x) = \Phi^{-1}(\pi_x) = \eta_x, \text{ ή } \pi_x = \Phi(\beta_0 + \beta_1 x) = \Phi(\pi_x)$$

Σημειώνεται ότι ανεξάρτητα από τις τιμές των β_0 και β_1, το $\Phi(\beta_0 + \beta_1 x)$ θα εξακολουθεί να είναι ένας αριθμός στο διάστημα [0,1] - είναι μια πιθανότητα. Η συνάρτηση $g = \Phi^{-1}$ είναι μια συνεχώς παραγωγίσιμη‘‘1 – 1’’ συνάρτηση στο (0, 1) στη γραμμή των πραγματικών αριθμών R.

Αυτή η σύντομη εισαγωγή δείχνει πώς η θεωρία των γενικευμένων γραμμικών μοντέλων επεκτείνει το κλασσικό γραμμικό μοντέλο και περιλαμβάνει επίσης τα μοντέλα $logit$, Poisson και $probit$ ως ειδικές περιπτώσεις. Τα παρακάτω σημεία συνοψίζουν μερικούς από τους κύριους τρόπους με τους οποίους αυτή η θεωρία γενικεύει την κλασσική προσέγγιση της γραμμικής μοντελοποίησης:

- Η σχέση μεταξύ του μέσου μ_i μιας τυχαίας παρατήρησης Y_i και άλλων επεξηγηματικών μεταβλητών $x_{i1}, \ldots, x_{ip}$ μπορεί να είναι γενικότερη από αυτή στο γραμμικό μοντέλο. Αντί να υποθέσουμε ότι $\mu_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ (δηλαδή, ο μέσος όρος είναι ένας γραμμικός συνδυασμός των μεταβλητών πρόβλεψης), υποθέτουμε ότι υπάρχει μια συνάρτηση g (που ονομάζεται συνάρτηση σύνδεσης) που συνδέει το μέσο μ_i με τον γραμμικό παράγοντα πρόβλεψης $\eta_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ μέσω της $g(\mu_i) = \eta_i$.
- Η διακύμανση μιας παρατήρησης μπορεί να εξαρτάται από το μέσο όρο. Στην πραγματικότητα, θα δειχθεί ότι $Var(Y_i) = (\varphi | A_i)\ V(\mu_i)$, όπου το A_i είναι το γνωστό εκ των προτέρων βάρος και το φ είναι μια παράμετρος κλίμακας ανεξάρτητη από το i. Το $V(\mu)$ ονομάζεται συνάρτηση διακύμανσης και δείχνει πώς η διακύμανση μιας παρατήρησης εξαρτάται από το μέσο όρο της.
- Η κατανομή της μεταβλητής απόκρισης Y μπορεί να είναι οποιοδήποτε μέλος της εκθετικής οικογένειας κατανομών. Αυτή η οικογένεια πιο συγκεκριμένα περιλαμβάνει τις Κανονική, Poisson, Διωνυμική και Γάμμα κατανομές.

Η ανάπτυξη του γενικευμένου γραμμικού μοντέλου ξεκινά με μια ανασκόπηση στην Ενότητα 2.2 των κανονικών γραμμικών μοντέλων. Τα γενικευμένα γραμμικά μοντέλα επεκτείνονται από τη κατηγορία γραμμικών μοντέλων σε μια κατηγορία κατανομών γνωστών ως εκθετικές οικογένειες.

2.2 Πολλαπλή γραμμική παλινδρόμηση και το γραμμικό μοντέλο

Στο κλασσικό κανονικό γραμμικό μοντέλο υποθέτουμε ότι $Y = Y_1 + \cdots + Y_n$ είναι ανεξάρτητες κανονικές τυχαίες μεταβλητές, όπου $Y_i \sim N(\mu_i, \sigma^2)$ και

$\mu = E(Y) = x^T\beta$.

Συγκεκριμένα, το μοντέλο έχει τη μορφή $Y = X\beta + \varepsilon$, όπου

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},\ X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix},\ \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},\ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Οι όροι σφάλματος ε_i είναι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την κατανομή $N(0,\ \sigma^2)$. Μια παρατήρηση Y_i διαιρείται σε συνιστώσες που συμβολίζονται με $x_i^T\beta = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ και ε_i ένα τυχαίος παράγοντας. Το κομμάτι $(\eta_i = x_i^T\beta)$ είναι μια γραμμική πρόβλεψη για το $E(Y_i)$. Το X συνήθως αναφέρεται ως ο $n\ x\ p$ πίνακας σχεδιασμού, όπου n είναι το πλήθος των τυχαίων παρατηρήσεων και p είναι ο αριθμός των επεξηγηματικών μεταβλητών.

Εδώ, οι επεξηγηματικές μεταβλητές p μπορεί να είναι κατηγορικές, συνεχείς ή συνδυασμός και των δύο, και αναφερόμαστε σε αυτήν την κατηγορία μοντέλων ως το γενικό γραμμικό μοντέλο. Εάν υπάρχει σταθερά στο γραμμικό μοντέλο, τότε η πρώτη στήλη του πίνακα σχεδιασμού X αποτελείται από μονάδες. Το πιο βασικό τέτοιο μοντέλο είναι το απλό μοντέλο γραμμικής παλινδρόμησης όπου

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

Ο όρος ανάλυση διακύμανσης ή ANOVA χρησιμοποιείται συχνά για να περιγράψει ένα μοντέλο όπου οι επεξηγηματικές μεταβλητές είναι όλες κατηγορικές. Συνήθως, θέλουμε να συγκρίνουμε ομάδες που ορίζονται από διαφορετικά επίπεδα αυτών των κατηγορικών μεταβλητών ή παραγόντων. Για παράδειγμα, μπορεί να θέλουμε να συγκρίνουμε το log μίας απαίτησης (υποθέτοντας ότι οι αξιώσεις κατανέμονται σύμφωνα με τη λογαριθμοκανονική) σε σχέση με το φύλο και τη διεύθυνση κατοικίας ενός οδηγού. Όταν υπάρχει ένα μείγμα συνεχών και κατηγορικών μεταβλητών στο κανονικό γραμμικό μοντέλο, συχνά χρησιμοποιείται ο όρος ανάλυση συνδιακύμανσης ή ANCOVA για να το περιγράψει. Σε ένα τέτοιο μοντέλο, ενδιαφερόμαστε να συγκρίνουμε τις ομάδες που ορίζονται από διαφορετικά επίπεδα κατηγορικών μεταβλητών, αλλά θέλει επίσης να ενσωματώσει την πιθανότητα η απάντηση μέσα σε μια ομάδα να χρειαστεί προσαρμογή για ορισμένες συμμεταβλητές ή συνεχείς μεταβλητές. (Boland, 2007)

Σε ένα κανονικό γραμμικό μοντέλο, η μέση τιμή $E(Y_i) = \mu_i$ διαμορφώνεται ως γραμμική συνάρτηση των επεξηγηματικών μεταβλητών με συντελεστές β. Η παράμετρος σ^2 (διακύμανση) αναφέρεται συχνά ως παράμετρος όχλησης, αλλά φυσικά μπορεί να είναι χρήσιμη στη δημιουργία συμπερασμάτων σχετικά με το μ_i. Για ένα δείγμα παρατηρήσεων y, η συνάρτηση πιθανότητας παίρνει τη μορφή

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T \text{Q})^2\right\}.$$

Είναι γνωστό πως οι μέγιστες εκτιμήσεις των Q και σ^2 δίνονται από τους τύπους

$$\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \quad \text{και}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T \hat{\text{Q}})^2$$

Συνήθως στο κανονικό γραμμικό μοντέλο οι εκτιμήσεις των παραμέτρων Q λαμβάνονται με τη μέθοδο των ελαχίστων τετραγώνων, ελαχιστοποιώντας την παρακάτω συνάρτηση σε σχέση με το Q:

$$\sum_{i=1}^{n}(y_i - x_i^T \text{Q})^2,$$

Αυτό δίνει αποτελέσματα πανομοιότυπα με αυτά που χρησιμοποιείται στη μέθοδο μεγίστης πιθανοφάνειας, και αυτοί οι εκτιμητές είναι αμερόληπτοι. Για την εκτίμηση του $\boldsymbol{\sigma}^2$, χρησιμοποιείται συνήθως ο διαιρέτης $(n - p)$ (ο αριθμός των βαθμών ελευθερίας) αντί για το n. Ο τελικός εκτιμητής είναι ένας αμερόληπτος εκτιμητής του $\boldsymbol{\sigma}^2$, και συνήθως καλείται μέσο άθροισμα τετραγώνων των υπολοίπων ή άθροισμα μέσων τετραγώνων. Το γνωστό θεώρημα των Gauss – Markov αναφέρει ότι οι ελάχιστες εκτιμήσεις τετραγώνων (ή η μέγιστη πιθανοφάνεια) είναι οι καλύτερες γραμμικές αμερόληπτες εκτιμήσεις του Q (στις παρατηρήσεις y_i).

Συχνά υπάρχουν πολλές πιθανές επεξηγηματικές μεταβλητές που μπορεί να συμπεριληφθούν σε ένα μοντέλο, και επομένως είναι σημαντικό να προσδιοριστεί ένα υποσύνολο που εξηγεί ένα μεγάλο ποσοστό της διακύμανσης της μεταβλητής απόκρισης, αλλά οδηγεί σε βέλτιστο μοντέλο. Στην πολλαπλή γραμμική παλινδρόμηση, μπορεί κανείς να κάνει χρήση της μεθόδου step-wise παλινδρόμησης για τον προσδιορισμό κατάλληλων μεταβλητών. Ουσιαστικά, κάποιος χρησιμοποιεί μια ανάλυση του πίνακα διακύμανσης για ένα πιθανό μοντέλο, αποφασίζοντας με βάση τα μέσα αθροίσματα τετραγώνων και την κατανομή πιθανότητας χ^2 εάν μια συγκεκριμένη υποψήφια μεταβλητή (ή σύνολο μεταβλητών) πρέπει να προστεθεί ή όχι σε ένα υπάρχον μοντέλο. Σε γενικευμένα γραμμικά μοντέλα, ωστόσο, κάποιος χρησιμοποιεί την έννοια της απόκλισης για να αποφασίσει ποιες μεταβλητές πρέπει να συμπεριληφθούν σε ένα μοντέλο. Η συνάρτηση απόκλισης στα GLM είναι μία από τις σημαντικότερες εφαρμογές της θεωρίας της πιθανότητας μέσω της χρήσης του στατιστικού λόγου πιθανότητας. Η απόκλιση μπορεί επίσης να χρησιμοποιηθεί για τη μέτρηση της έλλειψης προσαρμογής ενός μοντέλου. (Boland, 2007)

Η ανάλυση υπολοίπων είναι φυσικά ένα βασικό εργαλείο της μοντελοποίησης. Αν το $\hat{\mu}_i$ είναι μία πρέπουσα τιμή για ένα δεδομένο μοντέλο που αντιστοιχεί στο x_i^T, τότε το i-οστό υπόλοιπο δίνεται από τη συνάρτηση:

$$\varepsilon_i = y_i - x_i^T\beta = y_i - \hat{\mu}_i.$$

Οι ασυνήθιστες παρατηρήσεις έχουν γενικά μεγάλα υπόλοιπα, αλλά είναι συχνά κατατοπιστική η τυποποίηση τους για τον εντοπισμό τάσεων ή προτύπων που μπορεί να προκαλέσουν ενδιαφέρον ή να προβληματισμό. Το συνηθισμένο i-οστό τυποποιημένο υπόλοιπο δίνεται από τον τύπο:

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{(1-h_{ii})}}.$$

Όπου το h_{ii} είναι το i-οστό διαγώνιο στοιχείο στον πίνακα εκτιμώμενων συντελεστών $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. (Boland, 2007)

2.3 Η δομή των γενικευμένων γραμμικών μοντέλων

2.3.1 Εκθετικές οικογένειες

Η έννοια της εκθετικής οικογένειας είναι θεμελιώδης για τη θεωρία των γενικευμένων γραμμικών μοντέλων, αλλά δυστυχώς ο συμβολισμός που χρησιμοποιείται ποικίλλει σημαντικά.

Πολλές από τις πιο χρήσιμες μονοπαραμετρικές κατανομές που χρησιμοποιούνται στη μοντελοποίηση ανήκουν σε αυτό που ονομάζεται γενική εκθετική οικογένεια. Η τυχαία μεταβλητή Y έχει κατανομή που ανήκει σε εκθετική μονοπαραμετρική οικογένεια εάν έχει συνάρτηση πυκνότητας (ή πιθανότητας) f_Y που μπορεί να εκφραστεί με τη μορφή

$$f_Y(y;\theta;) = \exp\left(\frac{A[y-\gamma()]}{\varphi} + r\left(y, \frac{\varphi}{A}\right)\right) \qquad (2.2)$$

ή ισοδύναμα, ο λογάριθμος τής πυκνότητας (συνάρτησης πιθανότητας) $l(\theta)$ μπορεί να εκφραστεί ως:

$$l(\theta) = \log(\theta,;y) = log f_Y(y;\theta,) = \frac{A[y-\gamma()]}{\varphi} + r\left(y, \frac{\varphi}{A}\right) \qquad (2.3)$$

Η παράμετρος θ αναφέρεται ως φυσική (ή κανονική) παράμετρος, και σχετίζεται με το μέσο μέσω του $E(Y) \equiv \mu = \gamma'(\theta)$. Το είναι μια παράμετρος κλίμακας ή παράμετρος διασποράς που μπορεί να είναι γνωστή και σχετίζεται με τη διακύμανση του Y μέσω του τύπου $Var\ (Y) = \gamma''(\theta)/A$. Αυτό επιτρέπει στη διακύμανση του Y να διαφοροποιείται από το μέσο όρο ($\mu = \gamma'(\theta)$) με την παράμετρο διασποράς , και αυτό είναι μια σημαντική ιδιότητα των εκθετικών οικογενειών στη γενικευμένη γραμμική μοντελοποίηση. Τα Y και r είναι γνωστές συναρτήσεις και το A είναι γνωστό εκ των προτέρων βάρος. (Boland, 2007)

Η συνάρτηση U είναι η μερική παράγωγος της λογαριθμικής συνάρτησης πιθανότητας l του Y σε σχέση με το θ, και συμβολίζεται με $U(\theta) = (\partial/\ \partial\theta)l$. Ως συνάρτηση του Y, έχει μέσο μηδέν αφού

$$E_Y\left(\frac{\partial}{\partial} l\right) = \int f_Y \frac{\partial}{\partial} \log f_Y\, dy = \int f_Y \frac{1}{f_Y}\frac{\partial f_Y}{\partial} dy = \int \frac{\partial f_Y}{\partial} dy = \frac{\partial}{\partial}\int f_Y\, dy = \frac{\partial}{\partial} 1 = 0. \quad (2.4)$$

Χρησιμοποιώντας τον τύπο (2.4) κάποιος μπορεί να πει ότι

$$E_Y\left(\frac{\partial}{\partial\theta} l\right) = E_Y\left(\frac{A}{\varphi}[Y - \gamma'(\theta)]\right),$$

που σε συνδυασμό με το (2.4) συνεπάγεται ότι για κατανομές στην εκθετική οικογένεια έχουμε $E(Y) = \mu = \gamma'(\theta)$.

Επίσης,

$$E_Y\left[\frac{\partial^2}{\partial\theta^2} l + \left(\frac{\partial}{\partial\theta} l\right)^2\right] = \int f_Y \left[\frac{\partial}{\partial\theta}\left(\frac{1}{f_Y}\frac{\partial f_Y}{\partial\theta}\right) + \left(\frac{1}{f_Y}\frac{\partial f_Y}{\partial \text{P}}\right)^2\right] dy$$

$$= \int f_Y \frac{1}{f_Y^2}\left[-\left(\frac{\partial f_Y}{\partial\theta}\right)^q + \frac{\partial^2 f_Y}{\partial\theta^2} f_Y + \left(\frac{\partial f_Y}{\partial\theta}\right)^2\right] dy = \int \frac{\partial^2 f_Y}{\partial\theta^2} dy$$

$$= \frac{\partial^2}{\partial\theta^2}\int f_Y\, dy = \frac{\partial^2}{\partial\theta^2} 1 = 0.$$

Η πληροφορία του Fisher $I(\theta)$ στο Y για θ προσδιορίζεται από $I(\theta) = E\left(\frac{\partial}{\partial\theta} l\right)^2$, και για αυτό από τα παραπάνω έχουμε δύο διαφορετικούς τρόπους για να γράφουμε τον τύπο αυτό

$$I(\theta) = E\left(\frac{\partial}{\partial\theta} l\right)^2 = -E\left(\frac{\partial^2}{\partial^2} l\right).$$

Επίσης, επακόλουθα

$$E\left(\frac{\partial^2}{\partial^2} l\right) = -\frac{\gamma''()}{\varphi/A} \qquad \text{και} \qquad E\left(\frac{\partial}{\partial\theta} l\right)^2 = E\left(\frac{Y-\gamma'()}{\varphi/A}\right)^2 = \frac{Var(Y)}{(\frac{\varphi}{A})^2}$$

Και συνεπώς,

$$Var(Y) = /A\, \gamma''(\theta) \equiv /A\, V(\mu)$$

Χρησιμοποιούμε το $Var(Y) = \gamma''(\theta)$ για να υποδηλώσουμε πως η διακύμανση της συνάρτηση σχετίζεται με την οικογένεια. Να σημειωθεί ότι όταν $/A = 1$, η συνάρτηση $\gamma(\theta)$ έχει την ιδιότητα $E(Y) = \gamma'(\theta)$ και $Var(Y) = \gamma''(\theta) = V(\mu)$. Συνεπώς, το $\gamma(\theta)$ κάποιες φορές αναφέρεται ως συνάρτηση κατανομής του Y. (Boland, 2007)

2.3.2 Συναρτήσεις σύνδεσης και γραμμικοί παράγοντες πρόβλεψης

Στα GLM, ορίζουμε μια σχέση μεταξύ της μέσης απόκρισης $E(Y) = \mu$ και ενός γραμμικού συνδυασμού επιλεγμένων επεξηγηματικών μεταβλητών $\boldsymbol{x^T}$ μέσω μίας συνάρτησης g, όπου:

$$g(E(Y)) = \boldsymbol{x^T}\mathrm{Q} = \eta.$$

Μια τέτοια συνάρτηση g ονομάζεται συνάρτηση σύνδεσης, και παρόλο που υπάρχουν πολλές δυνατότητες να χρησιμοποιηθούν ως σύνδεσμοι σε κάθε δεδομένη κατάσταση, για κάθε δεδομένη εκθετική οικογένεια υπάρχει μια κανονική συνάρτηση σύνδεσης. Στη συνάρτηση (2.2) για τη συνάρτηση πυκνότητας (ή πιθανότητας) μιας τυχαίας μεταβλητής Y η οποία ανήκει σε μια εκθετική μονοπαραμετρική οικογένεια κατανομών, αναφέρεται η παράμετρος θ ως φυσική ή κανονική. Αυτή η παράμετρος θ θα σχετίζεται με το μέσο όρο μ του Y μέσω κάποιας συνάρτησης g_N όπου $g_N(\mu) = \theta$. Εάν χρησιμοποιήσουμε αυτήν τη συνάρτηση g_N ως σύνδεσμό μας με το γραμμικό παράγοντα πρόβλεψης στο μοντέλο, τότε ο σύνδεσμός μας ονομάζεται κανονικός ή φυσικός σύνδεσμος. Ο παρακάτω πίνακας δίνει τις κανονικές συναρτήσεις σύνδεσης και διακύμανσης ορισμένων κατανομών. (Boland, 2007)

Κανονική σύνδεση και διασπορά συναρτήσεων.

Οικογένεια	Κανονική σύνδεση	Όνομα	Διακύμανση V(μ)	Κλίμακα
Κανονική	μ	$\tau a u r o r \iota \kappa \eta$	1	σ^2
Poisson	$log\mu$	log	μ	1
Διωνυμική	$log[\mu/(1-\mu)]$	$logit$	$\mu(1-\mu)$	1
Γάμμα	$-1/\mu$	$\alpha\nu\tau\acute{\iota}\sigma\tau\rho o\eta$	μ^2	$1/\alpha$

Δεν είναι υποχρεωτική η χρήση των παραπάνω κανονικών συναρτήσεων σύνδεσης κατά τη χρήση μίας συγκεκριμένης εκθετικής οικογένειας. Παράγει λογικά αποτελέσματα τις περισσότερες φορές. Ωστόσο, κάθε περίπτωση θα πρέπει να κρίνεται με βάση τα πλεονεκτήματά της. Για τη διωνυμική οικογένεια η συνάρτηση probit ($g(\mu) = \Phi^{-1}(\mu)$) ή η συμπληρωματική συνάρτηση log log (($g(\mu) = log\{-log\ (1-\mu)\}$) χρησιμοποιούνται συχνά. Για την οικογένεια Poisson, αντί για τον (κανονικό) log σύνδεσμο, μερικές φορές χρησιμοποιείται η ταυτοτική συνάρτηση σύνδεσης ($g(\mu) = \mu$) ή τετραγωνικής ρίζας ($g(\mu) = \sqrt{\mu}$). Για την οικογένεια Γάμμα κατανομής, είναι σύνηθες να χρησιμοποιείται ο σύνδεσμος $1/\mu$ αντί για $-1/\mu$, και σε ορισμένες περιπτώσεις χρησιμοποιείται ο ταυτοτικός σύνδεσμος ή ο σύνδεσμος log. Σε γενικές γραμμές, απαιτείται η συνάρτηση σύνδεσης g να είναι συνάρτηση ‘’1-1’’ η οποία μπορεί να παραγωγηθεί. Αυτή είναι μια τεχνική απαίτηση η οποία είναι απαραίτητη για να διασφαλιστεί ότι η αριθμητική διαδικασία που χρησιμοποιείται για τον υπολογισμό των εκτιμήσεων μέγιστης πιθανοφάνειας λειτουργεί σωστά. (Boland, 2007)

Έχοντας αποφασίσει πως κατανέμονται τα δεδομένα και έχοντας μια συνάρτηση σύνδεσης g, αποφασίζεται ποιες από τις διαθέσιμες επεξηγηματικές μεταβλητές (παράγοντες και συμμεταβλητές) έχουν σημαντική επίδραση στην ανταπόκριση και ως εκ τούτου θα πρέπει να συμπεριληφθούν σε ένα καλό μοντέλο (αυτό γίνεται συχνά συγκρίνοντας πολλά πιθανά μοντέλα). Αλληλεπιδράσεις, είτε μεταξύ παραγόντων ή συμμεταβλητών, είτε μεταξύ ενός παράγοντα και μιας συμμεταβλητής μπορεί να εμφανίζονται επίσης στον παράγοντα πρόβλεψης. Ο γραμμικός δείκτης πρόβλεψης είναι $\eta = \boldsymbol{x}^{\boldsymbol{T}}\mathrm{Q}$, όπου οι παράμετροι Q πρέπει να εκτιμηθούν (για τα GLM, αυτό γίνεται μέσω της μέγιστης πιθανοφάνειας). Στη συνέχεια, για κάθε δεδομένο σύνολο επεξηγηματικών μεταβλητών, μπορούμε να καθορίσουμε την αντίστοιχη γραμμική πρόβλεψη ή και να προβλέψουμε την τιμή της μεταβλητής απόκρισης να είναι $g^{-1}(\eta) = E(Y)$. Πριν προχωρήσουμε περαιτέρω, ορίζουμε ακριβώς τι εννοούμε με τους παράγοντες, τις μεταβλητές και τις αλληλεπιδράσεις. (Boland, 2007)

2.3.3 Παράγοντες και συνδιακυμάνσεις

Οι συντελεστές είναι κατηγορικές μεταβλητές και οι τιμές που λαμβάνει ένας παράγοντας ονομάζονται επίπεδα. Για παράδειγμα, το φύλο είναι ένας παράγοντας με δύο επίπεδα (άντρας/γυναίκα), το αν καπνίζει κάποιος είναι ένας παράγοντας με δύο επίπεδα (ναι/όχι) και το επίπεδο εκπαίδευσης είναι ένας παράγοντας με τέσσερα επίπεδα (πρωτοβάθμια, δευτεροβάθμια, τριτοβάθμια και ακαδημαϊκή). Σε ένα μοντέλο ενός παράγοντα πρέπει να εκτιμήσουμε μία παράμετρο για κάθε επίπεδο που λαμβάνει ο παράγοντας. Για κάθε επιπλέον συντελεστή B (ας πούμε με τα επίπεδα b) που προσθέτουμε στο μοντέλο, πρέπει να εκτιμήσουμε επιπλέον $b - 1$ παραμέτρους. Αναφερόμαστε συλλογικά στις κύριες επιδράσεις του παράγοντα στις παραμέτρους που αντιπροσωπεύουν έναν παράγοντα και δεν εμπλέκονται σε αλληλεπίδραση. Οι συμμεταβλητές είναι αριθμητικές μεταβλητές. Για παράδειγμα, η ηλικία, ο χρόνος ενός ασφαλιστηρίου συμβολαίου, ο αριθμός των ατυχημάτων τα τελευταία τρία χρόνια, ο βαθμός εξέτασης και ο μισθός είναι συμμεταβλητές. Εάν υπάρχει γραμμική σχέση μεταξύ της γραμμικής πρόβλεψης και της συμμεταβλητής, πρέπει να εκτιμηθεί μόνο μία παράμετρος για την μεταβλητή.

Επί παραδείγματι, ένας αναλυτής ο οποίος εργάζεται στις ασφαλίσεις αυτοκινήτων, έχει αναπτύξει ένα GLM για να υπολογίσει το καθαρό ασφάλιστρο που πρέπει να χρεωθεί στον ασφαλισμένο μιας δεδομένης ηλικίας, φύλου και VRG (Vehicle Rating Group).Η εταιρεία χρησιμοποιεί 10 κατηγορίες. Είναι γνωστό ότι η συνολική απαίτηση ενός ατόμου ενός συγκεκριμένου προφίλ (ηλικίας, φύλου και VRG) μοντελοποιείται καλά από μια Γάμμα τ.μ. . Ας υποθέσουμε ότι ο αναλυτής δεν βρίσκει αλληλεπιδράσεις μεταξύ αυτών των παραγόντων και γραμμική επίδραση της ηλικίας. Τότε το τελικό μοντέλο θα έχει γραμμική πρόβλεψη της μορφής:

$$\eta = \alpha_i + \gamma_j + \beta x_{age}$$

Όπου α_i είναι η παράμετρος του i-οστού επιπέδου του φύλου($i = 1$ για άνδρα, ($i = 2$ για γυναίκα), γ_j είναι η παράμετρος του j-οστού επιπέδου VRG($1 \leq j \leq 10$) και β είναι η παράμετρος της ηλικίας (x_{age}). (Boland, 2007)

2.3.4 Αλληλεπιδράσεις

Υπάρχει αλληλεπίδραση μεταξύ δύο παραγόντων B και C όταν το αποτέλεσμα στη γραμμική πρόβλεψη της αλλαγής του επιπέδου του παράγοντα Β ποικίλλει ανάλογα με το επίπεδο του παράγοντα C. Ας υποθέσουμε ότι ο παράγοντας Β έχει b επίπεδα και ο C έχει c επίπεδα. Τότε το μοντέλο περιλαμβάνει τους παράγοντες B και C αλλά όχι την αλληλεπίδρασή τους έχει $b + c - 1$ παράγοντες. Ωστόσο, το μοντέλο που περιλαμβάνει τα B και C και την αλληλεπίδρασή τους $B.\ C$ έχει bc παραμέτρους. Επομένως, προσθέτοντας έναν όρο αλληλεπίδρασης στα αποτελέσματα του μοντέλου στην εκτίμηση προστίθενται επιπλέον $bc - (b + c - 1) = (b - 1)\ (c - 1)$ παράμετροι!

Συνεχίζοντας με το παράδειγμά μας για την αξιολόγηση ασφαλίστρων αυτοκινήτου, ας υποθέσουμε ότι ο αναλυτής συμπεριέλαβε στο προηγούμενο μοντέλο του μια αλληλεπίδραση μεταξύ φύλου και VRG. Αυτό σημαίνει ότι η διαφορά (στον γραμμικό παράγοντα πρόβλεψης) μεταξύ των αντρών και των γυναικών δεν είναι σταθερή μέσω του VRG. Ο γραμμικός προγνωστικός δείκτης για ένα τέτοιο μοντέλο μπορεί να γραφτεί ως εξής:

$$\eta = a_i + \gamma_j + (\alpha\gamma)_{ij} + \beta x_{age},$$

όπου η παράμετρος $(\alpha\gamma)_{ij}$ αντιπροσωπεύει το φαινόμενο αλληλεπίδρασης στο i-οστό επίπεδο του φύλου και στο j-οστό επίπεδο του VRG. Παρατηρήστε ξανά ότι το μοντέλο είναι υπερπαραμετροποιημένο. Ο τρόπος που γράφτηκε παραπάνω υποδηλώνει ότι πρέπει να εκτιμήσουμε 20 παραμέτρους αλληλεπίδρασης. Στην πραγματικότητα, μόνο (2 - 1) (10 - 1) = 9 μπορούν να εκτιμηθούν ανεξάρτητα από τα κύρια αποτελέσματα. (Boland, 2007)

Μπορούμε επίσης να έχουμε μια αλληλεπίδραση μεταξύ μιας συμμεταβλητής και ενός παράγοντα. Σε αυτή την περίπτωση, η επίδραση της συμμεταβλητής αλλάζει σε διαφορετικά επίπεδα του παράγοντα. Ας υποθέσουμε ότι υπάρχει μια αλληλεπίδραση μεταξύ μιας συμμεταβλητής x και ενός παράγοντα B (με bεπίπεδα) σε ένα μοντέλο. Εάν το x έχει γραμμική σχέση με το η για ένα δεδομένο επίπεδο B, τότε θα πρέπει να εκτιμήσουμε τις παραμέτρους β για το x (μία για κάθε επίπεδο του B). Για παράδειγμα, επεκτείνοντας το προηγούμενο μοντέλο για τα ασφάλιστρα αυτοκινήτων, ας υποθέσουμε ότι η επίδραση της ηλικίας ήταν διαφορετική για κάθε

VRG. Τότε θα υπήρχε αλληλεπίδραση μεταξύ ηλικίας και VRG και θα μπορούσαμε να γράψουμε

$$\eta = a_i + \gamma_j + (\alpha\gamma)_{ij} + \beta x_{age} + \beta_j x_{age} ,$$

όπου β υποδηλώνει τη (μέση) επίδραση της ηλικίας και β_j είναι η διαφορά μεταξύ αυτής (μέσης) επίδρασης και εκείνης για το VRG j $(1 \leq j \leq 10)$ μιας αύξησης ενός χρόνου της ηλικίας στο γραμμικό εκτιμητή η. Όλες οι άλλες παράμετροι στο μοντέλο είναι όπως πριν. Σημειώνεται για άλλη μια φορά ότι γράφοντας το μοντέλο με αυτόν τον (ενδελεχή) τρόπο, είναι αυστηρά υπερπαραμετροποιημένο.

Για μια μεταβλητή όπως η ηλικία, οι πολυώνυμοι συντελεστές μπορούν να εμφανιστούν στο γραμμικό παράγοντα πρόβλεψης. Τα παρακάτω περιγράφουν ένα μοντέλο που περιέχει τους παράγοντες φύλο και VRG και που υπάρχει τετραγωνική σχέση μεταξύ του γραμμικού παράγοντα πρόβλεψης και της ηλικίας:

$$\eta = a_i + \gamma_j + \beta_1 x_{age} + \beta_2 x_{age}^2$$

Ένας άλλος τρόπος για να το δούμε όμως είναι να θεωρήσουμε την x_{age}^2 ως άλλη μεταβλητή και με αυτόν τον τρόπο είναι σαφές ότι το η είναι «γραμμικό στις επεξηγηματικές μεταβλητές» (μία από τις απαιτήσεις είναι ότι το η να είναι ένας γραμμικός συνδυασμός των παραμέτρων Q). Είναι επίσης κατάλληλο για να χρησιμοποιήσουμε μια συνάρτηση ηλικίας ως μεταβλητή. Μπορεί να πιστεύουμε ότι υπάρχει μια γραμμική σχέση μεταξύ η και log(age). Υποθέτοντας πάλι ότι οι παράγοντες φύλου και VRG είναι στο μοντέλο, ένας πιθανός εκτιμητής θα ήταν:

$$\eta = a_i + \gamma_j + \beta \log(x_{age})$$

Ο παρακάτω πίνακας δίνει σύντομα μοντέλα που είναι χρήσιμα κατά την προσαρμογή των μοντέλων όταν χρησιμοποιείται κάποιο πρόγραμμα στον υπολογιστή όπως το SAS. (Boland, 2007)

Συμβολισμός για γραμμικούς παράγοντες πρόβλεψης .

Γραμμικοί παράγοντες πρόβλεψης μοντέλων GLM	
Μοντέλο πρόβλεψης	Συντομογραφίες
$\eta = a_i$	φύλο
$\eta = \alpha_i + \gamma_j$	φύλο + VRG
$\eta = \alpha_i + \gamma_j + \beta x_{age}$	φύλο + VRG + ηλικία
$\eta = \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \beta x_{age}$	φύλο + VRG + φύλο.VRG + ηλικία
$\eta = \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \beta x_{age} + \beta_j x_{age}$	φύλο + VRG + φύλο.VRG + ηλικία + ηλικία.VRG
$\eta = \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \beta_j x_{age}$	φύλο + VRG + φύλο.VRG + ηλικία.VRG
$\eta = \alpha_i + \gamma_j + \beta_1 x_{age} + \beta_2 x_{age}^2$	φύλο + VRG + ηλικία + ηλικία^2
$\eta = \alpha_i + \gamma_j + \beta \log(age)$	φύλο + VRG + log(ηλικία)

2.3.5 Ελάχιστα επαρκή στατιστικά στοιχεία

Εάν έχουμε ένα δείγμα n ανεξάρτητων παρατηρήσεων $y_1, \ldots, y_n$ όπου κάθε y_i έχει κατανομή από μια δεδομένη εκθετική οικογένεια με παραμέτρους $(\theta_i,)$, τότε η log-πιθανοφάνεια για το δείγμα είναι

$$l(\theta, y) = \sum_1^n A_i \left[\frac{yii - \gamma(i)}{\Phi}\right] + r(y_i, \phi / A_i)$$

Αν g είναι μία συνάρτηση όπου $g(\mu_l) = \theta_l = \sum_1^p x_{ij}\beta_j$, τότε η log-πιθανοφάνεια για τις παραμέτρους $(\beta_1, \ldots, \beta_\rho)$ παίρνει τη μορφή

$$l(\beta, y) = (\sum_{j=1}^{p} \beta_j \sum_{i=1}^{n} A_i \frac{y_i x_{ij}}{\varphi} - \sum_{i=1}^{n} [A_1 \frac{y(_i)}{\varphi} - r(y_i, /A_i)])$$

Στη συνέχεια ακολουθεί (υποθέτοντας ότι τα $_1, \ldots, _\rho$ είναι γνωστά) από τοθεώρημα παραγοντοποίησης ότι

$$\{\sum_{i=1}^{n} A^{\underline{y_i x_{ij}}} : j = 1, \ldots, p\}$$

είναι ένα σύνολο ελάχιστα επαρκών στατιστικών συναρτήσεων για τις παραμέτρους $(\beta_1, \ldots, \beta_\rho)$. Ουσιαστικά, αυτές οι στατιστικές συναρτήσεις περιέχουν όλες τις απαραίτητες πληροφορίες για την εκτίμηση του Q, και η γνώση των επιμέρους τιμών του y_i δεν μας δίνει περισσότερες πρόσθετες πληροφορίες σχετικά με αυτό. Αυτός είναι ένας άλλος λόγος που η συνάρτηση g με την ιδιότητα που $g(\mu) = \theta = \eta$ για εκθετική οικογένεια ονομάζεται κανονικός σύνδεσμος για αυτήν την οικογένεια. Όπως έχει ήδη δειχθεί, $E(Y) = \mu = \gamma'(\theta)$, από την οποία προκύπτει ότι η κανονική συνάρτηση σύνδεσης g για μια οικογένεια ικανοποιεί την σχέση $g = (\gamma')^{-1}$. Η συνήθης κανονική σύνδεση και η συνάρτηση διακύμανσης σχετίζονται με το

$V(\mu) = (g^{-1})'(\mu)$ όπου ο τόνος (‘) δηλώνει τη παραγώγιση ως προς το θ.

(Boland, 2007)

2.4 Επιλογή μοντέλου και απόκλιση

Συνήθως πριν επιλεγεί ένα τελικό μοντέλο, υπάρχει ένας αριθμός υποψήφιων μεταβλητών που μπορεί να επηρεάσουν την απόκριση. Στο τελικό μοντέλο επιλέγεται ένα υποσύνολο από αυτά, το οποίο ελπίζουμε να δίνει την καλύτερη δυνατή (ή τουλάχιστον μία καλή) περιγραφή της υποκείμενης διαδικασίας. Δύο κριτήρια που χρησιμοποιούνται για να γίνουν οι επιλογές είναι η ευελιξία και η φειδωλότητα του μοντέλου (η ποιότητα των λίγων παραμέτρων που ταιριάζουν καλά στα δεδομένα). Θα θέλαμε το μοντέλο να παρέχει μία καλή περιγραφή των δεδομένων που χρησιμοποιήθηκαν για να προσαρμοστούν στο μοντέλο. Το πόσο καλή προσαρμογή έγινε (ή ίσως ακριβέστερα η έλλειψη προσαρμογής) συχνά μετριέται με την απόκλιση. Κάποιος θα ήθελε επίσης το τελικό μοντέλο να είναι όσο το δυνατόν πιο απλό. Τα μοντέλα με λιγότερες παραμέτρους είναι ευκολότερα στη χρήση. Επιπλέον, η εκτίμηση παραμέτρων για μεταβλητές ή παράγοντες που ουσιαστικά δεν έχουν επίδραση στην απόκριση είναι επιζήμια για το μοντέλο, καθώς αυξάνει τη διακύμανση των παραμέτρων που εκτιμήθηκαν, και ως εκ τούτου μειώνει την ακρίβεια των συμπερασμάτων που προκύπτουν. Δυστυχώς, εδώ βρίσκεται μία σύγκρουση συμφερόντων! Όσο περισσότερες παραμέτρους προσθέτει κάποιος στο μοντέλο, τόσο καλύτερα θα ταιριάζει με τα δεδομένα, αλλά τόσο λιγότερο λιτό θα είναι το μοντέλο που προκύπτει. Χρειαζόμαστε έναν τρόπο επίλυσης αυτών των δύο αντικρουόμενων στόχων.

Παρόλο που η απόκλιση μπορεί να χρησιμοποιηθεί για να δώσει ένα βαθμό προσαρμογής σε ένα μοντέλο, η κύρια χρήση είναι η σύγκριση πιθανών μοντέλων που είναι εμφωλευμένα (δηλαδή, όπου το σύνολο των επεξηγηματικών μεταβλητών ενός μοντέλου είναι υποσύνολο αυτών ενός άλλου). Μπορεί κανείς να δει την ανάλυση της απόκλισης ως ουσιαστικά μια γενίκευση της κλασσικής μεθόδου ανάλυσης διακύμανσης. (Boland, 2007)

2.4.1 Απόκλιση και το κορεσμένο μοντέλο

Το κορεσμένο μοντέλο παρέχει τέλεια προσαρμογή για τα δεδομένα. Ένα τέτοιο μοντέλο αναφέρεται συχνά ως μέγιστο ή πλήρες μοντέλο. Ο αριθμός των παραμέτρων (που υποδηλώνεται με n_S) για το κορεσμένο μοντέλο είναι σε πολλές περιπτώσεις ο ίδιος με τον μέγιστο αριθμό τάξεων συμμεταβλητών όταν τα δεδομένα είναι κατηγορικά ή ο αριθμός των μέσων των δεδομένων είναι συνεχής.

Για να συγκρίνουμε την καταλληλότητα ενός μοντέλου M με παραμέτρους διακύμανσης $p < n_s$ για ένα σύνολο δεδομένων y, συγκρίνουμε τη μέγιστη τιμή της συνάρτησης πιθανότητας με αυτήν του κορεσμένου μοντέλου. Η log-πιθανοφάνεια $l(\theta;\ y)$ σε ένα γενικευμένο γραμμικό μοντέλο εξαρτάται από τους γραμμικούς συντελεστές στον παράγοντα πρόβλεψης $\eta = \beta_1 x_1 + \cdots + \beta_p x_p$ μέσω της συνάρτησης σύνδεσης $g(\mu) = \eta$. Βρίσκει κανείς εκτιμήσεις μέγιστης πιθανότητας των συντελεστών στον γραμμικό παράγοντα πρόβλεψης, οι οποίοι από την ιδιότητα αμετάβλητου της μεθόδου της μέγιστης πιθανότητας αποδίδουν εκτιμήσεις μέγιστης πιθανοφάνειας του μ_i.

Έστω L_S και L_M, αντίστοιχα, οι μέγιστες τιμές των συναρτήσεων πιθανοφάνειας για τα δεδομένα που υπάρχουν για το κορεσμένο μοντέλο S και το μοντέλο υπό εξέταση

M. Χρησιμοποιώντας το $l_S = logL_S$ και $l_M = logL_M$, θεωρούμε το στατιστικό που δίνεται από τον τύπο $log\lambda_{S,M} = log\ (L_s\ /L_M\ ) = l_S - l_M$. Ισχύει $L_S \geq L_M$, και επομένως $l_S - l_M \geq 0$ αφού (υποθέτοντας ότι χρησιμοποιείται η ίδια εκθετική κατανομή οικογένειας και συνάρτηση σύνδεσης και για τα δύο μοντέλα) το κορεσμένο μοντέλο παρέχει μια πλήρη περιγραφή των δεδομένων. Εάν το μοντέλο M ταιριάζει αρκετά καλά με τα δεδομένα, τότε $L_S \doteq L_M$και $log\lambda_{S,M}$ θα πρέπει να είναι αρκετά μικρό. Μεγάλες τιμές $log\lambda_{S,M}$ υποδηλώνουν κακή προσαρμογή.

Ας υποθέσουμε προς το παρόν ότι δουλεύουμε με ένα μοντέλο όπου η παράμετρος διασποράς $= 1$ (για παράδειγμα, όταν η απόκριση είναι Poisson ή διωνυμική). Η μέγιστη εκτίμηση πιθανότητας για $E(Y_i) = \mu = \gamma'(\theta_i)$ στο κορεσμένο μοντέλο είναι σαφώς η ίδια η y_i, ή, ισοδύναμα, η εκτίμηση για θ_i στο κορεσμένο μοντέλο είναι $\theta(y_i) \equiv (\gamma')^{-1}(y_i)$. Έστω $\hat{\theta}$ η μέγιστη εκτίμηση πιθανοφάνειας της θ_i υπό το μοντέλο M. Στη συνέχεια, η απόκλιση του μοντέλου M βασίζεται στα δεδομένα y (ή δύο φορές το στατιστικό λόγο της log-πιθανοφάνειας για τον έλεγχο του μοντέλου M σε σχέση με το κορεσμένο μοντέλο S) παίρνει τη μορφή :

$$D_M = 2(l_S - l_M) = 2\sum_{i=1}^{ns} A_i\,\{y_i\theta(y_i) - \gamma(\theta(y_i))\} - \{y_i\hat{\theta} - \gamma(\hat{\theta}_i)\}$$

Όπου

$$D_M(\,y_i) = 2\,A_i(\{y_i\theta(y_i) - \gamma(\theta(y_i))\} - \{y_i\hat{\theta} - \gamma(\hat{\theta}_i)\})$$

είναι η ατομική συμβολή της παρατήρησης y_i στη συνολική απόκλιση D_M.

Η απόκλιση D_M μπορεί να θεωρηθεί ως ένα μέτρο της απόστασης μεταξύ του κορεσμένου μοντέλου και ενός προτεινόμενου μοντέλου M. Το απλούστερο μοντέλο είναι το λεγόμενο σταθερό μοντέλο όπου το $E(Y_i) = \mu$ είναι σταθερό για όλα τα i και μερικές φορές αναφέρεται ως το μηδενικό μοντέλο M_0 (με μία παράμετρο - τον άγνωστο μέσο όρο). Η απόκλιση του μηδενικού μοντέλου ονομάζεται συνήθως μηδενική απόκλιση. Όταν η παράμετρος κλίμακας διαφέρει από τη μονάδα, η ποσότητα D_M (όπου οι εκτιμήσεις μέγιστης πιθανοφάνειας των παραμέτρων λαμβάνουν υπόψη το) εξακολουθεί να αναφέρεται συνήθως ως απόκλιση, αν και δεν είναι πλέον διπλάσια του λόγου πιθανοφανειών. (Boland, 2007)

Όταν $\neq 1$, συχνά θεωρείται ότι η κλιμακωτή απόκλιση δίνεται από το κλάσμα D_M /. Για το κανονικό μοντέλο M όπου $Yi \sim N(\mu, \sigma^2)$ και

$$E(Y_i) = \mu_i = x_i^T \beta = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Η διακύμανση παίρνει τη μορφή $D_M = \sum_1^n (y_i - \hat{\mu}_i)^2$ και επιπλέον,

$$\frac{D_M}{\sigma^2} = \frac{\sum_1^n (y_i - \hat{\mu}_i)^2}{\sigma^2} \sim X^2_{n-p}.$$

Επομένως, στην κανονική περίπτωση όπου η διακύμανση είναι γνωστή, η (κλιμακωτή) απόκλιση μπορεί να χρησιμοποιηθεί για να ελέγξει εάν το μοντέλο M είναι αποδεκτό. Όταν η παράμετρος κλίμακας είναι άγνωστη, είναι συνήθης πρακτική η χρήση εκτίμησης.

Η παράμετρος κλίμακας (ή συντελεστής κλίμακας) μπορεί να χρησιμοποιηθεί για να μοντελοποιήσει την υπερδιασπορά στο μοντέλο. Η υπερβολική διασπορά είναι ένα σχετικά συνηθισμένο φαινόμενο που προκύπτει από την παρατήρηση περισσότερων διαφοροποιήσεων στα δεδομένα από ό,τι θα μπορούσε να αναμένεται σε ένα δεδομένο μοντέλο. Οι πιθανές αιτίες υπερδιασποράς περιλαμβάνουν την παρουσία υπερβολικών ποσοτήτων και έλλειψη επεξηγηματικών μεταβλητών στο μοντέλο.

Εάν τα δεδομένα δεν ακολουθούν την κανονική κατανομή και το είναι γνωστό, τότε το D_M / μπορεί να είναι περίπου X^2_{n-p}, όταν το M είναι μοντέλο p παραμέτρων. Σημειώνεται ότι η (κλιμακωτή) μηδενική απόκλιση είναι περίπου X^2 που κατανέμεται με $n_s - 1$ βαθμούς ελευθερίας όταν ισχύει το μηδενικό μοντέλο! Αν η κατανομή είναι κανονική, τότε το D_{M_0}/σ^2 ακολουθεί (ακριβώς) την κατανομή X^2_{n-1}.

(Boland, 2007)

2.4.2 Συγκρίνοντας μοντέλα με απόκλιση

Έστω ότι γνωρίζουμε ότι το σωστό μοντέλο για τα δεδομένα μας πρέπει οπωσδήποτε να περιέχει τις μεταβλητές p: $x_1, \ldots, x_p$,αλλά θα θέλαμε να μάθουμε αν (μερικές από) τις q μεταβλητές $x_{p+1}, \ldots, x_{p+q}$ θα πρέπει επίσης να εμφανίζονται στο σωστό μοντέλο. Έστω β_i η παράμετρος που αντιστοιχεί στη μεταβλητή x_i του μοντέλου. Το M_1 να υποδηλώνει το μοντέλο όπου $\beta_i = 0$ για $i > p$ και M_2 το (μεγαλύτερο) μοντέλο όπου $\beta_i \neq 0$ για τουλάχιστον ένα $i > p$. Η έλεγχος λόγου πιθανοφανειών H_0: M_1 έναντι H_A: M_2 βασίζεται στα στοιχεία του στατιστικού ελέγχου

$$2 \log \frac{L_{M_2}}{L_{M_1}} = 2(l_{M_2} - l_{M_1}) \sim X_q^2 \quad (2.5)$$

Όπου L_{M2} είναι η μέγιστη τιμή της πιθανοφάνειας κάτω από το μοντέλο με παραμέτρους $p + q$ με αντίστοιχη log-πιθανοφάνεια l_{M_2} (παρόμοια για L_{M_1} και l_{M_1}). Ουσιαστικά, εάν οι επιπλέον μεταβλητές στο μεγαλύτερο μοντέλο δεν έχουν πραγματικά καμία επίδραση, τότε η διπλή διαφορά μεταξύ των (μεγιστοποιημένων) log-πιθανοφανειών των δύο μοντέλων θα έχει, περίπου, κατανομή X^2 με q βαθμούς ελευθερίας!

Όπως αναφέρθηκε προηγουμένως, η απόκλιση είναι ένα μέτρο καλής προσαρμογής του μοντέλου. Διαισθητικά, αυτό έχει νόημα, διότι εάν το μοντέλο ενδιαφέροντος (με μόνο μεταβλητές p) δεν προσαρμόζεται καλά στα δεδομένα, το μοντέλο πιθανότατα θα προβλέψει προσαρμοσμένες τιμές αρκετά διαφορετικές από τις παρατηρούμενες τιμές. Με άλλα λόγια, τα δεδομένα δεν θα είναι συνεπή με το μοντέλο και συνεπώς η πιθανοφάνεια των παρατηρούμενων δεδομένων (και επίσης η log-πιθανοφάνεια l_{M1}) θα είναι χαμηλή στο υπό εξέταση μοντέλο. Η κορεσμένη log-πιθανοφάνεια l_s θα είναι τότε σημαντικά μεγαλύτερη από την l_{M1} και η απόκλιση θα είναι μεγάλη. Χρησιμοποιώντας παρόμοιο σκεπτικό, είναι λογικό η απόκλιση να είναι μικρή όταν το μοντέλο προσαρμόζεται καλά στα δεδομένα. Επομένως, κατά τη δοκιμή του μοντέλου M_1 έναντι του M_2, έχουμε από την εξίσωση (2.5) ότι όταν το M_1 είναι έγκυρο

$$2(l_{M_2} - l_{M_1}) = 2(l_s - l_{M_1}) - 2(l_s - l_{M_2}) = D_{M_1} - D_{M_2} \equiv \Delta D \sim X_q^2$$

Ως εκ τούτου, υπό την μηδενική υπόθεση ότι $\beta_i = 0$ για $i \geq p + 1$, έχουμε ότι η διαφορά απόκλισης ΔD μεταξύ του μικρότερου και του μεγαλύτερου μοντέλου είναι (περίπου) μια ^{2} τυχαία μεταβλητή με q βαθμούς ελευθερίας. Εάν η μηδενική υπόθεση είναι ψευδής (με το M_2 να είναι το σωστό μοντέλο), θα μπορούσαμε να περιμένουμε ότι η διαφορά απόκλισης θα είναι μεγαλύτερη από αυτή της μηδενικής υπόθεσης. Σε επίπεδο σημαντικότητας α%, είναι λογικό να συμπεράνουμε ότι μερικές από τις επιπλέον μεταβλητές θα πρέπει να είναι στο σωστό μοντέλο εάν $\Delta D = D_{M_1} - D_{M_2} > X^2_{1-a,q}$ (δηλαδή, για έναν έλεγχο 5% θα συμπεριλάμβανονταν οι επιπλέον όροι στο μοντέλο εάν η διαφορά απόκλισης ΔD υπερβαίνει το 95% της κατανομής X^2 με q βαθμούς ελευθερίας). Πρέπει να σημειωθεί ότι τα μοντέλα συνήθως συγκρίνονται δοκιμάζοντας έναν επιπλέον όρο τη φορά! (Boland, 2007)

2.4.3 Ανάλυση υπολοίπων για γενικευμένα γραμμικά μοντέλα

Έχοντας επιλέξει ένα μοντέλο, ίσως χρησιμοποιώντας μια ανάλυση απόκλισης όπως περιγράφεται στην τελευταία ενότητα, πρέπει να ελεγχθεί εάν οι παραδοχές του μοντέλου είναι δικαιολογημένες και αν το μοντέλο παρέχει καλή προσαρμογή στα δεδομένα.

Μια βασική υπόθεση για τα GLM είναι ότι τα σημεία δεδομένων $y_1, \ldots, y_n$ είναι ανεξάρτητες τυχαίες παρατηρήσεις από την ίδια εκθετική οικογένεια. Θα πρέπει επίσης να ελέγξουμε, για να δούμε ότι το συγκεκριμένο GLM που έχει προταθεί είναι κατάλληλο, ότι έχουμε χρησιμοποιήσει μια λογική (φυσική;) συνάρτηση σύνδεσης και ότι έχουμε συμπεριλάβει στο μοντέλο όλους τους διαθέσιμους σημαντικούς παράγοντες/μεταβλητές, έχοντας κατά νου την αρχή της φειδωλότητας. Συχνά δίναται να εντοπίστουν κάποια ελαττώματα σε ένα μοντέλο με τη χρήση των διαγραμμάτων υπολοίπων.

Τα υπόλοιπα βασίζονται στις διαφορές μεταξύ των παρατηρούμενων σημείων δεδομένων και προσαρμοσμένων τιμών που προβλέπονται από το μοντέλο. Το i-οστό υπόλοιπο Pearson για ένα δεδομένο μοντέλο δίνεται από τον τύπο

$$r_{P_i} = (y_i - \hat{\mu}_i)/\sqrt{Var(\hat{\mu}_i)}$$

όπου $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$ και $Var(\hat{\mu}_i) = (\frac{\varphi}{A_i})V(\hat{\mu}_i)$ (εδώ, είναι η εκτιμώμενη διακύμανση του Y δεδομένου του μέσου όρου $\hat{\mu}_i$). Από την άλλη πλευρά, το i-οστό υπόλοιπο δίνεται από τον τύπο

$$r_{d_i} = sgn(y_i - \hat{\mu}_i)\sqrt{d_i}$$

όπου d_i είναι η συμβολή του i-οστής παρατήρησης που δείχνει την (κλιμακωτή) απόκλιση για το μοντέλο και $sgn(y_i - \hat{\mu}_i) = 1$ αν $y_i - \hat{\mu}_i > 0$ και -1 διαφορετικά.

Τα υπόλοιπα Pearson έχουν συχνά ασυμμετρία για μη κανονικά δεδομένα, και αυτό καθιστά την ερμηνεία τέτοιων διαγραμμάτων υπολοίπων πιο δύσκολη. Τα υπόλοιπα των αποκλίσεων είναι πιο πιθανό να κατανέμονται σχεδόν κανονικά, και ως εκ τούτου συχνά προτιμώνται (για ασφαλιστικές και αναλογιστικές εφαρμογές). Κάποιος μπορεί να δείξει ότι για τα κανονικά κατανεμημένα δεδομένα, τα υπόλοιπα Pearson και οι αποκλίσεις είναι ίδιες.

Σύμφωνα με τις υποθέσεις των GLM, αναμένουμε ότι τα διαγράμματα υπολοίπων με ένα καλό μοντέλο δεν θα δείχνουν κανένα μοτίβο. Μετά την επιλογή του μοντέλου, θα πρέπει να σχεδιάσουμε τα υπόλοιπα έναντι των παραγόντων και των μεταβλητών στο μοντέλο. Αν οποιαδήποτε μοτίβο εντοπιστεί μπορεί να υποδηλώνει ότι έχει χαθεί ένα αποτέλεσμα που θα έπρεπε να είχε συμπεριληφθεί στο μοντέλο. Οι παρατηρήσεις των δεδομένων που έχουν εξαιρετικά μεγάλα υπόλοιπα μπορεί να είναι ακραίες. Ένα ιστόγραμμα υπολοίπων μπορεί να χρησιμοποιηθεί για τον έλεγχο της μορφής κατανομής στα GLM. (Boland, 2007)

ΚΕΦΑΛΑΙΟ 3: Τιμολόγηση ασφαλίστρων

Σε αυτό το κεφάλαιο θα επιχειρηθεί η τιμολόγηση ασφαλίστρων, ασφαλιστηρίων συμβολαίων στον Κλάδου του «Πυρός» και πιο συγκεκριμένα ασφαλιστηρίων συμβολαίων κατά του κινδύνου του σεισμού με τη χρήση γενικευμένων γραμμικών μοντέλων και με τη βοήθεια του στατιστικού πακέτου της R. Θα γίνει αρχικά μια παρουσίαση των παραδοχών που χρησιμοποιήθηκαν για την κατασκευή του μοντέλου και έπειτα η κατάδειξη του πλέον κατάλληλου μοντέλου με βάση τους παράγοντες επηρεασμού που έχουν επιλεχθεί και τα παρεχόμενα πρωτογενή στοιχεία.

3.1 Παραδοχές

3.1.1 Προέλευση και περιεχόμενα πρωτογενών στοιχείων

Τα πρωτογενή στοιχεία προέρχονται από ελληνική ασφαλιστική εταιρία και αφορούν οικίες και όχι επιχειρήσεις καθώς τα ασφάλιστρα για επιχειρήσεις όπως π.χ. εργοστάσια έχουν μεγαλύτερο μέσο ύψος και έτσι δεν θα μπορούσαν να μας δώσουν ασφαλή στατιστικά συμπεράσματα. Τα στοιχεία αυτά επίσης αφορούν τα τελευταία 5 ολοκληρωμένα οικονομικά έτη. Δηλαδή από τις ημερομηνίες 1/1/2016 έως τις 31/12/2020 και περιείχαν μεταξύ άλλων ασφάλιστρα ανά έτος ,αριθμούς συμβολαίων, κατηγορίες κτηρίων (με βάση του πώς κάνει την κατηγοριοποίηση η συγκεκριμένη εταιρία) και ταχυδρομικούς κώδικες.

3.1.2 Μεταβλητές του μοντέλου

Η πρώτη ανεξάρτητη μεταβλητή που χρησιμοποιείται είναι η ζώνη επικινδυνότητας η οποία συμβολίζεται με X_1.Η δεύτερη ανεξάρτητη μεταβλητή είναι ο τρόπος οικοδόμησης του κτηρίου που συμβολίζεται με X_2 και τέλος η εξαρτημένη μεταβλητή είναι το ασφάλιστρο το οποίο συμβολίζεται με Y.

3.1.3 Τροποποιήσεις των μεταβλητών

- Τα ασφάλιστρα έχουν αναπροσαρμοστεί με βάση το οικονομικό έτος. Δηλαδή αν ένα ασφαλιστήριο συμβόλαιο έχει ασφάλιστρο αξίας 100 ν.μ. και ημερομηνία έναρξης την 1/7/2021 και λήξης την 1/7/2022 τότε για το οικονομικό 2021 αντιστοιχούν 50 ν.μ. ασφαλίστρου. Αυτό ονομάζεται δεδουλευμένο ασφάλιστρο για το 2021 και άρα το Υ για την διενέργεια των υπολογισμών παίρνει την τιμή 50.
- Σύμφωνα με την Νέα Αρχιτεκτονική της Αυτοδιοίκησης και της Αποκεντρωμένης Διοίκησης - Πρόγραμμα Καλλικράτης(Νόμος 3852 ΦΕΚ Α' 87/7.6.2010) το οποίο εφαρμόστηκε από το 2011 και έπειτα, αρκετοί Δήμοι ενσωματώθηκαν σε άλλους Δήμους το οποίο σημαίνει ότι σε ότι αφορά την κατηγοριοποίηση αυτών σε ζώνες επικινδυνότητας σύμφωνα με το ΦΕΚ 1154 / ΕΑΚ 2003 οι προσαρτημένοι Δήμοι έχουν λάβει το επίπεδο επικινδυνότητας του Δήμου στον οποίο έχουν προσαρτηθεί.

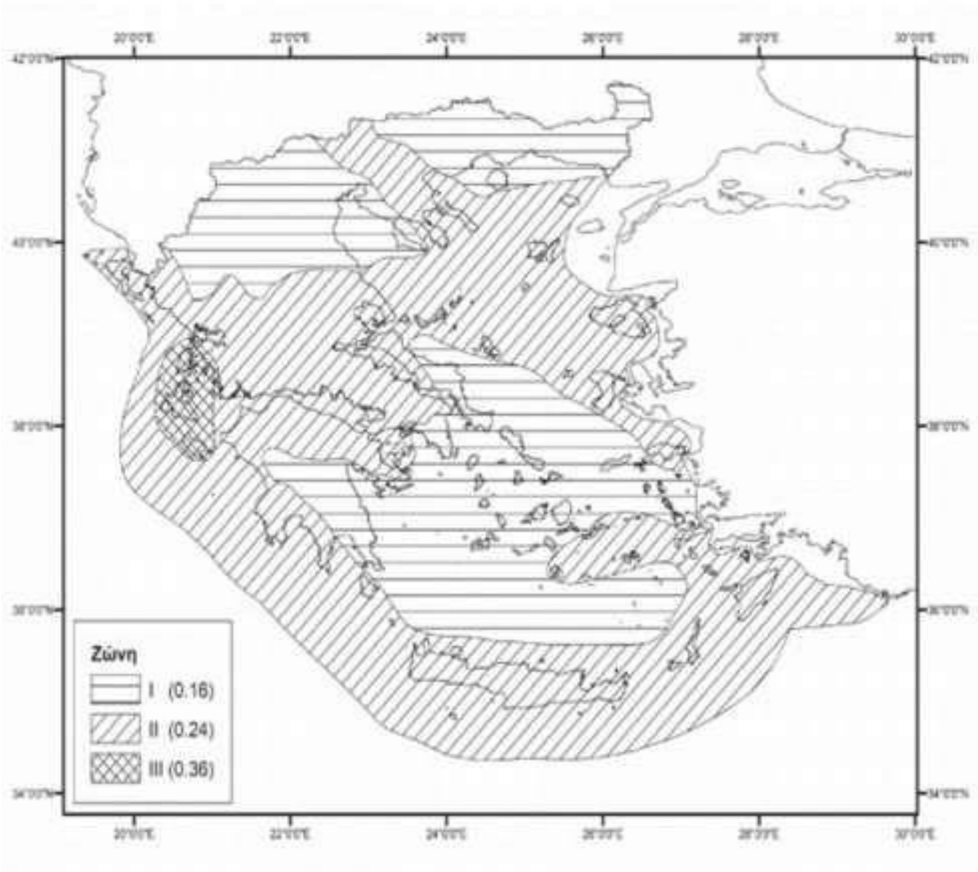
3.1.4 Τιμές των μεταβλητών

Ανεξάρτητη μεταβλητή X_1 : Η μεταβλητή X_1 είναι ποσοτική και παίρνει τις τιμές 0.16 , 0.24 και 0.36 σύμφωνα με τον παρακάτω πίνακα του ΦΕΚ 1154 / ΕΑΚ 2003

Ζώνη Σεισμικής Επικινδυνότητας	I	II	III
α	0.16	0.24	0.36

Πηγή:https://www.civilprotection.gr/sites/default/gscp_uploads/fek_1154b_2003anatheorisixarthseismikhsepikindinotitas_0.pdf

Στον παρακάτω χάρτη παρουσιάζεται η διαφοροποίηση των Ζωνών Σεισμικής Επικινδυνότητας ανά την Ελλάδα



Πηγή:https://www.civilprotection.gr/sites/default/gscp_uploads/fek_1154b_2003anatheorisixarthseismikhsepikindinotitas_0.pdf

Ανεξάρτητη μεταβλητή X_2: Η μεταβλητή X_2 είναι κατηγορικού τύπου και λαμβάνει τις τιμές «Από πέτρα» , «Από τούβλα» , «Από ξύλο».

3.2 Αποτελέσματα

Τα αποτελέσματα μετά από τον έλεγχο τριών Γενικευμένων Γραμμικών Μοντέλων είναι τα εξής:

```
> model_1 <- glm(Premium~Post_code+Type,data=train,family=gaussian())
> summary(model_1)

Call:
glm(formula = Premium ~ Post_code + Type, family = gaussian(),
    data = train)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
  -167    -146    -119     -57   54744

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  170.712     11.748  14.531   <2e-16 ***
Post_code         NA         NA      NA       NA
Type          -3.226      7.691  -0.419    0.675
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1315800)

    Null deviance: 5.594e+10  on 42515  degrees of freedom
Residual deviance: 5.594e+10  on 42514  degrees of freedom
AIC: 719708

Number of Fisher Scoring iterations: 2
```

```
> model_2 <- glm(Premium~Post_code+Type,data=train,family=Gamma())
> summary(model_2)

Call:
glm(formula = Premium ~ Post_code + Type, family = Gamma(), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2118  -1.5524  -1.0392  -0.3933  25.8494

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0058511  0.0004276  13.682   <2e-16 ***
Post_code          NA         NA      NA       NA
Type        0.0001194  0.0002848   0.419    0.675
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 47.53611)

    Null deviance: 108597  on 42515  degrees of freedom
Residual deviance: 108589  on 42514  degrees of freedom
AIC: 502919

Number of Fisher Scoring iterations: 11
```

```
> model_3 <- glm(Premium~Post_code+Type,data=train,family=quasi())
> summary(model_3)

Call:
glm(formula = Premium ~ Post_code + Type, family = quasi(), data = train)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
  -167    -146    -119     -57   54744

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  170.712     11.748  14.531   <2e-16 ***
Post_code         NA         NA      NA       NA
Type          -3.226      7.691  -0.419    0.675
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasi family taken to be 1315800)

    Null deviance: 5.594e+10  on 42515  degrees of freedom
Residual deviance: 5.594e+10  on 42514  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 2
```

Οι Συντελεστές προσδιορισμού R^2 για τα 3 μοντέλα είναι οι κάτωθι:

R_1^2:$2.96e^{-5}$

R_2^2:$2.93e^{-5}$

R_3^2:$2.95e^{-5}$

Βλέπουμε ότι καλύτερη προσαρμογή (αν και με ελάχιστη απόκλιση μεταξύ τους) στα δεδομένα εμφανίζουν τα Μοντέλα 1 και 3 με βάση τον συντελεστή προσδιορισμού. Θα θεωρήσουμε ότι το μοντέλο 3 λοιπόν είναι το βέλτιστο.

Ο Συντελεστής Προσδιορισμού (R^2) είναι ο πιο εύκολος τρόπος εκτίμησης την ερμηνευτικής δύναμης ενός γραμμικού μοντέλου είναι ο συντελεστής προσδιορισμού (coefficient of determination) που συνήθως συμβολίζεται με r^2 ή R^2. Ο συντελεστής αυτός μετρά πόση διακύμανση της εξαρτημένης μεταβλητής κατάφεραν να ερμηνεύσουν οι ανεξάρτητες μεταβλητές. Ουσιαστικά είναι το πιο απλό μέτρο που μετρά την ικανότητα ενός συνόλου παραγόντων να ερμηνεύσουν ένα φαινόμενο.

Ο συντελεστής προσδιορισμού R^2 είναι ο λόγος της διακύμανσης των εκτιμημένων τιμών της εξαρτημένης μεταβλητής προς τη διακύμανση των πραγματικών τιμών της εξαρτημένης μεταβλητής και υπολογίζεται ως εξής:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

όπου n είναι ο αριθμός των παρατηρήσεων, y_i είναι οι πραγματικές τιμές της εξαρτημένης μεταβλητής Y, $\bar{y}$ είναι η μέση τιμής της μεταβλητής Y και $\hat{y}$ είναι οι εκτιμημένες τιμές της Y. Οι τιμές του συντελεστή προσδιορισμού R^2 κυμαίνονται από το 0 ως το 1 και προφανώς όσο η τιμή πλησιάζει προς το 1 τόσο καλύτερη προσαρμογή έχει το μοντέλο. Η ερμηνεία των παραπάνω ορίων έχει ως εξής:

- $R^2 = 1$ σημαίνει ότι οι ερμηνευτικές μεταβλητές εξηγούν το 100% της διακύμανσης της εξαρτημένης μεταβλητής και άρα έχουμε ένα τέλειο μοντέλο

- $R^2 = 0$ σημαίνει ότι οι ερμηνευτικές μεταβλητές δεν εξηγούν καθόλου την διακύμανση της εξαρτημένης μεταβλητής. (Davis, 2002)

Τέλος συνεπάγεται ότι η ευθεία παλινδρόμησης που που περιγράφει καλύτερα τα δεδομένα μας είναι η $g(\mu) = \beta_0 + \beta_2 X_2$ με $\beta_0 = 170{,}712$ $\beta_2 = -3{,}226$.Η μεταβλητή X_1 είναι απόλυτα συσχετισμένη για το συγκεκριμένο εύρος τιμών συνεπώς το μοντέλο δεν την θεωρί ως στατιστικά σημαντική.

Βιβλιογραφία

Ανάκτηση από https://repository.kallipos.gr/bitstream/11419/5035/1/02_chapter_6.pdf

Boland, P. J. (2007). *Statistical and Probabilistic Methods in Actuarial Science.* Ιρλανδία: Chapman & Hall/CRC.

Cosmote. . Ανάκτηση από https://www.cosmote.gr/insurance/house/asfaleia-spitiou-seismos

Davis, J. (2002). *Statistics and Data Analysis in Geology: Third Edition.* New York: Jon Wiley & Sons.

E.A.E.E. (2021). Ανάκτηση από http://www1.eaee.gr/sites/default/files/oikmel_premium5months2021gr.pdf

Hartwig, R. P., & Wilkinson, C. (2010, Ιούλιος). *HURRICANE KATRINA:THE FIVE YEAR ANNIVERSARY.* Ανάκτηση από https://www.iii.org/sites/default/files/1007Katrina5Anniversary.pdf

Insurance Information Institute . (2020, Απρίλιος 2). Ανάκτηση από https://www.iii.org/article/background-on-earthquake-insurance-and-risk

Insurance Information Institute. (2021). Ανάκτηση από Facts + Statistics: Flood insurance: https://www.iii.org/fact-statistic/facts-statistics-flood-insurance

Insurance Information Instutute. (2021, Απρίλιος 2). Ανάκτηση από https://www.iii.org/article/background-on-earthquake-insurance-and-risk

Moon, C. (2021, Αύγουστος 3). *What are FAIR Plans and How Do You Insure a High-Risk Home?* Ανάκτηση από What are FAIR Plans and How Do You Insure a High-Risk Home?

Wikipedia. (2021, Αύγουστος 24). Ανάκτηση από https://en.wikipedia.org/wiki/Hurricane_Andrew

Νεκτάριος, Μ. (2010). *Διοικητική Κινδύνων και Ασφαλίσεις Επιχειρήσεων.* Αθήνα: Εκδόσεις Σταμούλη Α.Ε.

Τζώρτζη, Ε. (2019, Αυγούστου 20). Αποζημιώσεις σε 3.500 ακίνητα λόγω του σεισμού.

Τόγιας, Κ. (2019, Αύγουστος). Ανάκτηση από https://www.asfaleia-spitiou.gr/2019/08/02/%ce%b1%cf%83%cf%86%ce%ac%ce%bb%ce%b9%cf%83%ce%b7-%cf%83%ce%b5%ce%b9%cf%83%ce%bc%ce%bf%cf%8d/

Παράρτημα

Ο κώδικας που χρησιμοποιήθηκε για την εξαγωγή των αποτελεσμάτων είναι ο εξής:

```
install.packages("readxl")

library("readxl")

## ΚΑΘΑΡΙΣΜΟΣ ΚΑΙ ΜΟΡΦΟΠΟΙΗΣΗ ΠΡΩΤΟΓΕΝΩΝ ΔΕΔΟΜΕΝΩΝ ##

data <- read.table("/home/thanos /Downloads/premiums2010_2019.txt", sep = '~', skip
=1,fileEncoding="latin1")

data$V7 <- gsub(',', '.', data$V7)

data$V7 <- as.numeric(data$V7)

data$V8 <- gsub(',', '.', data$V8)

data$V8 <- as.numeric(data$V8)

data$V9 <- gsub(',', '.', data$V9)

data$V9 <- as.numeric(data$V9)

data$V2 <- as.Date(data$V2 , format = "%d/%m/%Y")

data$V3 <- as.Date(data$V3 , format = "%d/%m/%Y")


# a<-separate(file_in, 'V1', paste("n", 1:7, sep="_"), sep="~", extra="drop")

# b<-separate(a, 'V2', paste("n", 8:9, sep="_"), sep="~", extra="drop")

# c<-separate(b, 'V3', paste("n", 10:11, sep="_"), sep="~", extra="drop")

# d<-separate(c, 'V4', paste("n", 12:15, sep="_"), sep="~", extra="drop")

# final<-separate(d, 'V5', paste("n", 16:19, sep="_"), sep="~", extra="drop")


# data$V2 <- format(as.Date(data$V2), "%d/%m/%Y")

# data$V2 <- strptime(as.character(data$V2), "%d/%m/%Y")

# data$V3 <- format(as.Date(data$V3), "%d/%m/%Y")
```

```
# data$V3 <- strptime(as.character(data$V3), "%d/%m/%Y")



filter_years <- subset(data, V2 >= "2016-01-01" & V2 < "2021-01-01")
filter_zeros <- subset(filter_years, V7 > 0)
final_data <- filter_zeros[, c('V1', 'V2', 'V3', 'V7', 'V8')]

last_year <- function(x)
{
  lastyear <- as.numeric(format(x,'%Y')) -1
  last_date <- paste(lastyear, 12, 31, sep = "-")
}

final_data$V3 <- sapply(final_data$V3, last_year)
final_data$V3 <- as.Date(final_data$V3 , format = "%Y-%m-%d")
final_data$diff_in_days<- difftime(final_data$V3 ,final_data$V2 , units = c("days"))
final_data$premium <- (final_data$V7 + final_data$V8) * final_data$diff_in_days / 365

colnames(final_data) <- c("Number", "Start_date", "End_date", "Price", "Extra", "Diff_days",
"Premium")

## ΕΙΣΑΓΩΓΗ ΤΑΧΥΔΡΟΜΙΚΩΝ ΚΩΔΙΚΩΝ ΑΠΟ ΑΡΧΕΙΟ ##
tk_2 <- read.csv('/home/thanos/Downloads/TK_2.csv', sep=',', header = FALSE)
tk_3 <- read.csv('/home/thanos/Downloads/TK_3.csv', sep=',', header = FALSE)

data_2020 <- read_excel('/home/thanos/Downloads/EQ_TK_5 Years.xlsx', sheet = '2020-12')
```

```
data_2019 <- read_excel('/home/thanos/Downloads/EQ_TK_5 Years.xlsx', sheet = '2019-12')

data_2018 <- read_excel('/home/thanos/Downloads/EQ_TK_5 Years.xlsx', sheet = '2018-12')

data_2017 <- read_excel('/home/thanos/Downloads/EQ_TK_5 Years.xlsx', sheet = '2017-12')

data_2016 <- read_excel('/home/thanos/Downloads/EQ_TK_5 Years.xlsx', sheet = '2016-12')


data_years <- rbind(data_2016,data_2017,data_2018,data_2019,data_2020)

## ΚΑΘΟΡΙΣΜΟΣ ΚΑΤΗΓΟΡΙΑΣ ΕΠΙΚΙΝΔΥΝΟΤΗΤΑΣ ##

tk_class <- function(x) {


 if (x %in% tk_2$V1){

  x <- 2

 } else if (x %in% tk_3$V1){

    x <- 3

  } else{

    x <- 1

  }

 }


df <- merge(data_years, final_data, by.x = 'ΞΞ¥ΞœΞ’ΞŸΞ›Ξ‘Ξ™ΞŸ', by.y = 'Number', all.x = TRUE, all.y = TRUE)

df <- df[,c("ΞΞ¥ΞœΞ’ΞŸΞ›Ξ‘Ξ™ΞŸ" , "Ξ¤Ξ‘Ξ§.ΞšΞ©Ξ”.", "Premium")]

df$Ξ¤Ξ‘Ξ§.ΞšΞ©Ξ”. <- sapply(df$Ξ¤Ξ‘Ξ§.ΞšΞ©Ξ”., tk_class)

df <- na.omit(df)

df$Type <- as.numeric(df$ΞΞ¥ΞœΞ’ΞŸΞ›Ξ‘Ξ™ΞŸ)

df$Post_code <- as.numeric(df$Ξ¤Ξ‘Ξ§.ΞšΞ©Ξ”.)

df$Premium <- as.numeric(df$Premium)
```

```
df <- subset(df, Premium>0)
df <- df[,c("Type" , "Post_code", "Premium")]

train_ind <- sample(seq_len(nrow(df)), size = 0.8*nrow(df))

train <- df[train_ind, ]
test <- df[-train_ind, ]

RSQUARE <- function(y_actual,y_predict){
  cor(y_actual,y_predict)^2
}
## ΕΙΣΑΓΩΓΗ ΤΩΝ GLM ΚΑΙ ΕΛΕΓΧΟΣ R ΤΕΤΡΑΓΩΝΟ ΕΛΕΓΧΟΣ##
model_1 <- glm(Premium~Post_code+Type,data=train,family=gaussian())
summary(model_1)
predictions <- predict(model_1, test, type='response')
RSQUARE(test$Premium,predictions)

model_2 <- glm(Premium~Post_code+Type,data=train,family=Gamma())
summary(model_2)
predictions <- predict(model_2, test, type='response')
RSQUARE(test$Premium,predictions)

model_3 <- glm(Premium~Post_code+Type,data=train,family=quasi())
summary(model_3)
predictions <- predict(model_3, test, type='response')
RSQUARE(test$Premium,predictions)
```