ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΣΤΑΤΙΣΤΙΚΑ ΜΟΝΤΕΛΑ ΓΙΑ ΤΗΝ ΕΞΕΛΙΞΗ ΤΟΥ ΣΚΟΡ ΚΑΙ ΤΟ ΤΕΛΙΚΟ ΑΠΟΤΕΛΕΣΜΑ ΣΕ ΕΝΑΝ ΑΓΩΝΑ ΜΠΑΣΚΕΤ

Εμμανουήλ Γ. Ταβλάκης

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

Πειραιάς
Σεπτέμβριος 2013

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

Τα μέλη της Επιτροπής ήταν:
- Αναπληρωτής Καθηγητής Πολίτης Κωνσταντίνος (Επιβλέπων)
- Επίκουρος Καθηγητής Τζαβελάς Γεώργιος
- Λέκτορας Ευαγγελάρας Χαράλαμπος

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

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

STATISTICAL MODELS FOR THE SCORE DEVELOPMENT AND THE FINAL OUTCOME OF A BASKETBALL GAME

By

Emmanouil G. Tavlakis

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

Piraeus, Greece
September 2013

Ευχαριστίες

Θα ήθελα να ευχαριστήσω όλους τους ανθρώπους που συνέβαλαν στην ολοκλήρωση αυτής της διπλωματικής εργασίας. Κατά κύριο λόγο θα ήθελα να ευχαριστήσω τον Επιβλέποντα Αναπληρωτή Καθηγητή κ. Πολίτη Κωνσταντίνο που μου έδωσε την ευκαιρία να ασχοληθώ με ένα τόσο ενδιαφέρον και πρωτότυπο θέμα, να συνδυάσω την Στατιστική με τον Αθλητισμό, δύο πράγματα που αγαπάω. Τον ευχαριστώ για την απεριόριστη υπομονή του και τη διαρκή καθοδήγηση που είχα καθ' όλη τη διάρκεια της συγγραφής αυτής της μελέτης. Τέλος, δεν θα μπορούσα να μην ευχαριστήσω τα αγαπημένα μου οικογενειακά, συγγενικά και φιλικά μου πρόσωπα, για την αμέριστη υλική και ηθική υποστήριξη που μου παρείχαν με τη βοήθεια του Θεού, καθ' όλη την διάρκεια των μεταπτυχιακών μου σπουδών.

Περίληψη

Η Στατιστική είναι πλέον αναπόσπαστο κομμάτι και στον Αθλητισμό, βρίσκει απεριόριστες εφαρμογές σε όλα τα ατομικά ή ομαδικά σπορ και ήδη αθλητικά σωματεία δαπανούν μεγάλα ποσά για να πάρουν στατιστική πληροφόρηση. Χρησιμοποιείται από προπονητές και ειδικούς στατιστικολόγους των σπορ, με σκοπό την παρακολούθηση τόσο της ατομικής απόδοσης των παικτών μιας ομάδας, όσο και της ομάδας ως σύνολο, καθώς και από τις επιχειρήσεις (εταιρίες στοιχήματος ή εξειδικευμένες στα σπορ) κατασκευάζοντας στατιστικά μοντέλα για να προβλέψουν πόσο πιθανό είναι να συμβεί μια πληθώρα γεγονότων και να εξάγουν κέρδος.

Στην παρούσα μελέτη, αρχικά θα αναφερθούν εν συντομία μεθοδολογίες που χρησιμοποιούν οι ερευνητές στο άθλημα του μπάσκετ. Χρησιμοποιώντας πραγματικά δεδομένα από το μεγαλύτερο Ευρωπαϊκό πρωτάθλημα μπάσκετ, θα αναλύσουμε με πίνακες και γραφήματα τα κύρια στατιστικά στοιχεία των παιχνιδιών, θα κάνουμε κάποιους στατιστικούς ελέγχους και στη συνέχεια θα προσαρμόσουμε Μοντέλα Λογιστικής Παλινδρόμησης για να βρούμε ποια είναι τα σημαντικότερα στοιχεία που επηρεάζουν το τελικό αποτέλεσμα ενός αγώνα και κατά πόσο μπορούμε αυτό να το προβλέψουμε.

Abstract

Statistical Science has become a major part of Sports. It applies to all the individual or team sports and athletic clubs already spend large amounts of money to get valuable statistical information. Statistics is in use by trainers and sports statisticians in order to monitor either the players' individual performance or the team as a whole. Furthermore, betting companies build statistical models to predict how likely an event is to occur and make a profit.

In this study, statistical techniques used by the researchers into basketball will be briefly reported at first. Using actual data from the most prestigious European basketball tournament, we will analyze the games' main statistics with the help of tables and charts and next we will perform a couple of statistical tests. Afterwards, we will employ Logit and Probit models aiming to find the key factors affecting the final outcome of a game and to what extent we are able to predict it.

Περιεχόμενα

Κατάλογος πινάκων

Κατάλογος σχημάτων

Κατάλογος συντομογραφιών

β.ε.	βαθμοί ελευθερίας
δ.ε.	διάστημα εμπιστοσύνης
ε.σ.	επίπεδο σημαντικότητας
τ.μ.	τυχαία μεταβλητή
AD	Στατιστική συνάρτηση των Anderson-Darling
AIC	Akaike Information Criterion – Πληροφοριακό Κριτήριο του Akaike
AUC	Area Under Curve – Εμβαδόν Κάτω από την Καμπύλη
BIC	Bayesian Information Criterion – Πληροφοριακό Κριτήριο του Bayes
CHISQ	Στατιστικό χ^2 ελέγχου
CVM	Στατιστική συνάρτηση των Cramer-Von Mises
GLM	General Linear Models – Γενικευμένα Γραμμικά Μοντέλα
IRLS	Iterative Reweighted Least Squares
KS	Kolmogorov-Smirnov
LLH	Λογάριθμος της πιθανοφάνειας
LR	Logistic Regression – Λογιστική Παλινδρόμηση
LRT	Likelihood Ratio Test – Έλεγχος του Λόγου των Πιθανοφανειών
MLE	Maximum Likelihood Estimation – Εκτίμηση Μέγιστης Πιθανοφάνειας
OLS	Ordinary Least Squares – Μέθοδος Ελαχίστων Τετραγώνων
ROC	Receiver Operating Characteristic Curve – Χαρακτηριστική Καμπύλη
WLS	Weighted Least Squares – Μέθοδος Σταθμισμένων Ελαχίστων Τετραγώνων

Εισαγωγή

Τα τελευταία χρόνια, μια πληθώρα στατιστικών μοντέλων χρησιμοποιούνται για την πρόβλεψη αποτελεσμάτων σε αθλητικούς αγώνες. Ειδικά για το ποδόσφαιρο, πολύπλοκα στατιστικά μοντέλα επιστρατεύονται για την πρόβλεψη του ακριβές σκορ ενός αγώνα. Η παρούσα ανάλυση, αφορά αποκλειστικά το άθλημα του μπάσκετ.

Η δομή αυτής της μελέτης αποτελείται από 6 κεφάλαια. Στο $1^{ο}$ Κεφάλαιο επιχειρείται μία επισκόπηση των μοντέλων που χρησιμοποιούνται, ειδικά για την εξέλιξη του σκορ και το συνολικό αποτέλεσμα σε έναν αγώνα μπάσκετ. Στη συνέχεια γίνεται μια περιγραφή των συλλεχθέντων στατιστικών στοιχείων των παιχνιδιών, όπου και θα χρησιμοποιήσουμε για τις απαραίτητες αναλύσεις. Στο $2^{ο}$ Κεφάλαιο μελετώνται εξονυχιστικά τα στατιστικά στοιχεία των ομάδων, παραθέτοντας πίνακες και γραφήματα τόσο για τους νικητές, όσο και για τους ηττημένους. Θα ανακαλύψουμε κρυφές σχέσεις που υπάρχουν στα δεδομένα και θα προκύψουν χρήσιμα συμπεράσματα για τον προσδιορισμό του νικητή.

Στη συνέχεια, στο $3^{ο}$ Κεφάλαιο θα γίνουν εκτιμήσεις για την κατανομή που ακολουθούν οι πόντοι τόσο οι τελικοί, όσο και του σκορ ανά περίοδο ενός αγώνα. Τα Κεφάλαια 4 & 5 είναι αφιερωμένα στα λογιστικά και probit μοντέλα για την πρόβλεψη του τελικού αποτελέσματος. Αρχικά παρουσιάζεται η θεωρία που τα διέπει, καθώς και μέθοδοι αξιολόγησης της ποιότητας των μοντέλων αυτών. Στη συνέχεια, εξετάζεται συγκριτικά η προσαρμογή γενικευμένων γραμμικών μοντέλων στα πραγματικά δεδομένα και αναζητείται ο βέλτιστος συνδυασμός των μεταβλητών που θα προβλέψουν και θα ερμηνεύσουν το τελικό αποτέλεσμα ενός αγώνα μπάσκετ.

Στο τελευταίο Κεφάλαιο, θα γίνει μια ανακεφαλαίωση όλων των σημαντικών αποτελεσμάτων αυτής της μελέτης, ώστε να δοθεί μια πλήρης εικόνα στον αναγνώστη.

ΚΕΦΑΛΑΙΟ 1

Διαθέσιμες μέθοδοι και Δεδομένα

Σε αυτό το κεφάλαιο, θα αναφερθούν οι μέθοδοι που έχουν εφαρμογή στο μπάσκετ για την πρόβλεψη του σκορ και του τελικού αποτελέσματος και ακολούθως, θα δοθεί μια περιγραφή για την επιλογή, το είδος και την ποιότητα των δεδομένων που συλλέχθησαν.

1.1 Επισκόπηση στατιστικών μεθόδων

Παρακάτω, αναφέρονται συνοπτικά οι μέθοδοι που μπορούν να εφαρμοστούν για το τελικό αποτέλεσμα και την εξέλιξη του σκορ στο άθλημα του μπάσκετ.

α) Γραμμική Παλινδρόμηση (Linear Regression) και ανάλυση Ancova

Μπορούν να χρησιμοποιηθούν για να προβλέψουμε τις τιμές μιας συνεχούς μεταβλητής, όπως οι πόντοι που σημειώνονται ανά περίοδο, καθώς και η διαφορά αυτών για να μελετήσουμε ποια είναι τα χαρακτηριστικά που οδηγούν μια ομάδα να προηγείται στο σκορ ή ποια είναι αυτά που συμμετέχουν περισσότερο στην τελική διαμόρφωση των πόντων.

Στη συνήθη παλινδρόμηση για να συμπεριλάβουμε ως ανεξάρτητες μεταβλητές κατηγορικές θα πρέπει να δημιουργήσουμε δείκτριες μεταβλητές (dummy variables), τόσες όσες n-1 τα επίπεδα για μια μόνο κατηγορική μεταβλητή. Αυτό συνιστά μια χρονοβόρα διαδικασία, έτσι εναλλακτικά μπορούμε να προσαρμόσουμε ένα μοντέλο μέσω της ανάλυσης Ancova, όπου μας δίνεται η δυνατότητα να χρησιμοποιήσουμε και κατηγορικές και συνεχείς μεταβλητές μαζί.

Σε πρόσφατη έρευνα (Witkos, 2010) χρησιμοποιήθηκαν αυτά τα μοντέλα, ως μεταβλητή απόκρισης θεωρήθηκε οι κατατάξεις (rankings) που είχαν οι ομάδες την προηγούμενη σεζόν και ως ερμηνευτικές μεταβλητές, τα χαρακτηριστικά ενός αγώνα

(ευστοχία, λάθη κλπ). Ενώ, μία παλαιότερη έρευνα (Smith & Schwertman, 1999), μέσω απλών μοντέλων παλινδρόμησης, προσπάθησε να προβλέψει το περιθώριο νίκης (margin of victory), βάσει βαθμών κατάταξης στη διοργάνωση (seeds).

β) Μοντέλα Λογιστικής και Probit Παλινδρόμησης (Logit & Probit Models)

Ανήκουν στην κατηγορία των Γενικευμένων Γραμμικών Μοντέλων, ανάλυση και εφαρμογή των οποίων θα γίνει σε επόμενο κεφάλαιο. Μέσω αυτών τον μοντέλων, μπορούμε να εξετάσουμε ποια στοιχεία και πως αυτά επιδρούν στο τελικό αποτέλεσμα ενός αγώνα. Προϋπόθεση είναι η μεταβλητή που θέλουμε να προβλέψουμε να είναι δίτιμη, αδιαφορώντας για το είδος των ερμηνευτικών μεταβλητών. Τα τελευταία χρόνια ερευνητές, προσάρμοσαν ένα μοντέλο λογιστικής παλινδρόμησης (Kvam & Sokol, 2006) για να εκτιμήσουν την πιθανότητα, αν μία ομάδα με περιθώριο νίκης x πόντων στην έδρα της, είναι πράγματι καλύτερη από τον αντίπαλο.

Επίσης, μια μέθοδος που χρησιμοποιείται πολλές φορές για παρόμοιας φύσεως προβλήματα, είναι η διαχωριστική ανάλυση (Discriminant Analysis). Σε σχετικά πρόσφατη έρευνα (Ibáñez et al, 2009), εκτιμήθηκε ποιος είναι ο καλύτερος συνδυασμός παραγόντων, που διαχωρίζει της νικήτριες από τις ηττημένες ομάδες σε 3 σερί παιχνίδια.

γ) Αλγόριθμοι Κατηγοριοποίησης (Classification Algorithms)

Οι πιο γνωστοί ανήκουν στην κατηγορία των *Αλγόριθμων Μηχανικής Μάθησης (Machine Learning)*. Εδώ ανήκουν τα Δένδρα Απόφασης (Decision Trees) – αλγόριθμοι CHAID & CART, όπου μπορούν να κατασκευάσουν απλούς και πολύπλοκους κανόνες που να προβλέπουν και να ερμηνεύουν το αποτέλεσμα. Ενώ, εφαρμόζονται και πιο σύνθετοι αλγόριθμοι όπως SVM (Support Vector Machine) και Νευρωνικά Δίκτυα (Neural Nets). Το πλεονέκτημά τους είναι ότι, έχουν την ικανότητα να χειριστούν μεγάλου όγκου δεδομένα και μεγάλο αριθμό μεταβλητών, ανεξαρτήτως τύπου.

δ) Μοντέλα Κίνησης Brown (Brownian Motion Models)

Σύμφωνα με μελέτη (Stern, 1994), αυτά τα μοντέλα, μπορούν να χρησιμοποιηθούν για τη μελέτη της εξέλιξης του σκορ ενός αγώνα, με το ποια είναι η πιθανότητα νίκης σε ένα συγκεκριμένο χρονικό σημείο (π.χ. στο ημίχρονο). Σε αντίθεση με το

ποδόσφαιρο όπου ο αριθμός των γκολ ανά αγώνα έχει βρεθεί ότι ακολουθεί την κατανομή Poisson, στο μπάσκετ έχουμε συνεχείς μεταβολές στο σκορ ανά λεπτό, με αποτέλεσμα να εφαρμόζονται ικανοποιητικά μοντέλα κίνησης Brown.

Η διαφορά ανάμεσα στα σκορ των γηπεδούχων και των φιλοξενούμενων ομάδων σε μια διοργάνωση μπάσκετ, μοντελοποιείται ως μια διαδικασία κίνησης Brown, ορισμένη στο $t \in (0, 1)$, με μια τάση μ πόντων υπέρ της γηπεδούχου ομάδας και με μια διακύμανση σ^2. Το μοντέλο εκφράζει μια απλή σχέση, ανάμεσα στο προβάδισμα l της γηπεδούχου ομάδας στο χρόνο t και τη πιθανότητα αυτής να νικήσει.

ε) Μοντέλα Τυχαίου Περιπάτου (Random Walk Models)

Μελέτες (Gabel & Redner, 2011) έδειξαν ότι, το σκοράρισμα στο μπάσκετ περιγράφεται από έναν σχεδόν αμερόληπτο, συνεχούς χρόνου τυχαίο περίπατο. Ο χρόνος μεταξύ διαδοχικών σκοραρισμάτων, ακολουθεί μια εκθετική κατανομή με μικρή μνήμη, μεταξύ διαφορετικών διαστημάτων σκοραρίσματος. Το μοντέλο τυχαίου περιπάτου περιγράφει την εξέλιξη (στις διαφορές) του σκορ ανάμεσα στις αντίπαλες ομάδες, σαν μια συνάρτηση του χρόνου.

Λαμβάνοντας υπόψιν, την κατανομή της διαφοράς των πόντων μεταξύ των αντιπάλων και το χρονικό σημείο του αγώνα που η μία ομάδα προηγείται στο σκορ και παράλληλα εισάγοντας και την ετερογένεια στη δυναμική των ομάδων, το μοντέλο εκτιμάει με επιτυχία την πιθανότητα να προκύψει μια συγκεκριμένη διαφορά στο σκορ, την πιθανότητα μία ομάδα να προηγηθεί σε ένα συγκεκριμένο χρονικό σημείο του αγώνα και την πιθανότητα νίκης.

στ) Μαρκοβιανά μοντέλα (Markov Models)

Έρευνες (Shirley, 2007) έδειξαν ότι ένα παιχνίδι μπάσκετ, μοντελοποιείται ως μία ακολουθία μεταβάσεων μεταξύ διακριτών καταστάσεων. Συγκεκριμένα, το μοντέλο είναι μια αλυσίδα Markov που ορίζει ότι η κατανομή πιθανότητας της επόμενης κατάστασης, εξαρτάται μόνο από τη σημερινή κατάσταση. Αυτά τα μοντέλα, μπορούν να γίνουν πολύ πολύπλοκα. Μέσω προσομοίωσης, μπορούν να εκτιμηθούν οι πιθανότητες νίκης για μια ομάδα ή η μεταβολή στην πιθανότητα νίκης σαν συνάρτηση διαφόρων γεγονότων (π.χ. του αριθμού των κατοχών σε ένα παιχνίδι).

Οι καταστάσεις ορίστηκαν σε όρους τριών παραγόντων. α. Ποια ομάδα έχει την κατοχή (2): γηπεδούχος, φιλοξενούμενη, β. Πως η ομάδα απέκτησε την κατοχή (5): επαναφορά μπάλας, κλέψιμο, επιθετικό και αμυντικό ριμπάουντ, ελεύθερες βολές και

γ. Ο αριθμός των πόντων που επετεύχθησαν στην προηγούμενη κατοχή (4): 0, 1, 2, 3. Έτσι το μεγαλύτερο μοντέλο μπορεί να έχει 2 x 5 x 4 = 40 καταστάσεις, στην πράξη όμως είναι δυνατές οι 30 και με εξάλειψη των σπάνιων γεγονότων μειώνονται σε 18. Οι υπόλοιπες πιθανότητες μετάβασης μπορούν να εκτιμηθούν, χρησιμοποιώντας στατιστικά όπως λάθη, ευστοχία σε σουτ 2 και 3 πόντων κλπ.

ζ) Μπευζιανά μοντέλα (Bayesian Models) και Πιθανολογικά μοντέλα (Probabilistic Models)

Τελευταία, είναι σύνηθες η χρήση Μπευζιανών μοντέλων, για την πρόβλεψη αποτελεσμάτων στα σπορ. Εφαρμογές στη βιβλιογραφία (Parmesan & Mooney, 2011) έχουν το Bayesian Network και το Bayesian Logic. Σε μια από τις εφαρμογές του Bayesian Network, δίνονται δύο προσεγγίσεις, όπου η μία είναι χρήση μόνο των στατιστικών των παικτών και η άλλη είναι σε συνδυασμό με απλά στατιστικά της ομάδας.

Τέλος, βασικός σκοπός των πιθανολογικών / μοντέλων πιθανότητας (Probability Models) είναι, να ενσωματώσουν τη σχετική δύναμη των ομάδων στον υπολογισμό της πιθανότητας νίκης για κάθε ομάδα σε κάθε παιχνίδι. Μία μέθοδος είναι η Σταθμισμένη Πιθανοφάνεια (Weighted Likelihood) για να εκτιμηθούν τα ποσοστά νίκης της ομάδας στα τέλη της σεζόν.

Σε αυτήν τη μελέτη, θα γίνει εφαρμογή και λεπτομερής ανάλυση Λογιστικών και Probit μοντέλων, στην προσπάθειά μας να βρούμε τους καταλυτικούς παράγοντες που επιδρούν συνδυαστικά στο τελικό αποτέλεσμα ενός παιχνιδιού μπάσκετ.

1.2 Συλλογή στοιχείων

Για τις ανάγκες αυτής της μελέτης συλλέξαμε στοιχεία από 185 παιχνίδια του κορυφαίου Ευρωπαϊκού πρωταθλήματος καλαθοσφαίρισης Euroleague της περιόδου 2010-11. Σε αυτή τη διοργάνωση συμμετέχουν οι (θεωρητικά) καλύτερες 24 ομάδες της Ευρώπης, όπου είναι χωρισμένες σε 6 γκρουπ δυναμικότητας. Για την πρώτη φάση της διοργάνωσης, σχηματίζονται 4 όμιλοι και ο κάθε όμιλος περιλαμβάνει μία

ομάδα από κάθε γκρουπ δυναμικότητας (βλ. Παράρτημα). Σε αυτή τη φάση διεξάγονται 120 αγώνες.

Τα παιχνίδια μεταξύ των ομάδων είναι διπλά, δηλαδή διεξάγονται εντός και εκτός έδρας. Οι 4 πρώτες ομάδες από κάθε όμιλο περνάνε στην επόμενη φάση, αυτή των 16 ομάδων όπου και πάλι χωρίζονται σε 4 ομίλους από τους οποίους προκρίνονται οι 2 πρώτες. Σε αυτή τη φάση διεξάγονται 48 αγώνες. Στην επόμενη φάση των 8 (playoffs) προκρίνεται η ομάδα που θα επιτύχει 3 νίκες. Στη συνέχεια είναι η φάση των 4 (final four), όπου 2 ζευγάρια από τις 4 καλύτερες ομάδες αναμετρώνται σε έναν (knockout) αγώνα. Οι νικητές και οι χαμένοι των ζευγαριών συναντώνται στο «μεγάλο» και «μικρό» τελικό αντίστοιχα, σε έναν και μοναδικό αγώνα.

Για την ιστορία την συγκεκριμένη διοργάνωση την κατέκτησε μια Ελληνική ομάδα, ο Παναθηναϊκός, όπου αγωνιζόμενος στη Βαρκελώνη κέρδισε τη Μακάμπι Τελ Αβίβ από το Ισραήλ με σκορ 78-70, αυξάνοντας έτσι τον αριθμό διακρίσεων της Ελλάδας στο συγκεκριμένο άθλημα.

Να αναφέρουμε ότι ένα παιχνίδι μπάσκετ στην Ευρώπη αποτελείται από 4 περιόδους των 10 λεπτών. Στους αγώνες μπάσκετ δεν υπάρχουν ισόπαλα αποτελέσματα. Έτσι σε κάθε ισόπαλο αγώνα ακολουθεί παράταση 5 λεπτών και σε περίπτωση νέας ισοπαλίας ακολουθεί και άλλη πεντάλεπτη παράταση μέχρι την ανάδειξη νικητή.

1.3 Δεδομένα

Όπως σε όλα τα αθλήματα, η έδρα της ομάδας θεωρείται ότι παίζει σπουδαίο ρόλο στον καθορισμό του νικητή και η γηπεδούχος ομάδα έχει το πλεονέκτημα. Έτσι για την ανάλυση αυτή, δε συλλέξαμε στοιχεία από τους 4 αγώνες της τελικής φάσης, καθώς οι αγώνες διεξάγονται σε ουδέτερη έδρα και δεν ευσταθούν οι όροι γηπεδούχος και φιλοξενούμενη ομάδα. Το ζητούμενο είναι ποια χαρακτηριστικά επηρεάζουν τη γηπεδούχο ομάδα, στο να φτάσει στη νίκη (ή ήττα) σε έναν αγώνα μπάσκετ. Επομένως, χρειαζόμαστε πληροφορίες τόσο για την γηπεδούχο ομάδα, όσο όμως και για τον αντίπαλο.

Συλλέξαμε τα διαθέσιμα στατιστικά στοιχεία για την γηπεδούχο ομάδα μετά το πέρας του αγώνα, όπως πόντους, ευστοχία, λάθη κλπ. Επίσης, με σκοπό να πάρουμε έμμεσα πληροφορίες και για την αντίπαλη ομάδα (φιλοξενούμενη), χρησιμοποιήσαμε

μεταβλητές με τις διαφορές αυτών των στοιχείων. Οι μεταβλητές όμως αυτές, αφορούν μόνο το τελικό αποτέλεσμα (συμπεριλαμβάνεται η τυχόν παράταση). Έτσι, θα εισάγουμε και τις διαφορές του σκορ στις περιόδους, για να εξετάσουμε την εξέλιξη του σκορ στον αγώνα.

Τα στατιστικά στοιχεία που συλλέχθησαν εκφράζονται σε 32 μεταβλητές, όπου οι 29 είναι ποσοτικές και οι 3 είναι ποιοτικές. Επίσης δεν υπήρχε καμία ελλιπής τιμή στα δεδομένα (missing value). Ακολουθεί συγκεντρωτικός πίνακας αυτών των στοιχείων.

Πίνακας 1.1 *Ορισμός μεταβλητών*

Συνεχείς μεταβλητές

Pts: Συνολικοί πόντοι ομάδας (Points)
X2Fg: Ποσοστό ευστοχίας σουτ 2 πόντων (2-point Field Goals Percentage)
X2Fg_at: Προσπάθειες για σουτ 2 πόντων (2-point Field Goals Attempted)
X3Fg: Ποσοστό ευστοχίας σουτ 3 πόντων (3-point Field Goals Percentage)
X3Fg_at: Προσπάθειες για σουτ 3 πόντων (3-point Field Goals Attempted)
Ft: Ποσοστό ευστοχίας ελευθέρων βολών (Free Throw Percentage)
Ft_at: Προσπάθειες ελευθέρων βολών (Free Throws Attempted)
Reb_o: "Ριμπάουντ" επιθετικά (Rebounds Offensive)
Reb_d: "Ριμπάουντ" αμυντικά (Rebounds Defensive)
As: "Ασίστ" (Assists)
St: "Κλεψίματα" (Steals)
To: "Λάθη" (Turnovers)
Bl_fv: "Κοψίματα" υπέρ της ομάδας (Blocks In Favor)
Bl_ag: "Κοψίματα" κατά της ομάδας (Blocks Against)
Fl_cm: "Φάουλ" που διέπραξε η ομάδα (Fouls Committed)
Fl_rv: "Φάουλ" που δέχθηκε η ομάδα (Fouls Received)
Rkg: Ειδική αξιολόγηση επίδοσης ομάδας (Ranking)
dif_2Fg: Διαφορά στην ευστοχία σουτ 2 πόντων
dif_3Fg: Διαφορά στην ευστοχία σουτ 3 πόντων
dif_Ft: Διαφορά στην ευστοχία σουτ ελευθέρων βολών
dif_As: Διαφορά στις "ασίστ" των δυο ομάδων
dif_Reb: Διαφορά στα συνολικά "ριμπάουντ"
dif_St: Διαφορά στα "κλεψίματα"
dif_To: Διαφορά στα "λάθη"
dif_Rkg: Διαφορά στην αξιολόγηση
dif_Q1: Διαφορά πόντων μετά το πέρας της $1^{ης}$ περιόδου
dif_Q2: Διαφορά πόντων μετά το πέρας και της $2^{ης}$ περιόδου
dif_Q3: Διαφορά πόντων μετά το πέρας και της $3^{ης}$ περιόδου
dif_Pts: Διαφορά πόντων τελικού σκορ

Ποιοτικές μεταβλητές

Win: Αποτέλεσμα παιχνιδιού
Group: Γκρουπ δυναμικότητας
dif_Group: Διαφορά των γκρουπ δυναμικότητας των δυο ομάδων

Η ονομαστική μεταβλητή Win είναι δίτιμη, παίρνοντας τιμές 0 για ήττα και 1 για νίκη της γηπεδούχου ομάδας αντίστοιχα. Η διατάξιμη μεταβλητή Group έχει 6 επίπεδα, με το $1^{ο}$ επίπεδο να είναι το υψηλότερο ιεραρχικά και το $6^{ο}$ το χαμηλότερο (βλ. Παράρτημα). Η διατάξιμη dif_Group $\in$[-5,5] έχει 11 επίπεδα, με αρνητικές τιμές να εκφράζουν ότι η γηπεδούχοι ανήκουν σε υψηλότερο γκρουπ δυναμικότητας από τον αντίπαλο, θετικές τιμές το ακριβώς αντίθετο και μηδέν αν ανήκουν στο ίδιο γκρουπ.

Η αξιολόγηση της Euroleague λαμβάνει υπόψιν της αρκετούς παράγοντες και διαμορφώνεται με βάση την παρακάτω εξίσωση:

$$\text{Rkg} = \big(\text{Pts} + (\text{Reb\_o} + \text{Reb\_d}) + \text{As} + \text{St} + \text{Bl\_fv} + \text{Fl\_rv}\big) - \big((2\text{Fg\_at} - 2\text{Fg}) + \big(3\text{Fg\_at} - 3\text{Fg}\big) + \big(\text{Ft\_at} - \text{Ft}\big) + \text{To} + \text{Bl\_ag} + \text{Fl\_cm}\big)$$

Όπου Ft, 2Fg, 3Fg τα επιτυχημένα καλάθια 1, 2 και 3 πόντων αντίστοιχα.

Οι μεταβλητές dif_Pts και dif_Rkg δε μπορούν να θεωρηθούν στοχαστικές, παρά «ντετερμιστικές», γιατί οι τιμές τους προσδιορίζουν τον τελικό νικητή.

Τέλος, για όλους τους στατιστικούς ελέγχους θα χρησιμοποιηθεί επίπεδο σημαντικότητας 0.05 και η επεξεργασία των δεδομένων θα πραγματοποιηθεί μέσω του λογισμικού ανοιχτού κώδικα «R» (www.r-project.org).

ΚΕΦΑΛΑΙΟ 2

Περιγραφική Ανάλυση

Στο παρόν κεφάλαιο, θα γίνει μια περιγραφή για τις τιμές που μπορεί να πάρει κάθε μεταβλητή και τις σχέσεις που έχουν αυτές μεταξύ τους. Τα διαγράμματα και οι πίνακες συνάφειας, θα δώσουν χρήσιμες πληροφορίες για τον προσδιορισμό της νικήτριας ομάδας

2.1 Βασικά στατιστικά περιγραφικά μέτρα

Σε αυτήν την ενότητα, θα περιγράψουμε κάποια στοιχεία που παρουσιάζουν ενδιαφέρον. Ο πλήρης πίνακας (μαζί με τις απόλυτες διαφορές των μεταβλητών) με τα κυριότερα στατιστικά περιγραφικά μέτρα (μέση τιμή, διάμεσος, τυπική απόκλιση, μέγιστη - ελάχιστη τιμή), τόσο για τους νικητές (Win=No) και τους ηττημένους (Win=Yes), όσο και για το σύνολο των παιχνιδιών (Total), παρατίθεται στο Παράρτημα. Εκεί βρίσκεται και πίνακας με τα στατιστικά του νικητή της διοργάνωσης.

Ο παρακάτω πίνακας (σε δύο μέρη) παρουσιάζει μια συνοπτικότερη εκδοχή του πλήρους πίνακα, με κάποιες επιλεγμένες μεταβλητές.

Πίνακας 2.1 *Βασικά περιγραφικά μέτρα των μεταβλητών*

Win		*Pts*	*dif_Q1*	*dif_Q2*	*dif_Q3*	*dif_Pts*	*Rkg*	*dif_Rkg*	*dif_Reb*	*dif_As*	*dif_St*	*dif_To*
No	N	65										
	Mean	69.08	-2.78	-3.92	-5.11	-8.72	66.35	-19.89	-2.34	-2.00	-1.49	1.46
	SD	7.85	5.63	8.54	8.29	5.62	12.60	16.13	8.08	5.14	4.17	5.98
	Min	54	-14	-23	-27	-21	29	-56	-22	-14	-12	-14
	Max	89	10	12	10	-1	98	15	17	9	7	17
Yes	N	120										
	Mean	81.13	4.25	7.02	9.33	13.04	93.02	30.27	3.58	3.63	1.82	-2.45
	SD	9.06	6.33	7.88	9.05	10.44	15.82	25.61	8.46	5.45	4.02	4.65
	Min	60	-8	-7	-8	1	61	-20	-15	-10	-8	-12
	Max	104	20	38	39	49	137	128	28	20	15	10
Total	N	185										
	Mean	76.89	1.78	3.17	4.25	5.39	83.65	12.65	1.50	1.65	.65	-1.08
	SD	10.38	6.95	9.64	11.16	13.79	19.49	33.04	8.78	5.98	4.36	5.47
	Min	54	-14	-23	-27	-21	29	-56	-22	-14	-12	-14
	Max	104	20	38	39	49	137	128	28	20	15	17

Win		**2Fg**	**2Fg_at**	**3Fg**	**3Fg_at**	**Ft**	**Ft_at**	**dif_2Fg**	**dif_3Fg**	**dif_Ft**
No	N	65								
	Mean	48.68	37.66	31.13	20.45	75.79	17.78	-5.04	-5.10	2.25
	Std. Dev	7.83	6.21	9.75	4.44	10.99	5.81	9.47	14.66	18.07
	Min	31.8	22	5.0	11	33.3	7	-23.6	-51	-50.0
	Max	66.6	51	52.9	30	100.0	33	15.6	22	43.7
Yes	N	120								
	Mean	53.59	40.70	36.33	19.36	74.83	21.65	5.58	6.27	2.95
	Std. Dev	7.76	6.26	10.37	4.49	10.68	7.39	11.65	14.11	15.19
	Min	37.2	29	10.0	9	40.9	9	-16.0	-33	-36.8
	Max	74.2	57	66.6	32	95.4	39	41.7	39	50.0
Total	N	185								
	Mean	51.86	39.63	34.50	19.74	75.17	20.29	1.85	2.27	2.70
	Std. Dev	8.11	6.39	10.43	4.49	10.77	7.10	12.03	15.27	16.21
	Min	31.8	22	5.0	9	33.3	7	-23.6	-51	-50.0
	Max	74.2	57	66.6	32	100.0	39	41.7	39	50.0

Παρατηρούμε ότι:

❖ Από συνολικά 185 παιχνίδια, οι γηπεδούχοι νίκησαν στα 120 και έχασαν στα 65.

❖ Στην πλειοψηφία των μεταβλητών η διάμεσος με την μέση τιμή είχαν μικρή διαφορά.

❖ Η γηπεδούχος ομάδα επιτυγχάνει κατά μέσο όρο 77 πόντους με τυπική απόκλιση 10 πόντων, οι νικητές 81 και οι ηττημένοι 69 πόντους. Οι ελάχιστοι πόντοι που επετεύχθησαν ήταν 54 και οι μέγιστοι 104.

❖ Η ομάδα που κατάφερε να επιτύχει από 90 και άνω πόντους ήταν και νικήτρια. Ενώ από τις ομάδες που επέτυχαν 80-89 πόντους, μόνο οι 5 από αυτές (9.6%) δεν κατάφεραν να στεφθούν νικήτριες. Δύο ομάδες πέτυχαν 104 πόντους, με τη μία από αυτές να είναι η φιναλίστ της διοργάνωσης έχοντας ένα ιδιαίτερα αυξημένο μέσο όρο πόντων (89).

❖ Στις 3 πρώτες περιόδους η γηπεδούχος ομάδα προηγήθηκε κατά μέσο όρο με 2, 3 και 4 πόντους αντίστοιχα.

❖ Τα παιχνίδια έληξαν με μια μέση διαφορά περίπου 5 πόντων υπέρ του γηπεδούχου, με τυπική απόκλιση 14 πόντων και μέγιστη διαφορά 49 πόντων. Σε σχετικές έρευνες (Harville & Smith, 1994), εκτιμήθηκε ότι το προβάδισμα για τους γηπεδούχους στο τέλος του αγώνα ήταν 4.68 ±0.28 πόντους, πράγμα που σημαίνει ότι βρισκόμαστε πολύ κοντά σε αυτήν την εκτίμηση.

❖ Παρατηρούμε ότι οι νικητές είχαν κατά μέσο όρο περισσότερα ριμπάουντ, ασιστ και κλεψίματα και λιγότερα λάθη, ενώ το αντίστροφο ισχύει για τους ηττημένους. Στο σύνολο όμως οι γηπεδούχοι είχαν οριακά κατά μέσο όρο καλύτερα στατιστικά σε αυτούς τους τομείς από τους αντιπάλους.

❖ Στα δίποντα και στα τρίποντα οι νικητές γηπεδούχοι είχαν 5.6% και 6.3% αντίστοιχα παραπάνω ποσοστά ευστοχίας από τους αντιπάλους.

❖ Οι γηπεδούχοι είναι πιο εύστοχοι στις βολές (κατά μέσο όρο 2.7%) από τους φιλοξενούμενους είτε κερδίσουν, είτε χάσουν.

❖ Οι γηπεδούχοι που κέρδισαν είχαν κατά μέσο όρο παραπάνω προσπάθειες για σουτ ελευθέρων βολών και 2 πόντων και μια λιγότερη προσπάθεια στα σουτ 3 πόντων από τους γηπεδούχους που έχασαν.

❖ Οι γηπεδούχοι είναι φανερό ότι υπερτερούν στα στατιστικά σε όλους τους τομείς. Ο παράγοντας έδρα φαίνεται να παίζει σημαντικό ρόλο στην εξέλιξη ενός αγώνα.

Η πρωταθλήτρια ομάδα νίκησε στα 8 από τα 10 παιχνίδια που έδωσε στην έδρα της δηλαδή στο 80% αυτών, σημαντικά πάνω από το 64.9% επιτυχία όλων των ομάδων της διοργάνωσης. Παρατηρούμε ότι ο μέσος όρος καθώς και η διάμεσός των συνολικών πόντων είναι ελαφρώς μεγαλύτερος (κατά 2.51 και 1 πόντους αντίστοιχα) συγκριτικά με το σύνολο των ομάδων. Επίσης, η γηπεδούχος εκτελεί κατά μέσο όρο περισσότερες βολές (+6.31) και επομένως κερδίζει περισσότερα φάουλ (+4.02), έχει περισσότερα κλεψίματα (+1.14), κάνει λιγότερα λάθη (-1.07), έχει περισσότερα κοψίματα (+0.64), ενώ στα ριμπάουντ δεν παρουσιάστηκαν διαφορές. Τέλος κατάφερε να εξασφαλίσει προβάδισμα 6 πόντων στο ημίχρονο και 12 πόντων στο τέλος και της 3^{ης} περιόδου, ενώ νίκησε τους αντιπάλους της κατά μέσο όρο με 11 πόντους διαφορά.

2.2 Σχέσεις των μεταβλητών

Αρχικά θα υπολογιστούν όλες οι ανά δύο συσχετίσεις των μεταβλητών μεταξύ τους, τόσο των ποσοτικών όσο και των ποιοτικών μεταβλητών. Οι πλήρεις πίνακες των συντελεστών συσχέτισης βρίσκονται στο Παράρτημα.

Για την περιγραφή της έντασης της εξάρτησης μεταξύ των συνεχών μεταβλητών χρησιμοποιήθηκε ο γνωστός συντελεστής γραμμικής συσχέτισης του Pearson. Στη διεθνή βιβλιογραφία (Calkins, 2005) προτείνεται για την ένταση της σχέσης μεταξύ μιας δίτιμης μεταβλητής (όπως η Win) και μιας ποσοτικής μεταβλητής η χρήση του συντελεστή συσχέτισης Point-Biserial, που είναι μια ειδική περίπτωση του Pearson. Ενώ για την σχέση μιας διατάξιμης (όπως η Group) με μια ποσοτική μεταβλητή ο συντελεστής συσχέτισης Polyserial (ή ο συντελεστής του Spearman). Για την σχέση μιας δίτιμης και μιας διατάξιμης συστήνεται ο συντελεστής Rank Biserial, όπου η χρήση του έχει αντικατασταθεί από το Somers' D.

Οι τιμές όλων των παραπάνω συντελεστών συσχέτισης είναι ανάμεσα στο -1 και +1. Όσο μεγαλύτερες κατ' απόλυτη τιμή είναι οι τιμές των συντελεστών, τόσο πιο ισχυρές (αρνητικές / θετικές) θεωρούνται οι σχέσεις των μεταβλητών. Μια τιμή ίση ή πολύ κοντά στο 0, υποδηλώνει ότι δεν υπάρχει καμία σχέση, ενώ μία κοντά στο 1 ή -1 υποδηλώνει ότι η μία μεταβλητή μπορεί σχεδόν τέλεια να προβλέψει την άλλη. Το πρόσημο μας πληροφορεί για το εάν μία αύξηση της μίας μεταβλητής, οδηγεί σε

αύξηση (+) ή σε μείωση (-) της άλλης. Επίσης τετραγωνίζοντας τον συντελεστή συσχέτισης πληροφορούμαστε για το ποσό της μεταβλητότητας της μιας μεταβλητής που εξηγείται από την άλλη (R^2).

Ο χαρακτηρισμός για το βαθμό της έντασης της σχέσης είναι υποκειμενικός, ωστόσο ένας αποδεκτός πρακτικός κανόνας (Choudhury, 2009) είναι ο εξής: $|r| < 0.1$ καμία ή αμυδρή σχέση, $0.1 \le |r| < 0.3$ αδύναμη σχέση, $0.3 \le |r| < 0.5$ μέτριας εντάσεως σχέση και $|r| \ge 0.5$ ισχυρή. Αυτό ισχύει κυρίως, για τους Pearson συντελεστές συσχέτισης.

Οι μεταβλητές στη συντριπτική τους πλειοψηφία (σύμφωνα με το K-S τεστ) αποδεχόμαστε ότι ακολουθούν την κανονική κατανομή σε ε.σ. 5%, με ορισμένες εξαιρέσεις χωρίς να αποκλίνουν όμως σημαντικά από αυτήν (Ft, Reb_o, St, To, Bl_fv, Bl_ag). Έτσι καλύπτεται μια από τις βασικές υποθέσεις, αν και οι συντελεστές του Pearson είναι ιδιαίτερα ανθεκτικοί σε παραβιάσεις αυτών.

Παρατηρήσεις:

❖ Η Pts είναι έντονα θετικά συσχετισμένη κατά φθίνουσα σειρά με τις Rkg (0.886), dif_Pts, dif_Rkg, dif_Q3, As, X3Fg (0.516). Και αν για τις 4 πρώτες μεταβλητές δεν είχαμε κάποια έκπληξη, έχει ενδιαφέρον η διαπίστωση ότι οι πόντοι έχουν ισχυρή σχέση με τις Ασίστ και το ποσοστό ευστοχίας των σουτ 3 πόντων.

❖ Η διαφορά του σκορ που διαμορφώνεται στην $3^{η}$ περίοδο, φαίνεται να επηρεάζεται έντονα από τις διαφορές στις ασίστ και την ευστοχία στα δίποντα σουτ.

❖ Η Rkg συσχετίζεται έντονα θετικά κατά σειρά με τις Pts, dif_Rkg, dif_Pts, dif_Q3, As, Win, dif_As, dif_Q2, dif_2Fg, X2Fg, έτσι μεταβλητές που συνεισφέρουν πιο πολύ στη βαθμολογία είναι οι πόντοι, οι ασίστ και η ευστοχία στα σουτ 2 πόντων.

❖ Ενδιαφέρον επίσης έχει, η πολύ ισχυρή (αρνητική) σχέση ανάμεσα στη διαφορά των λαθών (dif_To) και στη διαφορά των κλεψιμάτων (dif_St), με τιμή του συντελεστή του Pearson ίση με -0.755, παρόλο που St και To εμφανίζονται να μην έχουν καθόλου σχέση με συντελεστή σχεδόν 0.

❖ Οι ασίστ (As), τα λάθη (To), τα ριμπάουντ (Reb_o, Reb_d) και τα κλεψίματα (St) δε φαίνεται να έχουν σημαντική σχέση μεταξύ τους με συσχετίσεις $|r|<0.1$, με εξαίρεση τον συντελεστή r(As,St)=0.251 που φανερώνει μια ασθενής σχέση ανάμεσα στην As και την St

❖ Μάλλον ένα περίεργο εύρημα είναι ότι, οι μεταβλητές που εκφράζουν ποσοστά ευστοχίας δε φαίνεται να συνδέονται μεταξύ τους (εξαίρεση αποτελούν οι X2Fg-dif_2Fg, X3Fg-dif_3Fg, Ft-dif_Ft), ενώ αντίθετα οι προσπάθειες έχουν κάποια σχέση.

❖ Οι μεταβλητές X2Fg_at, X3Fg_at, Bl_fv, Bl_ag, Fl_cm δε συσχετίζονται ισχυρά με καμία άλλη μεταβλητή.

❖ Βάσει των Point-Biserial συντελεστών, η μεταβλητή Win είναι έντονα θετικά συσχετισμένη με τις μεταβλητές dif_Pts (0.754), dif_Rkg (0.725), Rkg (0.653), dif_Q3 (0.617), Pts (0.554), dif_Q2 (0.542). Επίσης μέτριας εντάσεως είναι η σχέση του αποτελέσματος με τη διαφορά πόντων στην $1^{η}$ περίοδο, τη διαφορά στις ασίστ και την διαφορά ευστοχίας στα δίποντα (τιμές $0.4 \leq r_{pb} < 0.5$). Ενώ, οι αμέσως επόμενες σε τιμές (κοντά στο 0.350) είναι οι διαφορές στα κλεψίματα, στην ευστοχία τριπόντων και στα λάθη. Όπως είναι λογικό, η διαφορά των πόντων στο τέλος της $2^{ης}$ και $3^{ης}$ περιόδου και οι τελικοί πόντοι είναι πολύ σημαντικοί παράγοντες στον καθορισμό του αποτελέσματος.

❖ Το τελικό αποτέλεσμα (Win) δείχνει να μην έχει σχέση με τις μεταβλητές Ft, Reb_o, Bl_fv, Fl_cm και dif_Ft, ενώ εξαιρετικά ασθενής είναι η σχέση με X3Fg_at, Bl_ag και Fl_rv.

❖ Η τελική διαφορά πόντων (dif_Pts) που προσδιορίζει το τελικό αποτέλεσμα καταλήγει σε παρόμοια συμπεράσματα για τη σχέση της με τις υπόλοιπες μεταβλητές, όπου μετά τους πόντους, οι διαφορές στην ευστοχία διπόντων (0.628) και στις ασίστ (0.599) να ξεχωρίζουν. Έπειτα με μέτριας εντάσεως σχέση ακολουθούν οι διαφορές στα ριμπάουντ, στα κλεψίματα, στην ευστοχία τριπόντων και στα λάθη.

❖ Βάσει των Polyserial συντελεστών, η μεταβλητή Group είναι μέτρια αρνητικά συσχετισμένη (με τιμές $0.3 \leq r_b < 0.4$) με τις dif_Rkg, dif_Q3, Rkg, dif_2Fg. Και ακολουθούν οι dif_Pts dif_Q1, dif_Q2, dif_As, Pts, με τις υπόλοιπες να έχουν αμυδρή έως καθόλου σχέση με τα γκρουπ δυναμικότητας.

❖ Παρόμοια συμπεράσματα ισχύουν για τις διαφορές των γκρουπ (dif_Group).

❖ Η τιμή του συντελεστή Somers' D = -0.226 μαρτυρά μια αρνητική συσχέτιση ασθενούς εντάσεως μεταξύ της Win και Group. Τα νικητήρια αποτελέσματα (τιμές 1 της Win) τείνουν να αντιστοιχούν σε υψηλότερα γκρουπ δυναμικότητας (το $1^{ο}$ είναι το υψηλότερο).

Την σημαντικότητα της σχέσης των δύο μεταβλητών θα την επιβεβαιώσουμε με έναν χ^2 έλεγχο.

Πίνακας 2.2 *Έλεγχος χ^2 για τη σχέση αποτελέσματος και γκρουπ δυναμικότητας*

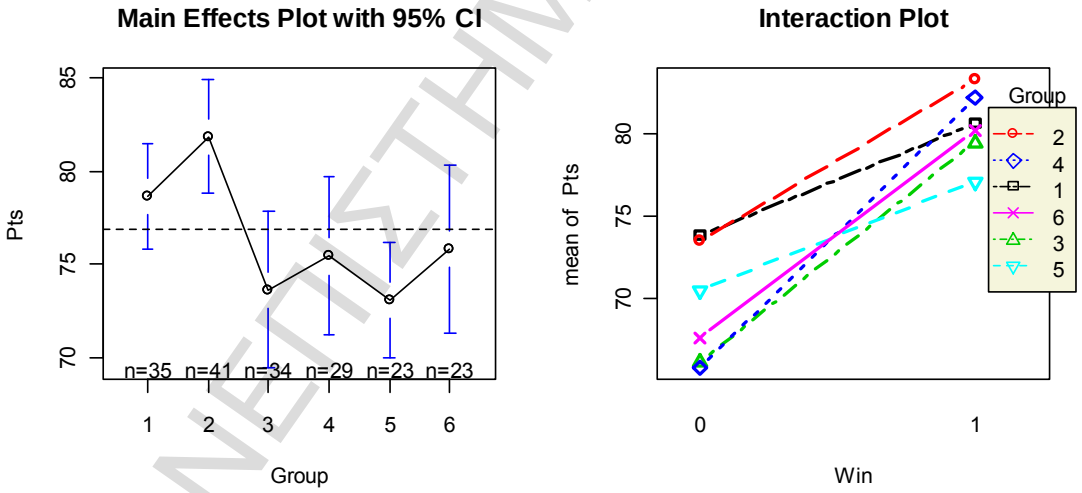
```
        Pearson's Chi-squared test

data:  contingency_table
X-squared = 16.6073, df = 5, p-value = 0.005308
```

Με p-value < 0.05, απορρίπτουμε τη μηδενική υπόθεση ότι το τελικό αποτέλεσμα δεν έχει σχέση με το γκρουπ δυναμικότητας που ανήκει η ομάδα.

Θα είχε ενδιαφέρον να ελέγξουμε τη σχέση ανάμεσα στους πόντους που επιτυγχάνονται (συνεχή μεταβλητή) και στα γκρουπ δυναμικότητας (κατηγορική μεταβλητή). Έτσι θα προχωρήσουμε σε ανάλυση της διασποράς (One-Way Anova). Αρχικά όμως, διαγραμματικά μπορούμε να πάρουμε κάποιες ενδείξεις.

Σχήμα 2.1 *Διαγράμματα κύριων επιδράσεων και αλληλεπιδράσεων των πόντων*



Στο 1^{ο} διάγραμμα οι πόντοι φαίνεται να επηρεάζονται από το γκρουπ δυναμικότητας, καθώς κάποιες μέσες τιμές των πόντων στα γκρουπ φαίνεται να διαφέρουν αρκετά μεταξύ τους και από το γενικό μέσο όρο των πόντων (διακεκομμένη γραμμή). Το 2^{ο} γκρουπ δυναμικότητας εμφανίζει τον μεγαλύτερο μέσο όρο πόντων, ενώ το 5^{ο} τον μικρότερο.

Στο 2^{ο} διάγραμμα οι ευθείες δεν εμφανίζονται παράλληλες και φαίνεται να υπάρχει αλληλεπίδραση των γκρουπ δυναμικότητας με το αποτέλεσμα του αγώνα στη διαμόρφωση των πόντων.

Πριν προχωρήσουμε σε ανάλυση Anova και σε post-hoc ελέγχους, θα ελέγξουμε πρώτα αν η μεταβλητή Pts ακολουθεί την Κανονική κατανομή και εάν έχει σταθερή διακύμανση σε κάθε επίπεδο της Group.

Πίνακας 2.3 *Ανάλυση διασποράς των πόντων στα επίπεδα των γκρουπ δυναμικότητας και έλεγχος του Tukey*

```
        Shapiro-Wilk normality test

data:  Pts[Group==1:6]
W = 0.9823, p-value = 0.8303
W = 0.9624, p-value = 0.1914
W = 0.9703, p-value = 0.4692
W = 0.9721, p-value = 0.6177
W = 0.9678, p-value = 0.6356
W = 0.9748, p-value = 0.8025


Levene's Test for Homogeneity of Variance (center = "mean")

       Df F value  Pr(>F)

group   5  1.9378 0.09022 .

      179

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Analysis of Variance Table

Response: Pts
           Df  Sum Sq Mean Sq F value   Pr(>F)
Group       5  1915.4  383.07  3.8255 0.002567 **
Residuals 179 17924.5  100.14
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Pts ~ Group)

$Group
```

```
          diff        lwr        upr     p adj
2-1  3.2209059  -3.413423  9.8552350 0.7278432
3-1 -5.0394958 -11.981218  1.9022264 0.2964816
4-1 -3.2433498 -10.482259  3.9955595 0.7898033
5-1 -5.5701863 -13.308249  2.1678760 0.3057981
6-1 -2.8310559 -10.569118  4.9070064 0.8986753
3-2 -8.2604017 -14.947157 -1.5736467 0.0062576
4-2 -6.4642557 -13.459039  0.5305275 0.0881354
5-2 -8.7910923 -16.301272 -1.2809130 0.0116021
6-2 -6.0519618 -13.562141  1.4582175 0.1909267
4-3  1.7961460  -5.490841  9.0831328 0.9805376
5-3 -0.5306905  -8.313748  7.2523666 0.9999591
6-3  2.2084399  -5.574617  9.9914970 0.9640181
5-4 -2.3268366 -10.376077  5.7224042 0.9610495
6-4  0.4122939  -7.636947  8.4615347 0.9999901
6-5  2.7391304  -5.761815 11.2400762 0.9387662
```

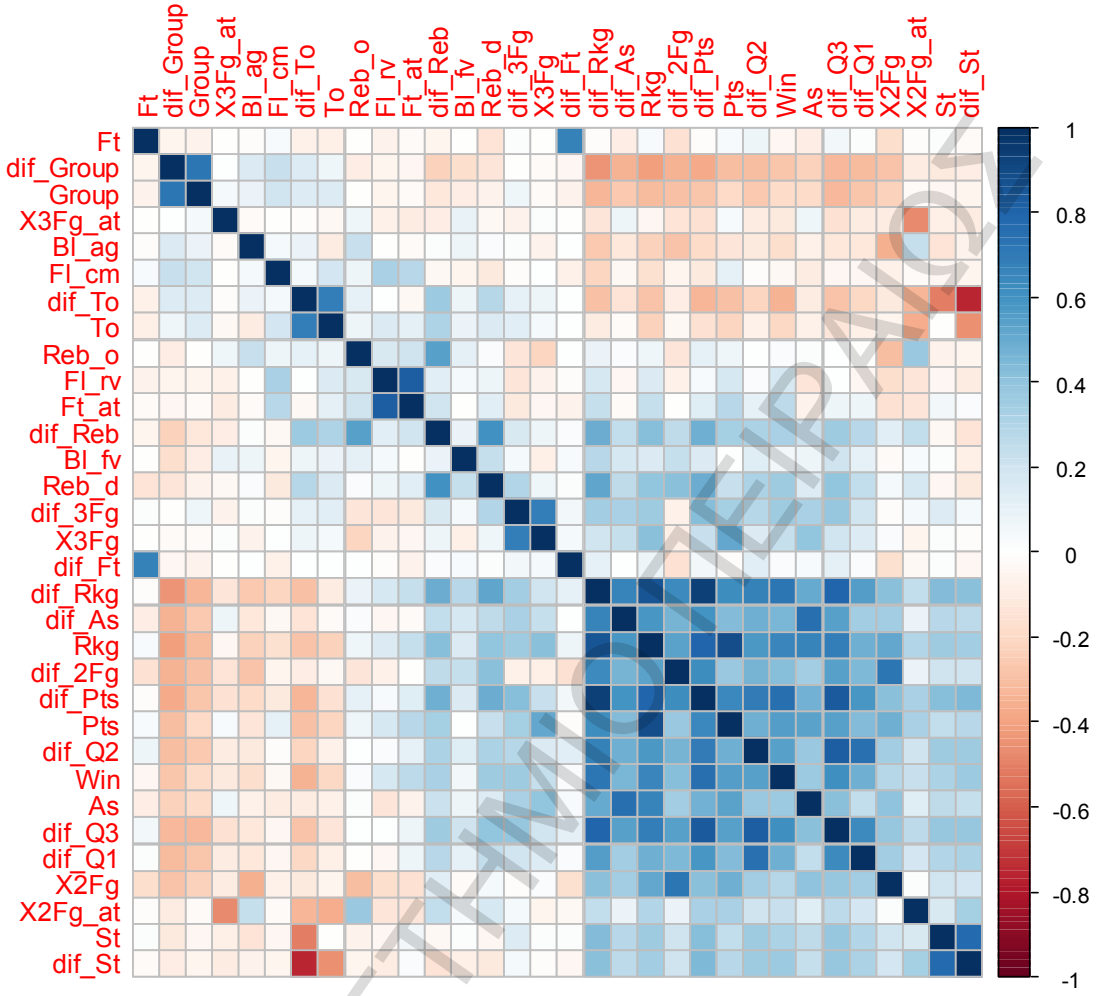
Από τον έλεγχο Shappiro-Wilk, δεν απορρίπτουμε ότι οι πόντοι στα γκρουπ δυναμικότητας ακολουθούν την Κανονική κατανομή και από το τεστ του Levene δεν απορρίπτουμε σε ε.σ. 5%, ότι οι πόντοι στα γκρουπ έχουν σταθερή διακύμανση και έτσι προχωράμε στον παραμετρικό έλεγχο Anova. Απορρίπτουμε ότι η μέση τιμή των πόντων σε κάθε γκρουπ δε διαφέρει. Από τον έλεγχο του Tukey, στατιστικά μόνο οι πόντοι στα γκρουπ 2-3 και 2-5 φαίνεται να διαφέρουν σημαντικά. Παρόμοια αποτελέσματα προκύπτουν και για την τελική διαφορά του σκορ με τα γκρουπ δυναμικότητας.

2.3 Γραφική Ανάλυση

Θα διερευνήσουμε γραφικά, τόσο τη συμπεριφορά των μεταβλητών σε σχέση με το τελικό αποτέλεσμα, όσο και τη σχέση των μεταβλητών μεταξύ τους

Σε αυτό το σημείο, θα παραθέσουμε ένα γράφημα το οποίο θα συγκεντρώσει όλη την πληροφορία των σχέσεων, μεταξύ των μεταβλητών που είδαμε παραπάνω. Είναι ένα γράφημα, το οποίο οπτικοποιεί έναν πίνακα συσχετίσεων και μάλιστα αναλυτές δεδομένων (βλ. Βιβλιογραφία – Σύνδεσμοι) το χρησιμοποιούν, ανεξαρτήτως του τύπου των μεταβλητών. Η θέση που πήραν οι μεταβλητές στο συγκεκριμένο γράφημα, στηρίχθηκε σε έναν αλγόριθμο γωνιακής διάταξης των ιδιοδιανυσμάτων, ώστε να βελτιωθεί το αισθητικό αποτέλεσμα.

Σχήμα 2.2 Συγκεντρωτικό *γράφημα συσχετίσεων*

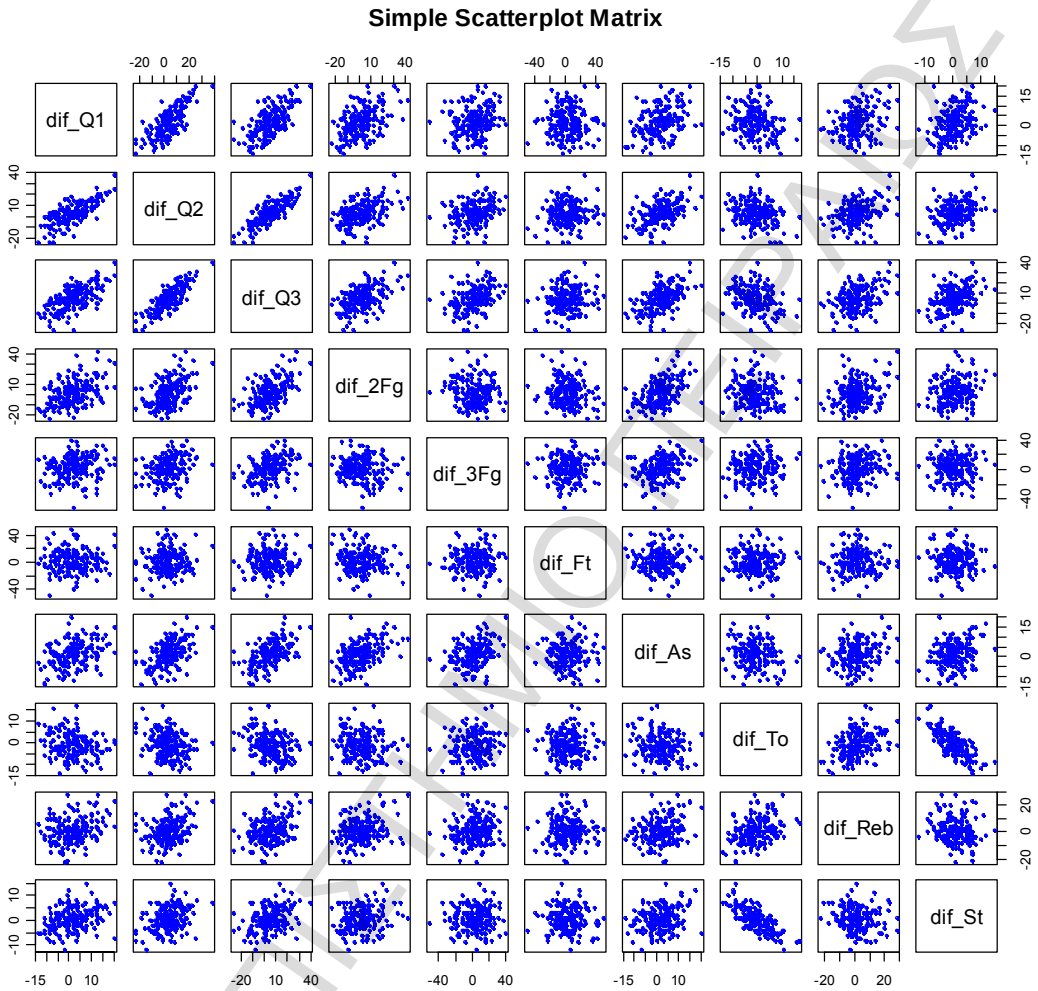


Μπορούμε να δούμε ότι, οι ισχυρότερες θετικές συσχετίσεις συγκεντρώνονται κυρίως στην κάτω δεξιά γωνία (μπλε χρώμα) και κοντά στη διαγώνιο, ενώ οι ισχυρότερες αρνητικές (κόκκινο χρώμα), συγκεντρώνονται στην πάνω δεξιά και στην κάτω αριστερή γωνία. Όσο πιο ανοικτός ο χρωματισμός, τόσο πιο αδύναμη η σχέση. Οι μεταβλητές στο γράφημα από την dif_Rkg έως και την X2Fg, εμφανίζονται να έχουν από μέτρια έως πολύ ισχυρή σχέση μεταξύ τους, κάτι λογικό αφού όλες έχουν άμεση σχέση με το σκοράρισμα.

Για να εξετάσουμε αν οι συνεχείς μεταβλητές, συνδέονται ανά δύο με κάποια άλλη σχέση μη γραμμική, θα πρέπει να κατασκευάσουμε διαγράμματα διασποράς για

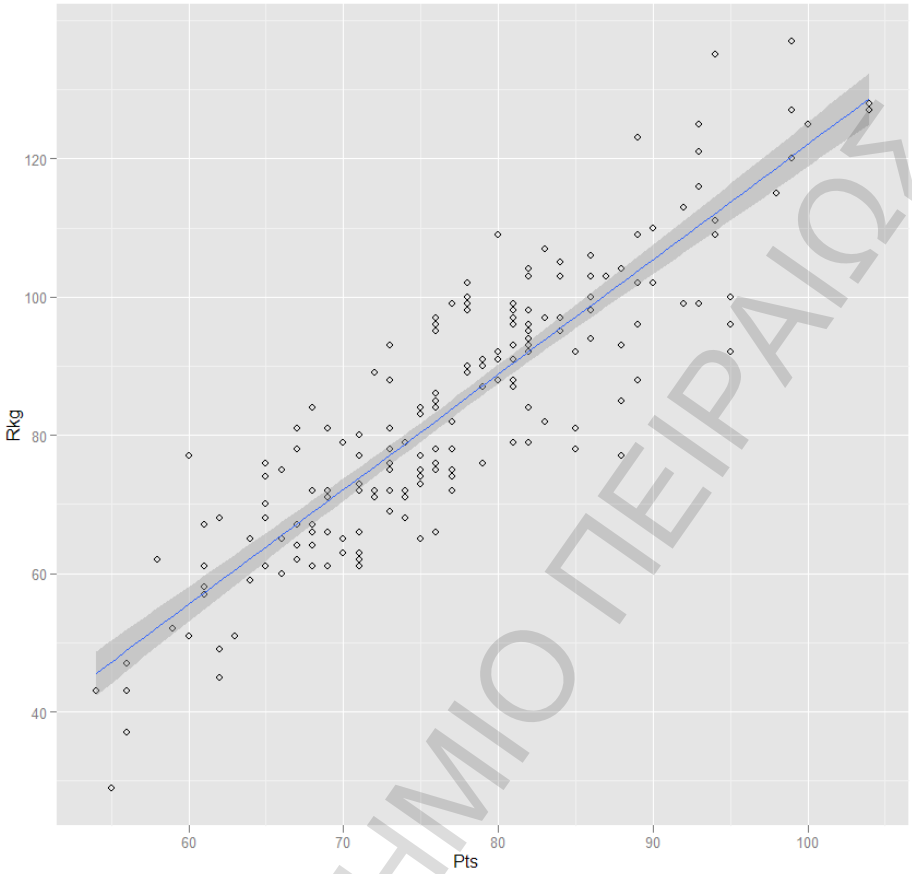
όλους τους ανά δύο συνδυασμούς. Παρατίθεται ένα μικρό δείγμα αυτών των ανά δύο σχέσεων, χρησιμοποιώντας τις μεταβλητές που εκφράζουν διαφορές.

Σχήμα 2.3 *Διαγράμματα διασποράς για όλες τις κύριες διαφορές*



Τα σημεία φαίνεται να σχηματίζουν μια γραμμή, στους συνδυασμούς των dif_Q και στο συνδυασμό dif_To-dif_St, ενώ τα υπόλοιπα φαίνεται περισσότερο να είναι διασκορπισμένα στο χώρο. Μια πιο αναλυτική περιγραφή, θα δοθεί στο παρακάτω γράφημα των πόντων με την ειδική αξιολόγηση.

Σχήμα 2.4 *Διάγραμμα διασποράς για τις Pts και Rkg*

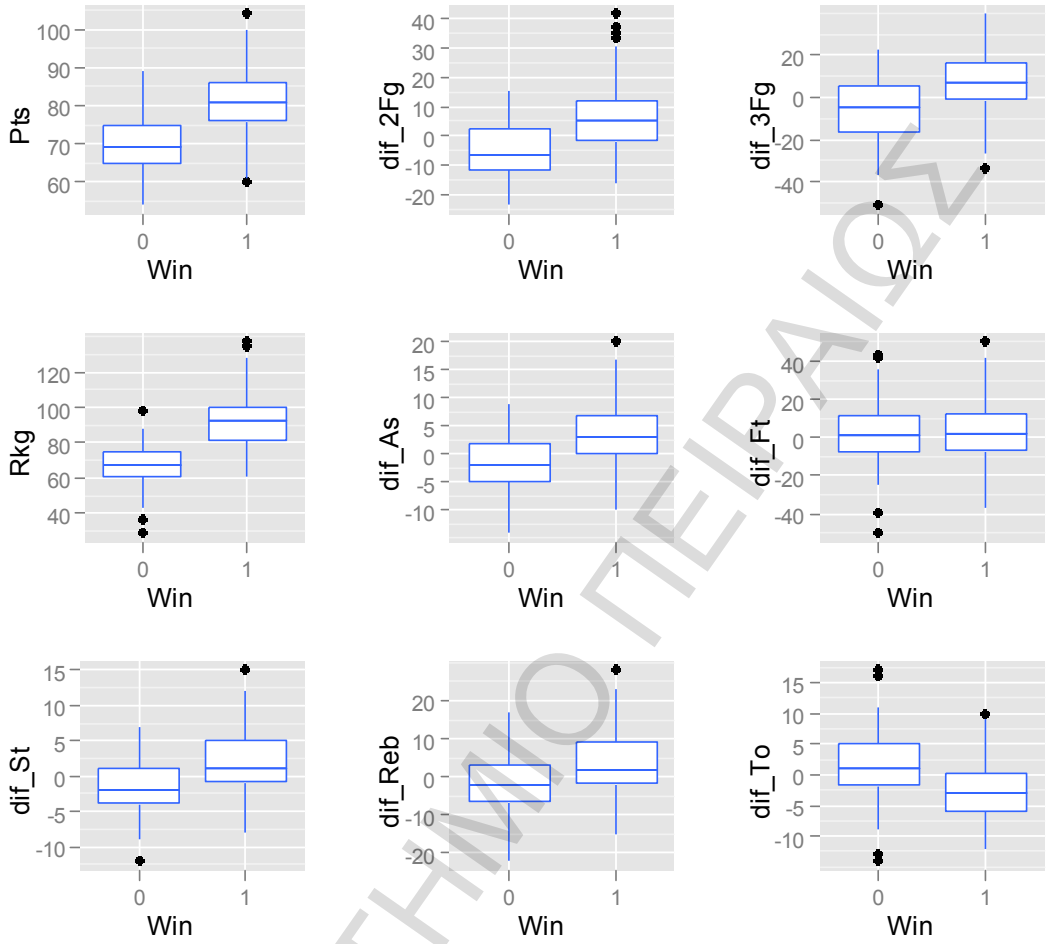


Η σχέση τους είναι γραμμική (το μοτίβο των παρατηρήσεων είναι μια ευθεία γραμμή), θετική (θετική κλίση ευθείας) και ισχυρή (συγκέντρωση των παρατηρήσεων γύρω από τη γραμμή). Η Pts επηρεάζει περισσότερο από όλους τους άλλους παράγοντες, τη δημιουργία της αξιολόγησης.

Αξίζει να αναφέρουμε ότι, κανένα διάγραμμα διασποράς σε κάθε ανά δύο συνδυασμό των συνεχών μεταβλητών, δεν απεικόνισε κάποιο μοτίβο διαφορετικό από το γραμμικό ή το τυχαίο.

Για να εξετάσουμε γραφικά, τη συμπεριφορά κάποιων βασικών συνεχών μεταβλητών, βάσει του τελικού αποτελέσματος της γηπεδούχου ομάδας, θα χρησιμοποιήσουμε θηκογράμματα (boxplots).

Σχήμα 2.5 *Θηκογράμματα βάσει της Win*



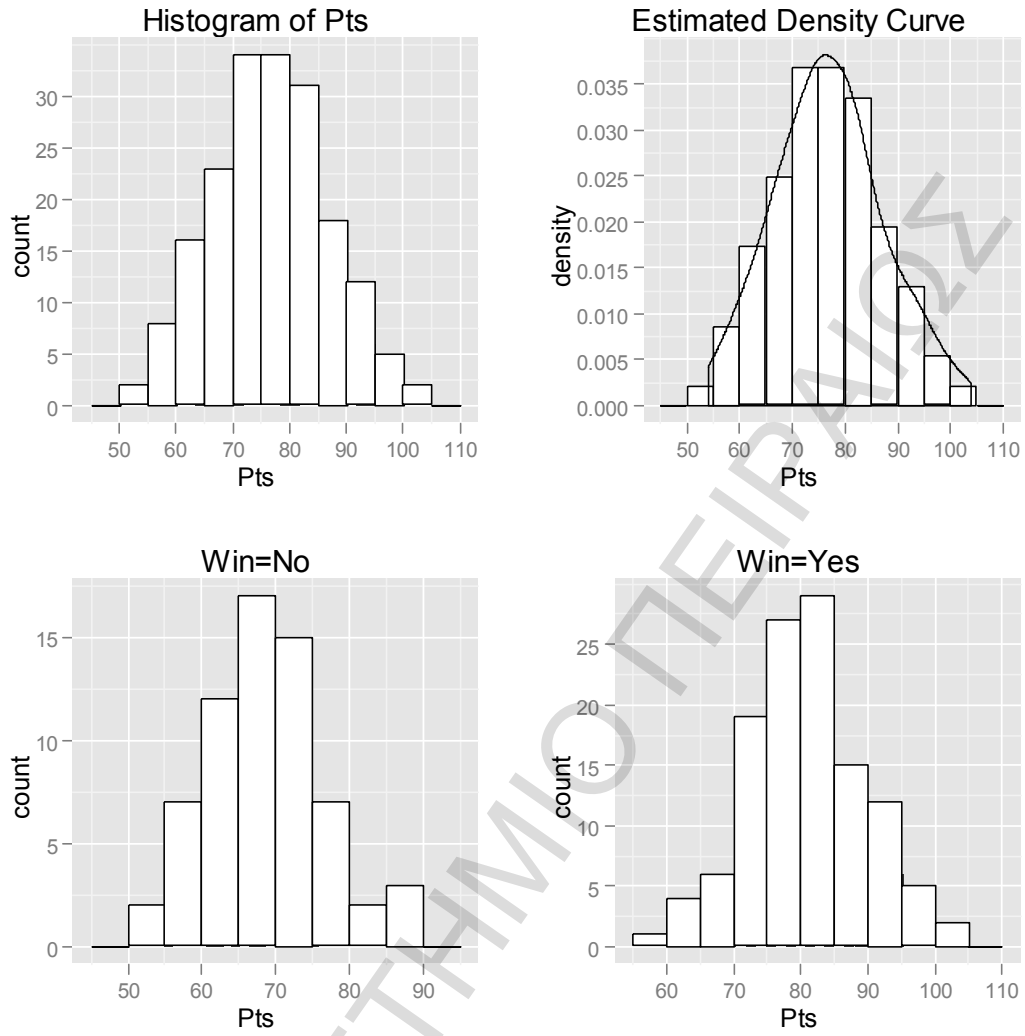
Τα παραπάνω θηκογράμματα των πόντων σε αντιστοιχία με το τελικό αποτέλεσμα μας παρέχουν τις εξής πληροφορίες:

- Η διάμεσος (μεσαία γραμμή ορθογωνίων τμημάτων) των πόντων διαφέρει σημαντικά και συγκεκριμένα κατά 14 πόντους στα νικηφόρα και μη παιχνίδια. Η μεγαλύτερη διαφορά διαμέσων, παρατηρείται στην ειδική αξιολόγηση, κάτι αναμενόμενο αφού αποτελεί ένα γενικό δείκτη απόδοσης της ομάδας. Επίσης σημαντική φαίνεται να είναι η διαφορά στην ευστοχία σε σουτ 2 και 3 πόντων, όπως και της διαφοράς λαθών. Έτσι, φαίνεται οι τιμές αυτών των μεταβλητών να επηρεάζουν το τελικό αποτέλεσμα. Από την άλλη πλευρά, εντελώς ασήμαντη φαίνεται να είναι η σχέση του αποτελέσματος, με τη διαφορά ευστοχίας ελ. βολών (ίδια διάμεσος και κοντά στο 0).

- Παρατηρούνται 3 ακραίες τιμές (κουκίδες) στο θηκόγραμμα των πόντων με τα νικηφόρα παιχνίδια, που μας πληροφορούν ότι μια ομάδα νίκησε με μόλις 60 πόντους (125η περίπτωση), ενώ δυο άλλες επέτυχαν 104 πόντους (41η και 139η περίπτωση). Αρκετές ακραίες τιμές, παρατηρούνται στη μεταβλητή dif_2Fg με τιμές πάνω από 30% διαφορά ευστοχίας για τους νικητές, ενώ δεν είναι λίγες και της dif_To με το ενδιαφέρον να εντοπίζεται στο ότι, υπήρχε ομάδα που ηττήθηκε αν και έκανε 14 λάθη λιγότερα από τον αντίπαλο. Επίσης, ενδιαφέρουσα περίπτωση είναι εκείνη, όπου ανιχνεύτηκε ομάδα που νίκησε και είχε λιγότερο από 30% ευστοχία στα τρίποντα από τους φιλοξενούμενους.
- Το εύρος (απόσταση από το ανώτερο μέχρι το κατώτερο τμήμα μαζί με τις ακραίες τιμές) για τις ηττημένες ομάδες είναι 35 πόντους, ενώ για τις νικήτριες είναι 44 πόντους αντίστοιχα. Οι πιο συνήθεις περιπτώσεις πόντων βρίσκονται εντός του ενδοτεταρτημοριακού εύρους (πλάτος ορθογώνιου σχήματος), για τους ηττημένους 65-75 πόντοι και για τους νικητές 76-86 πόντοι.
- Στις άκρες των κατακόρυφων τμημάτων, παρατηρούμε τις υψηλότερες (ή τις χαμηλότερες) τιμές που δεν μπορούν να θεωρηθούν ως ακραίες και αυτές είναι 54 και 89 πόντοι για τους ηττημένους και 61 και 100 πόντοι για τους νικητές αντίστοιχα.
- Η dif_Reb με την dif_St φαίνεται να έχουν αρκετά κοινά ως προς τις τιμές που παίρνουν, ενώ οι τιμές της dif_As, φαίνεται να διακρίνουν λίγο καλύτερα τους νικητές, από τους ηττημένους.

Παρακάτω, παρουσιάζονται κάποια ιστογράμματα για τους τελικούς πόντους (σε κλάσεις των 5 πόντων).

Σχήμα 2.6 *Ιστογράμματα για την Pts*



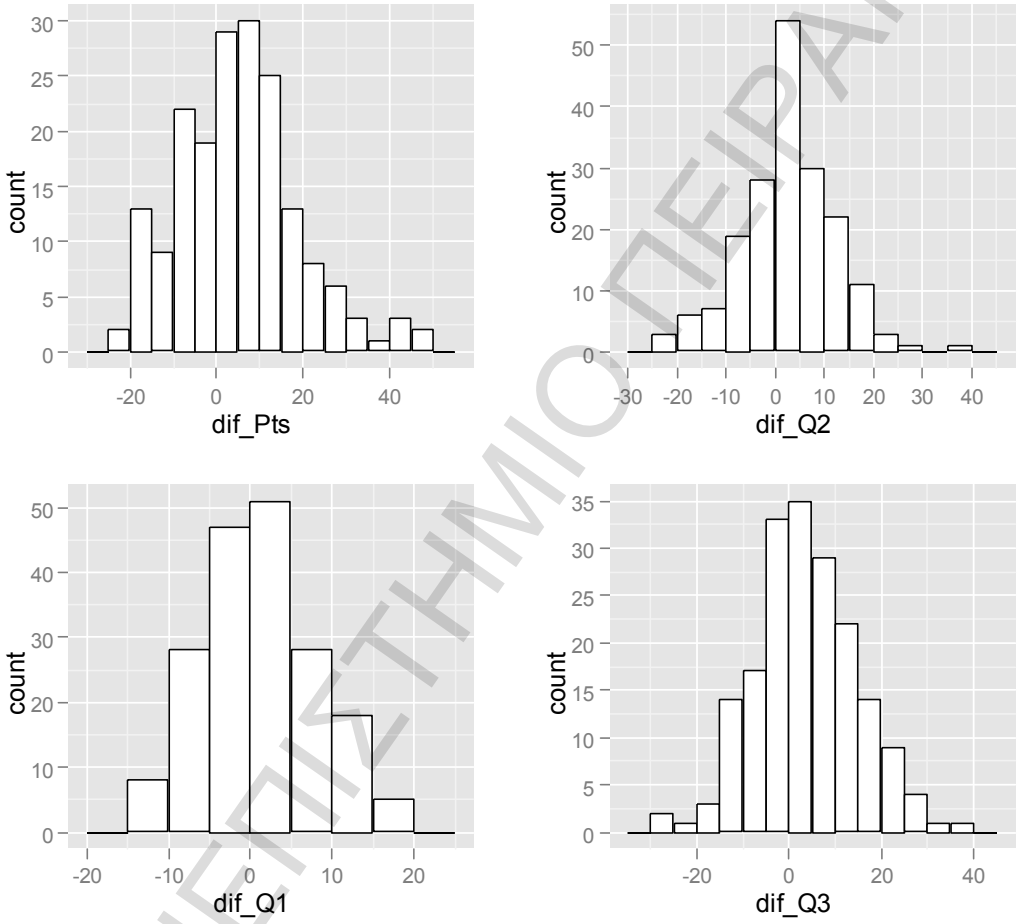
Θα μπορούσαμε να πούμε ότι, σαν πρώτη εικόνα οι πόντοι φαίνεται να κατανέμονται κανονικά, τόσο στο σύνολό των παιχνιδιών όσο και για τους νικητές και τους ηττημένους. Στο 2^{ο} ιστόγραμμα παρέχεται μια εκτίμηση της μορφής της καμπύλης, που προσεγγίζει τα δεδομένα, όπου η κορυφή εμφανίζεται αιχμηρή.

Επίσης, βάσει της συχνότητας των πόντων στα περισσότερα παιχνίδια οι γηπεδούχοι επιτυγχάνουν 70-80 πόντους, με μια συχνότητα περίπου 38.4% των παιχνιδιών (71 από τα 185). Υψηλή συχνότητα έχουν επίσης οι πόντοι 80-85 και 65-70. Οι πόντοι από 60 και κάτω, καθώς και από 95 και άνω έχουν πολύ μικρή συχνότητα. Από 100 πόντους και άνω επετεύχθησαν μόνο στο 1.6% των παιχνιδιών (3 από τα 185).

Όσον αφορά το ιστόγραμμα των ηττημένων, η μεγαλύτερη συχνότητα των πόντων παρουσιάζεται στους 65-70 πόντους (>15 παιχνιδιών), ενώ στους νικητές αντίστοιχα η μεγαλύτερη συχνότητα είναι 80-85 (30 παιχνίδια).

Παρακάτω, παρουσιάζονται κάποια ιστογράμματα για τη διαφορά του σκορ στο τέλος των 4 περιόδων.

Σχήμα 2.7 *Ιστογράμματα για την dif_Pts*



Η πιο συχνή διαφορά πόντων στο τελικό σκορ είναι 0-10 πόντους και πραγματοποιήθηκε σε 60 παιχνίδια από τα 120 που οι γηπεδούχοι νίκησαν, δηλαδή ακριβώς στο 50% αυτών και στο 32.4% στο σύνολο των παιχνιδιών (185). Μεγάλη συχνότητα όμως έχουν και εκείνοι που ηττήθηκαν με μέχρι 10 πόντους, αφού αυτό συνέβη σε 40 παιχνίδια από τα 65 (61.5%).

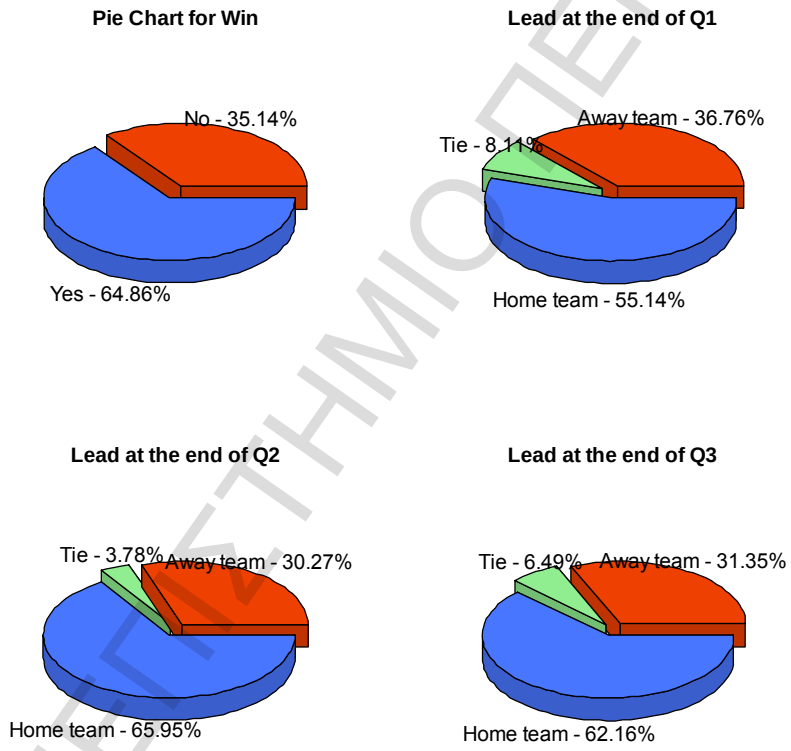
Στην πρώτη περίοδο την μεγαλύτερη συχνότητα έχει το διάστημα 0-5 πόντων υπέρ των γηπεδούχων, με μικρή όμως διαφορά από τους -5 πόντους εις βάρους τους. Στη δεύτερη περίοδο (ημίχρονο) πιο συχνή με σχεδόν 55 παιχνίδια είναι η διαφορά 0-5

πόντων, με πολύ μεγαλύτερη συχνότητα έναντι όλων των άλλων διαστημάτων. Στο τέλος της τρίτης περιόδου, η πιο συχνή διαφορά είναι 0-10 πόντους, ενώ και οι μεγαλύτερες διαφορές παρουσιάζουν αύξηση στη συχνότητα.

Η κατανομή της διαφοράς πόντων στις περιόδους φαίνεται να είναι κοντά στην κανονική, καθώς είναι συμμετρική, αλλά η περίοδος Q2 παρουσιάζει υψηλή κορυφή και έχει πιο λεπτόκυρτη μορφή από την Κανονική.

Παρουσιάζονται κάποιες πίτες, για να δούμε γραφικά σε τι ποσοστό οι γηπεδούχοι βρίσκονται μπροστά ή πίσω στο σκορ ανά περίοδο.

Σχήμα 2.8 *Πίτες για το τελικό αποτέλεσμα*



Η γηπεδούχος ομάδα κέρδισε σχεδόν στο 65% του συνόλου των παιχνιδιών. Επίσης, είχε το πάνω χέρι στο σκορ στο τέλος της $1^{ης}$ περιόδου στο 55% των αγώνων, στο τέλος της $2^{ης}$ στο 66% και στο τέλος της $3^{ης}$ στο 62% των αγώνων. Ισοπαλίες στο τέλος των τριών πρώτων περιόδων, σημειώθηκαν στο 8.1%, 3.78% και 6.49% αντίστοιχα των παιχνιδιών. Παρατηρείται ότι το ημίχρονο παρουσιάζει το μικρότερο ποσοστό στις ισοπαλίες και το μικρότερο ποσοστό που οι γηπεδούχοι ήταν πίσω στο σκορ.

Παράταση

Αξίζει να αναφέρουμε ότι στα 8 από τα 185 παιχνίδια, δηλαδή στο 4.3% αυτών, υπήρξε παράταση στην κανονική διάρκεια του αγώνα και συγκεκριμένα στις περιπτώσεις 10, 36, 96, 97, 110, 141, 151 και 156 των δεδομένων μας. Στα παιχνίδια αυτά η γηπεδούχος ομάδα κέρδισε στο 75% αυτών (6 παιχνίδια). Στα παιχνίδια της παράτασης επετεύχθησαν 85-95 πόντοι, υπήρξαν πολλές προσπάθειες για σουτ και τα ποσοστά ευστοχίας διπόντων και τριπόντων ήταν χαμηλά.

Βάσει των γκρουπ δυναμικότητας, οι γηπεδούχοι σε 3 παιχνίδια άνηκαν στο $2^{ο}$ γκρουπ, σε 3 παιχνίδια στο $6^{ο}$, και σε 2 στο $5^{ο}$ γκρουπ δυναμικότητας. Επίσης σε 3 παιχνίδια οι αντίπαλες ομάδες είχαν 1 γκρουπ διαφορά, σε 3 παιχνίδια 4 γκρουπ και σε 2 παιχνίδια διαφορά 2 γκρουπ δυναμικότητας.

2.4 Πίνακες συνάφειας με βάση το τελικό αποτέλεσμα

Με σκοπό να μελετήσουμε τις ανατροπές του σκορ αλλά και άλλων στοιχείων, θα κατασκευάσουμε πίνακες συνάφειας των σημαντικότερων μεγεθών ενός αγώνα μπάσκετ σε σύγκριση με τον τελικό νικητή. Στις γραμμές αναγράφονται οι απόλυτες και σχετικές (σε παρένθεση) συχνότητες βάσει κάποιας συνθήκης. Στις στήλες "Total" αναγράφεται το ποσοστό επί του συνόλου των παιχνιδιών, ενώ τα υπόλοιπα ποσοστά αφορούν μόνο την εκάστοτε συνθήκη.

Στους παρακάτω πίνακες αναφέρονται πόσες ομάδες νίκησαν ή ηττήθηκαν, ενώ ήταν μπροστά ή πίσω στο σκορ ανά περίοδο.

Πίνακας 2.4 *Πίνακες συχνοτήτων τελικού αποτελέσματος βάσει των περιόδων*

	Win		
Outcome of Q1	No	Yes	Total
Home < Away	42 (61.8%)	26 (38.2%)	68 (36.8%)
Home = Away	5 (33.3%)	10 (66.7%)	15 (8.1%)
Home > Away	18 (17.6%)	84 (82.4%)	102 (55.1%)
	65	120	185

	Win		
Outcome of Q2	No	Yes	Total
Home < Away	39 (69.6%)	17 (30.4%)	56 (30.3%)
Home = Away	3 (42.9%)	4 (57.1%)	7 (3.8%)
Home > Away	23 (18.9%)	99 (81.1%)	122 (65.9%)
	65	120	185

	Win		
Outcome of Q3	No	Yes	Total
Home < Away	43 (74.1%)	15 (25.9%)	58 (31.3%)
Home = Away	2 (16.7%)	10 (83.3%)	12 (6.5%)
Home > Away	20 (17.4%)	95 (82.6%)	115 (62.2%)
	65	120	185

	Win		
Conditions	No	Yes	Total
Q1, Q2 > 0	13 (14.8%)	75 (85.2%)	88 (47.6%)
Q2, Q3 > 0	14 (14.1%)	85 (85.9%)	99 (53.5%)
Q1, Q2, Q3 > 0	9 (12.2%)	65 (87.8%)	74 (40%)
Q1, Q2 < 0	34 (77.3%)	10 (22.7%)	44 (23.8%)
Q2, Q3 < 0	35 (87.5%)	5 (12.5%)	40 (21.6%)
Q1, Q2, Q3 < 0	31 (91.2%)	3 (8.8%)	34 (18.4%)

Οι παραπάνω πίνακες μας δίνουν τις εξής πληροφορίες:

Στην $1^{η}$ περίοδο η γηπεδούχος ομάδα ήταν πίσω στο σκορ σε 68 από τα 185 παιχνίδια, δηλαδή σε ποσοστό 36.8%, ισοπαλία σε 8.1% και προηγούταν σε 102 παιχνίδια, με ποσοστό 55.1%. Επίσης οι γηπεδούχοι που προηγούνταν στο σκορ (σε 84 από 102 παιχνίδια) κέρδισαν στο 82.4% αυτών. Αντίθετα όσοι έχαναν κέρδισαν μόλις στο 38.2% αυτών.

Μετά το πέρας της $2^{ης}$ περιόδου (ημίχρονο) το 30.3% των γηπεδούχων ήταν πίσω στο σκορ, το 3.8% ισοπαλία και το 65.9% ήταν μπροστά στο σκορ. Το 81.1% των γηπεδούχων που προηγούνταν νίκησαν και μόνο το 18.9% ηττήθηκαν.

Μετά το πέρας και της $3^{ης}$ περιόδου το 31.3% των γηπεδούχων ήταν πίσω στο σκορ, το 6.5% ισοπαλία και το 62.2% ήταν μπροστά στο σκορ. Το 82.6% των γηπεδούχων που προηγούνταν νίκησαν και μόνο το 17.4% ηττήθηκαν. Να σημειωθεί ότι στην περίπτωση ισοπαλίας οι γηπεδούχοι αναδείχθηκαν νικητές σε ποσοστό 83.3%.

Στη διεθνή βιβλιογραφία (Cooper et al, 1992) σε σχετική έρευνα αναφέρεται ότι η ομάδα που προηγούνταν στο σκορ στο ημίχρονο κέρδισε κατά προσέγγιση το 75% των παιχνιδιών, ενώ όταν αυτό συνέβαινε μετά το πέρας της 3^{ης} περιόδου το ποσοστό ήταν περίπου 80%. Το δείγμα μας μπορεί να επιβεβαιώσει απόλυτα τον παραπάνω ισχυρισμό, καθώς τα εμπειρικά ποσοστά ήταν 77.5% (=[39+99]/[56+122]) και 79.8% (=[43+95]/[58+115]) αντίστοιχα. Μετά το πέρας της 1^{ης} περιόδου το ποσοστό ήταν 74.1%.

Ο τελευταίος πίνακας μας πληροφορεί τι συνέβη όταν μια ομάδα προηγούνταν σε 2 ή 3 συνεχόμενες περιόδους. Η γηπεδούχος (φιλοξενούμενη) ομάδα που προηγούνταν και στις 3 πρώτες περιόδους κέρδισε το 87.8% (91.2%) των παιχνιδιών, ενώ το γεγονός αυτό συνέβη στο 40% (18.4%) του συνόλου των παιχνιδιών.

Παρακάτω παρουσιάζονται οι συχνότητες, όταν η διαφορά των πόντων σε κάθε περίοδο ήταν τουλάχιστον 5 υπέρ της μιας ομάδας.

Πίνακας 2.5 *Συχνότητες τελικού αποτελέσματος για ελάχιστη διαφορά 5 πόντων*

	Win		
Conditions	No	Yes	Total
dif_Q1 ≥ 5	6 (10.5%)	51 (89.5%)	57
dif_Q2 ≥ 5	9 (11.8%)	67 (88.2%)	76
dif_Q3 ≥ 5	7 (8.2%)	78 (91.8%)	85
dif_Q1 ≤ -5	26 (72.2%)	10 (27.8%)	36
dif_Q2 ≤ -5	29 (82.9%)	6 (17.1%)	35
dif_Q3 ≤ -5	33 (89.2%)	4 (10.8%)	37

Στο ημίχρονο οι γηπεδούχοι που προηγούνταν από 5 πόντους και άνω κέρδισαν τους αντιπάλους στο 88.2% των παιχνιδιών. Ενώ όταν αυτό συνέβαινε στην 1^{η} ή στην 3^{η} περίοδο το ποσοστό ήταν 89.5% και 91.8% αντίστοιχα.

Αντίθετα όταν οι γηπεδούχοι έχαναν με 5 και άνω πόντους στο ημίχρονο, κέρδισαν μόλις το 17.1% των παιχνιδιών, ενώ όταν συνέβαινε αυτό στην 1^{η} και 3^{η} περίοδο το ποσοστό ήταν 27.8% και 10.8% αντίστοιχα.

Η 3^{η} και προτελευταία περίοδος, όπως ήταν αναμενόμενο φαίνεται να παίζει σημαντικό ρόλο στην τελική έκβαση του αγώνα, καθώς τα ποσοστά είναι αυξημένα υπέρ ή κατά για κάθε ομάδα.

Το τι συνέβη στα γκρουπ δυναμικότητας μπορούμε να το δούμε στους δύο επόμενους πίνακες.

Πίνακας 2.6 *Συχνότητες σύμφωνα με τα γκρουπ δυναμικότητας των ομάδων*

	Win		
Group	No	Yes	Total
Home < Away	41 (46.1%)	48 (53.9%)	89 (48.1%)
Home = Away	2 (50%)	2 (50%)	4 (2.2%)
Home > Away	22 (23.9%)	70 (76.1%)	92 (49.7%)
	65	120	185

Όταν οι γηπεδούχοι ανήκαν σε υψηλότερο γκρουπ δυναμικότητας, κέρδισαν το 76.1% των αγώνων. Όσοι γηπεδούχοι ήταν σε μικρότερο γκρουπ, κέρδισαν το 53.9% των αγώνων. Όσοι ανήκαν στο ίδιο γκρουπ είχαν μοιρασμένα ποσοστά 50%.

Πίνακας 2.7 *Ποσοστά ανά γκρουπ δυναμικότητας*

		Win		
		No	Yes	Total
Group	1	10 (28.6%)	25 (71.4%)	35 (18.9%)
	2	6 (14.6%)	35 (85.4%)	41 (22.2%)
	3	15 (44.1%)	19 (55.9%)	34 (18.4%)
	4	12 (41.4%)	17 (58.6%)	29 (15.7%)
	5	14 (60.9%)	9 (39.1%)	23 (12.4%)
	6	8 (34.8%)	15 (65.2%)	23 (12.4%)
	Total	65	120	185

Στα περισσότερα παιχνίδια οι γηπεδούχοι άνηκαν στο 2^{ο} γκρουπ με ποσοστό 22.2%, μάλιστα οι δύο ομάδες από αυτές έφτασαν μέχρι τον τελικό. Από αυτό το γκρουπ παρατηρείται και το μεγαλύτερο ποσοστό νικών, σε 85.4% παιχνίδια από αυτά που έδωσαν, με το 1^{ο} γκρουπ να ακολουθεί με 71.4%. Το μικρότερο ποσοστό συμμετοχής στη διοργάνωση είχαν οι ομάδες από το 5^{ο} και 6^{ο} γκρουπ δυναμικότητας (12.4% στο καθένα), όπου βρίσκονταν και οι θεωρητικά αδύναμες, καθώς αποκλείστηκαν πρόωρα από την συνέχεια. Το 5^{ο} γκρουπ είχε και το μικρότερο ποσοστό νικών εντός έδρας με μόνο 39.1%.

Το τελικό αποτέλεσμα των ομάδων, σύμφωνα με κρίσιμους παράγοντες όπως τα λάθη και η ευστοχία, παρουσιάζεται στους δύο παρακάτω πίνακες.

Πίνακας 2.8 *Ποσοστά για ασίστ, ριμπάουντ, κλεψίματα και λάθη*

	Win		
Assists	No	Yes	Total
Home < Away	40 (62.5%)	24 (37.5%)	64 (34.6%)
Home = Away	4 (33.3%)	8 (66.7%)	12 (6.5%)
Home > Away	21 (19.3%)	88 (80.7%)	109 (58.9%)
Rebounds			
Home < Away	36 (46.8%)	41 (53.2%)	77 (41.6%)
Home = Away	8 (57.1%)	6 (42.9%)	14 (7.6%)
Home > Away	21 (22.3%)	73 (77.7%)	94 (50.8%)
Steals			
Home < Away	40 (54.1%)	34 (45.9%)	74 (40%)
Home = Away	8 (50%)	8 (50%)	16 (8.6%)
Home > Away	17 (17.9%)	78 (82.1%)	95 (51.4%)
Turnovers			
Home < Away	24 (23.1%)	80 (76.9%)	104 (56.2%)
Home = Away	3 (23.1%)	10 (76.9%)	13 (7%)
Home > Away	38 (55.9%)	30 (44.1%)	68 (36.8%)

Οι γηπεδούχοι που είχαν περισσότερες ασίστ από τους φιλοξενούμενους, κέρδισαν το 80.7% των παιχνιδιών, περισσότερα ριμπάουντ το 77.7%, περισσότερα κλεψίματα το 82.1% και λιγότερα λάθη το 76.9%. Ενδιαφέρον παρουσιάζει ότι, όταν είχαν και οι δυο ομάδες ίσα κλεψίματα το ποσοστό για νίκη ήταν στο 50-50, ενώ όταν είχαν ίσα ριμπάουντ οι γηπεδούχοι ηττήθηκαν σε 2 παιχνίδια παραπάνω από τους φιλοξενούμενους.

Πίνακας 2.9 *Ποσοστά για ευστοχία διπόντων, τριπόντων και ελευθέρων βολών*

Accuracy rates	Win		
2-point Field Goals	No	Yes	Total
Home < Away	44 (54.3%)	37 (45.7%)	81 (43.8%)
Home = Away	0 (0%)	3 (100%)	3 (1.6%)
Home > Away	21 (20.8%)	80 (79.2%)	101 (54.6%)
3-point Field Goals			
Home < Away	39 (52%)	36 (48%)	75 (40.5%)
Home = Away	0 (0%)	2 (100%)	2 (1.1%)
Home > Away	26 (24.1%)	82 (75.9%)	108 (58.4%)
Free Throws			
Home < Away	31 (36.5%)	54 (63.5%)	85 (45.9%)
Home = Away	1 (50%)	1 (50%)	2 (1.1%)
Home > Away	33 (33.7%)	65 (66.3%)	98 (53%)

Οι γηπεδούχοι ήταν πιο εύστοχοι από τους αντιπάλους τους (κυρίως στα τρίποντα) και συνήθως όταν συνέβαινε αυτό, κέρδισαν τα παιχνίδια με 79.2%, 75.9% και 66.3% για τα δίποντα, τα τρίποντα και τα σουτ ελ. βολών αντίστοιχα. Οι φιλοξενούμενες ομάδες που ήταν πιο εύστοχες στα σουτ δύο και τριών πόντων, κέρδισαν σε λίγο πάνω από τα μισά παιχνίδια με ποσοστά 54.3% και 52% αντίστοιχα, ενώ όταν τα ποσοστά ήταν τα ίδια κέρδισαν οι γηπεδούχοι. Όσον αφορά τις ελ. βολές το προβάδισμα για την νίκη το είχαν οι γηπεδούχοι με ποσοστό πάνω από 60% είτε ήταν πιο εύστοχοι, είτε όχι και μόνο στην περίπτωση που ήταν ίδιο το ποσοστό ευστοχίας, οι πιθανότητες ήταν μοιρασμένες.

Στους επόμενους και τελευταίους πίνακες, θα μελετηθεί η ειδική αξιολόγηση Rkg.

Πίνακας 2.10 *Πίνακας συχνοτήτων για την ειδική αξιολόγηση της Euroleague*

	Win	
Ranking	No	Yes	Total
Home < Away	57 (90.5%)	6 (9.5%)	63 (34%)
Home = Away	0 (0%)	2 (1.1%)	2 (1.1%)
Home > Away	8 (6.7%)	112 (93.3%)	120 (64.9%)
	65	120	185

Η ομάδα (γηπεδούχος ή μη) με μικρότερη βαθμολογία από την αντίπαλη, νίκησε μόλις σε 14 παιχνίδια (6 ως γηπεδούχος και 8 ως φιλοξενούμενη) από τα 185, δηλαδή σε ποσοστό 7.6%. Έτσι φαίνεται η ειδική αξιολόγηση της διοργάνωσης να προσδιορίζει ικανοποιητικά το τελικό αποτέλεσμα του αγώνα.

Πίνακας 2.11 *Περιπτώσεις λάθος κατηγοριοποίησης της ειδικής αξιολόγησης*

Misclassified cases														
	Lower Rkg - Win						*Higher Rkg - Defeat*							
Case	49	180	56	61	104	106	39	173	47	174	71	82	153	129
dif_Rkg	-1	-3	-6	-6	-10	-20	1	1	1	2	3	7	11	15
dif_Pts	3	9	4	4	1	1	-2	-2	-11	-4	-4	-4	-1	-3

Εντοπίζονται δύο περιπτώσεις όπου η γηπεδούχος ομάδα αν και είχε 10 ή 20 μονάδες μικρότερη αξιολόγηση από τον αντίπαλο, κέρδισε με 1 πόντο διαφορά, ενώ όταν η διαφορά στην αξιολόγηση ήταν μικρότερη κέρδισε με διαφορά 3-9 πόντων. Αντίθετα όταν είχε επιπλέον μονάδες αξιολόγησης 11 ή 15, ηττήθηκε με 1 και 3 πόντους διαφορά αντίστοιχα, ενώ υπάρχει και μια περίπτωση όπου ηττήθηκε με 11 πόντους διαφορά αν και είχε έναν πόντο αξιολόγησης μεγαλύτερο. Επίσης στα δύο παιχνίδια όπου dif_Rkg=0, η γηπεδούχος ομάδα κέρδισε με 1 και 2 πόντους διαφορά αντίστοιχα.

Τέλος θα προσπαθήσουμε να εξηγήσουμε τι συνέβη και παιχνίδια με dif_Rkg = -20 και 15, δεν είχαν το αναμενόμενο αποτέλεσμα. Στο 106^{ο} παιχνίδι παρότι η αντίπαλη ομάδα (Παναθηναϊκός) ήταν πιο εύστοχη στα σουτ 2 και 3 πόντων και είχε 9 παραπάνω ασίστ, σε όλα τα υπόλοιπα στοιχεία (με εξαίρεση τα ριμπάουντ που ήταν ίδια) κλεψίματα, λάθη, κοψίματα, φάουλ ήταν ελαφρώς χειρότερη από την γηπεδούχο. Στο 129^{ο} παιχνίδι η γηπεδούχος ήταν καλύτερη σε ευστοχία 2 πόντων και είχε 17 ριμπάουντ και 7 ασίστ παραπάνω από την φιλοξενούμενη, εντούτοις όμως υπήρχε σημαντική υστέρηση σε ευστοχία ελ. βολών (-16.8%), λάθη (+6), κλεψίματα (-5).

Σύμφωνα με τους παραπάνω πίνακες, είναι φανερό ότι ανατροπές του σκορ είναι δύσκολο να επιτευχθούν. Επίσης, οι σημαντικότεροι παράγοντες όπου φαίνεται οι τιμές τους να δίνουν ενδείξεις για το τελικό αποτέλεσμα του αγώνα είναι οι: πόντοι, ευστοχία διπόντων – τριπόντων, ασίστ, λάθη, ριμπάουντ, κλεψίματα και γκρουπ δυναμικότητας.

ΚΕΦΑΛΑΙΟ 3

Έλεγχοι καλής προσαρμογής του σκορ

Στο παρόν κεφάλαιο, θα μελετηθούν με στατιστικές μεθόδους οι κατανομές που περιγράφουν καλύτερα το σχηματισμό του σκορ ενός αγώνα.

3.1 Αναφορά μεθόδων καλής προσαρμογής

Σε αυτό το κεφάλαιο θα εξετάσουμε, ποια ή ποιες από τις περισσότερο γνωστές κατανομές προσεγγίζουν καλύτερα την κατανομή των συνολικών πόντων της γηπεδούχου ομάδας, καθώς και τις διαφορές πόντων του τελικού σκορ. Αυτή η πληροφορία είναι ιδιαίτερα χρήσιμη σε ερευνητές που κατασκευάζουν μοντέλα πρόβλεψης και υπολογίζουν πιθανότητες.

Έτσι θα κάνουμε κάποιους ελέγχους καλής προσαρμογής (goodness of fit tests) για την μεταβλητή Pts, καθώς και για τις διαφορές των τελικών πόντων dif_Pts, αλλά και τις απόλυτες τιμές αυτών. Επίσης, θα πραγματοποιήσουμε έλεγχο Κανονικότητας του σκορ ανά περίοδο.

Για τους ελέγχους εφαρμόσαμε τα κυριότερα στατιστικά τεστ, αυτά των Kolmogorov-Smirnov, Anderson-Darling, Cramer-Von Mises, Chi-Squared (χ^2) και Log Likelihood (λογαρίθμου πιθανοφάνειας). Και ύστερα ελέγχεται γραφικά η καταλληλότητα της επιλεχθείσας κατανομής. Όπως φαντάζει λογικό, θα ερευνήσουμε μόνο συνεχείς κατανομές, αφού το πλήθος των δυνατών τιμών του σκορ ενός αγώνα είναι αρκετά μεγάλο και θεωρητικά μπορεί να πάρει οποιαδήποτε τιμή.

Στον χ^2 έλεγχο τα δεδομένα ομαδοποιήθηκαν σε κλάσεις, εκείνες που «ταιριάζουν» με το ιστόγραμμα συχνοτήτων. Σε αυτό τον έλεγχο αναφέρεται το p-value και όχι το στατιστικό χ^2 (λόγω του ότι διαφέρουν οι β.ε. ανά κατανομή), εν αντιθέσει με τα 3 πρώτα τεστ που χρησιμοποιούν τις στατιστικές συναρτήσεις D, A^2 και W^2 αντίστοιχα. Τα κριτήρια θα αναφέρονται ως εξής:

KS: στατιστική συνάρτηση των Kolmogorov-Smirnov

CVM: στατιστική συνάρτηση των Cramer-Von Mises

AD: στατιστική συνάρτηση των Anderson-Darling

CHISQ: p-value του χ^2 ελέγχου

LLH: λογάριθμος της πιθανοφάνειας

Ισχύει ότι:

- Οι μικρότερες τιμές των στατιστικών συναρτήσεων KS, CVM και AD εκφράζουν καλύτερη προσαρμογή.
- Στον χ^2 έλεγχο η μεγαλύτερη τιμή του p-value είναι και η καλύτερη.
- Η μεγαλύτερη τιμή του λογαρίθμου πιθανοφάνειας είναι η καλύτερη.

Επιπλέον, θα αναφερθούν δύο γνωστά μέτρα (τυποποιημένοι συντελεστές) για την περιγραφή της μορφής μιας κατανομής πιθανότητας.

❖ Συντελεστής ασυμμετρίας / λοξότητας (Skewness)

Αν η κατανομή είναι συμμετρική (όπως η κανονική) τότε ο συντελεστής είναι 0. Αν είναι θετικός (αρνητικός) σημαίνει ότι η κατανομή είναι λοξή προς τα δεξιά (αριστερά).

❖ Συντελεστής κύρτωσης (Kurtosis)

Τιμή ίση με 0 αντιστοιχεί στην τυπική κανονική κατανομή. Θετική τιμή του συντελεστή δείχνει μια πιο «αιχμηρή» κορυφή και πιο μακριές, παχιές ουρές, ενώ μια αρνητική τιμή δείχνει μια πιο «στρογγυλεμένη» κορυφή και μικρότερες, λεπτότερες ουρές.

Οι παραπάνω συντελεστές, οι στατιστικές συναρτήσεις των κριτηρίων, όπως και οι συναρτήσεις πυκνότητας πιθανότητας των (όχι τόσο συνηθισμένων) προσαρμοσμένων κατανομών, παρατίθενται στο Παράρτημα. Επίσης, όλες οι κατανομές που προσαρμόζονται στις επόμενες ενότητες είναι στατιστικά σημαντικές, σε όλους τους παραπάνω ελέγχους με p-values > 0.05 (εδώ η μηδενική υπόθεση είναι ότι η μεταβλητή ακολουθεί την εξεταζόμενη κατανομή). Τέλος, η εκτίμηση των παραμέτρων των κατανομών έγινε μέσω της MLE μεθόδου (μεγιστοποίησης της πιθανοφάνειας).

3.2 Προσαρμογή κατανομής στους πόντους

Στον παρακάτω πίνακα αναφέρονται οι τυποποιημένοι συντελεστές κύρτωσης και ασυμμετρίας των πόντων.

Πίνακας 3.1 *Τυποποιημένοι συντελεστές κύρτωσης και ασυμμετρίας της Pts*

	Pts
Stnd. Skewness	0.20210
Stnd. Kurtosis	-0.20453

Οι τιμές είναι κοντά στο 0 και έτσι η κατάλληλη κατανομή θα μοιάζει στην Κανονική.

Ύστερα από δοκιμές προσαρμογής με τις πιο γνωστές συνεχείς κατανομές, επιλέχθηκαν οι παρακάτω με τις αντίστοιχες παραμέτρους.

Πίνακας 3.2 *Προσαρμοσμένες κατανομές για την Pts*

Fitted Distributions for Pts				
Normal	*Log-normal*	*Gamma*	*Logistic*	*Inverse Gaussian*
mean	log(mean)	shape	location	log(mean)
76.89189	4.33327	54.91347	76.67887	4.34240
sd	log(sd)	scale	scale	log(lambda)
10.35579	0.13562	0.71416	5.93519	8.32942

Οι συγκρίσεις στον παρακάτω πίνακα μπορούν να γίνουν μόνο «κάθετα», για κάθε έλεγχο ξεχωριστά και όχι «οριζόντια». Οι «καλύτερες» τιμές (όπως αυτές ορίστηκαν στην προηγούμενη ενότητα) τονίζονται.

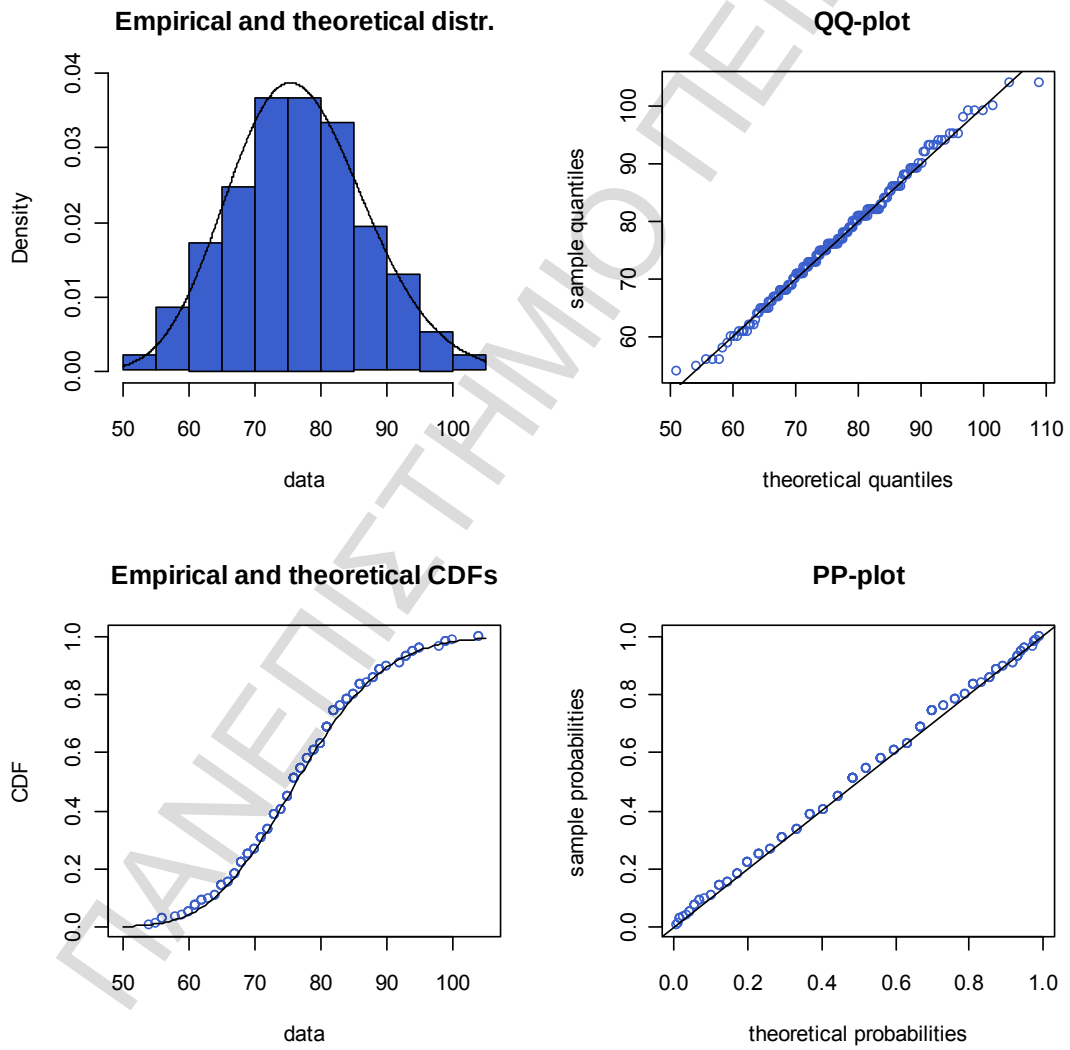
Πίνακας 3.3 *Έλεγχοι καλής προσαρμογής για την Pts*

Goodness of Fit Tests for Pts					
	KS	*CVM*	*AD*	*CHISQ*	*LLH*
Normal	0.05686	0.05043	0.34265	0.47725	-694.951
Log-normal	0.04828	0.04748	0.30429	**0.89130**	-694.544
Gamma	0.04522	**0.03644**	**0.23967**	0.86566	**-694.166**
Logistic	**0.04208**	0.04473	0.36639	0.42890	-697.235
Inverse Gaussian	0.04863	0.04816	0.30701	0.88074	-694.514

Εφόσον συμφωνούν τα 3 από τα παραπάνω κριτήρια, θα επιλέξουμε την κατανομή Γάμμα ως αυτή με την καλύτερη προσαρμογή για τους πόντους. Υπενθυμίζουμε ότι στον πίνακα μόνο στον χ^2 έλεγχο αναφέρονται τιμές p-value και η τιμή του στη Γάμμα κατανομή είναι πολύ υψηλή (0.86566), ενώ αν υπολογιστεί και το αντίστοιχο για το KS τεστ, αυτό είναι 0.84447 που επιβεβαιώνει την καλή προσαρμογή.

Θα επιβεβαιώσουμε την επιλογή μας με τα παρακάτω γραφήματα, στα οποία εξετάζουμε την προσαρμογή της κατανομής Γάμμα στα δεδομένα για τους πόντους της γηπεδούχου ομάδας.

Σχήμα 3.1 *Γραφήματα προσαρμογής Γάμμα κατανομής για την Pts*



➢ Η εκτιμώμενη καμπύλη της συνάρτηση πυκνότητας πιθανότητας προσαρμόζεται πολύ καλά στο ιστόγραμμα σχετικών συχνοτήτων. Η καμπύλη της Γάμμα κατανομής με αυτές τις παραμέτρους, έχει τη μορφή συμμετρικής κατανομής όπως η Κανονική.

➢ Από το Q-Q plot παρατηρούμε ότι μόνο η μικρότερη και μεγαλύτερη τιμή (ακραίες τιμές) αποκλίνουν σημαντικά από την ευθεία.

➢ Η εμπειρική συνάρτηση κατανομής των δεδομένων προσεγγίζει αρκετά ικανοποιητικά τη θεωρητική συνάρτηση κατανομής της Γάμμα.

➢ Τέλος στο P-P διάγραμμα οι δειγματικές με τις θεωρητικές πιθανότητες είναι πολύ κοντά, αφού όλα τα σημεία είναι πάνω ή πολύ κοντά σε μια ευθεία γραμμή.

Έτσι επιβεβαιώνεται και γραφικά η εξαιρετική προσαρμογή της Γάμμα κατανομής.

3.3 Προσαρμογή κατανομής για την διαφορά των πόντων

Παρακάτω αναφέρονται οι συντελεστές κύρτωσης και ασυμμετρίας των πόντων.

Πίνακας 3.4 *Τυποποιημένοι συντελεστές κύρτωσης και ασυμμετρίας της dif_Pts*

dif_Pts	
Stnd. Skewness	0.54527
Stnd. Kurtosis	0.47529

Αυτές οι θετικές τιμές υποδεικνύουν ότι η κατανομή αυτής της μεταβλητής θα ακολουθεί μια κατανομή, ελαφρώς λοξή προς τα δεξιά και με λίγο πιο αιχμηρή κορυφή σε σχέση με την κανονική.

Στη συνέχεια θα ερευνήσουμε την κατανομή που ακολουθεί η μεταβλητή με την τελική διαφορά του σκορ, η οποία παίρνει και αρνητικές τιμές. Οι εκτιμήσεις των παραμέτρων αναφέρονται στον επόμενο πίνακα.

Πίνακας 3.5 *Προσαρμοσμένες κατανομές για την dif_Pts*

Fitted Distributions for dif_Pts					
Normal	*Log-normal (3-parameter)*	*Logistic*	*Generalized Logistic*	*Loglogistic (3-parameter)*	*Gamma (3-parameter)*
mean	mean	location	shape	shape	shape
5.39460	5.38030	4.83940	2.97290	4.29008	11.05710
sd	sd	scale	scale	scale	scale
13.74796	13.76420	7.72017	9.95518	9.51855	0.24017
	location		location	location	location
	-66.52190		-9.31416	-68.71180	-40.66180

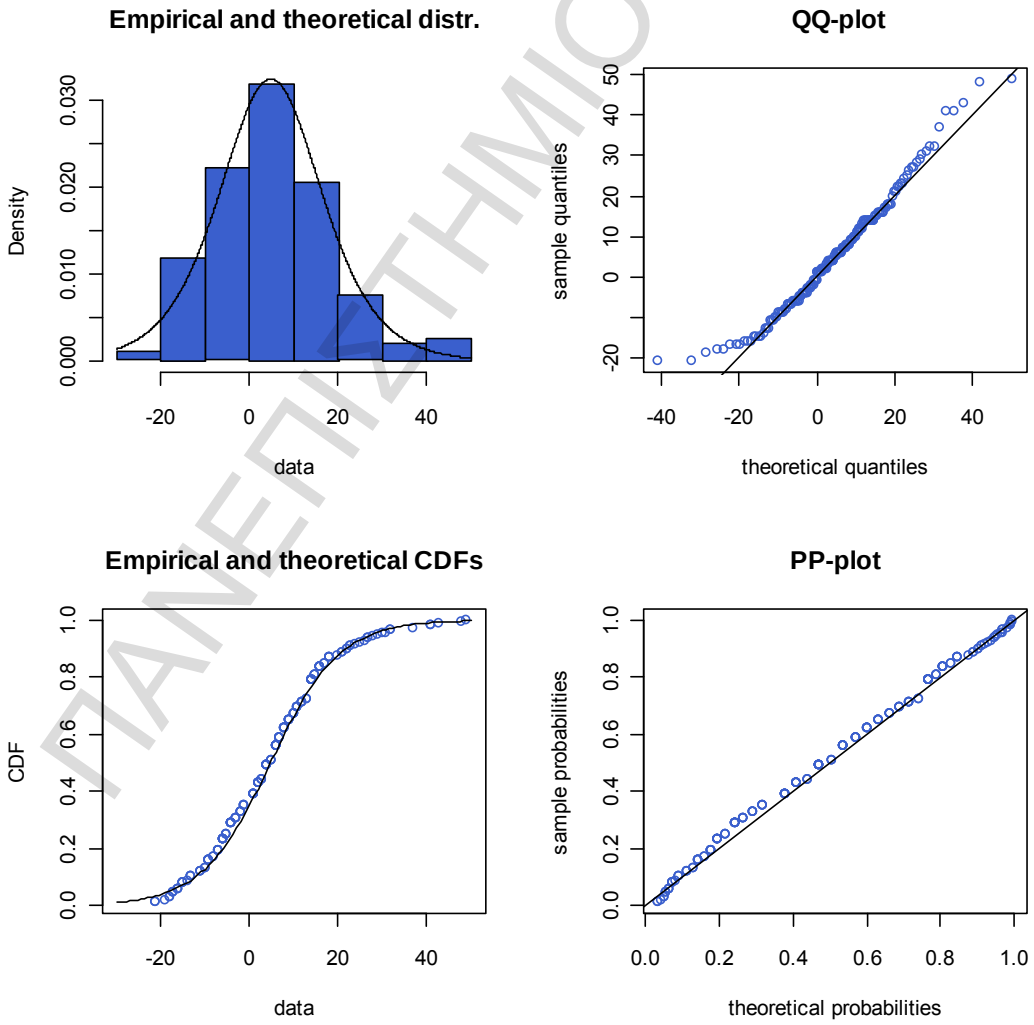
Στη συνέχεια, παρατίθενται οι τιμές από τους ελέγχους καλής προσαρμογής.

Πίνακας 3.6 *Έλεγχοι καλής προσαρμογής για την dif_Pts*

Goodness of Fit Tests for dif_Pts					
	KS	*CVM*	*AD*	*CHISQ*	*LLH*
Normal	0.05807	0.08958	0.74304	0.30924	-747.370
Lognormal (3-par)	0.05518	0.06974	**0.43551**	0.43556	-742.766
Logistic	**0.04508**	**0.04862**	0.49716	0.36497	-747.014
Generalized Logistic	0.05682	0.08075	0.49400	0.67559	-744.030
Loglogistic (3-par)	0.04826	0.07148	0.46652	**0.85563**	-744.714
Gamma (3-par)	0.05947	0.07944	0.47827	0.34802	**-742.609**

Θα επιλέξουμε τη Λογιστική κατανομή, εφόσον δύο από τους παραπάνω ελέγχους «συμφωνούν» και επιπλέον απαιτείται η εκτίμηση μόνο δύο παραμέτρων. Το p-value του χ^2 είναι το 4ο καλύτερο με 0.36. Το p-value της τιμής 0.04508 του KS τεστ για αυτήν την κατανομή είναι 0.84766, που υποδηλώνει πολύ καλή προσαρμογή. Σειρά έχουν οι απαραίτητοι γραφικοί έλεγχοι.

Σχήμα 3.2 *Γραφήματα προσαρμογής Λογιστικής κατανομής για την dif_Pts*



- Η καμπύλη φαίνεται να προσαρμόζεται καλά στο ιστόγραμμα (με το μεγαλύτερο πλήθος των τιμών της μεταβλητής), με εξαίρεση τις άκρες αυτής.
- Από το Q-Q διάγραμμα παρατηρούμε ότι υπάρχουν ακραίες τιμές, δηλαδή μεγάλες διαφορές στους πόντους στις οποίες η Λογιστική κατανομή δεν μπορεί να προσαρμοστεί καλά.
- Η συνάρτηση κατανομής φαίνεται να προσαρμόζεται αρκετά καλά στις τιμές της μεταβλητής.
- Στο P-P διάγραμμα λίγα σημεία αποκλίνουν από την ευθεία.

Η προσαρμογή κρίνεται αρκετά ικανοποιητική. Επίσης αξίζει να αναφέρουμε ότι, οι κατανομές με τις 3 παραμέτρους «χειρίστηκαν» καλύτερα τις ακραίες τιμές. Βέβαια μεγάλες διαφορές πόντων (κυρίως 25 και άνω) μπορεί να θεωρηθεί ένα μη συχνό φαινόμενο.

3.4 Προσαρμογή κατανομής στην απόλυτη διαφορά των πόντων

Η μορφή της κατανομής περιγράφεται στον επόμενο πίνακα.

Πίνακας 3.7 *Τυποποιημένοι συντελεστές κύρτωσης και ασυμμετρίας της |dif_Pts|*

|*dif_Pts*|
Stnd. Skewness	1.58919
Stnd. Kurtosis	3.14552

Οι παραπάνω σημαντικά μεγαλύτερες του μηδενός τιμές, περιγράφουν μια κατανομή με μακριά δεξιά ουρά και απότομη κορυφή.

Ύστερα στους δύο επόμενους πίνακες, θα δοθούν οι εκτιμήσεις των παραμέτρων των επιλεχθεισών κατανομών και θα βρούμε την κατανομή που προσαρμόζεται καλύτερα, η μεταβλητή που εκφράζει την απόλυτη διαφορά στο τελικό σκορ.

Πίνακας 3.8 *Προσαρμοσμένες κατανομές για την |dif_Pts|*

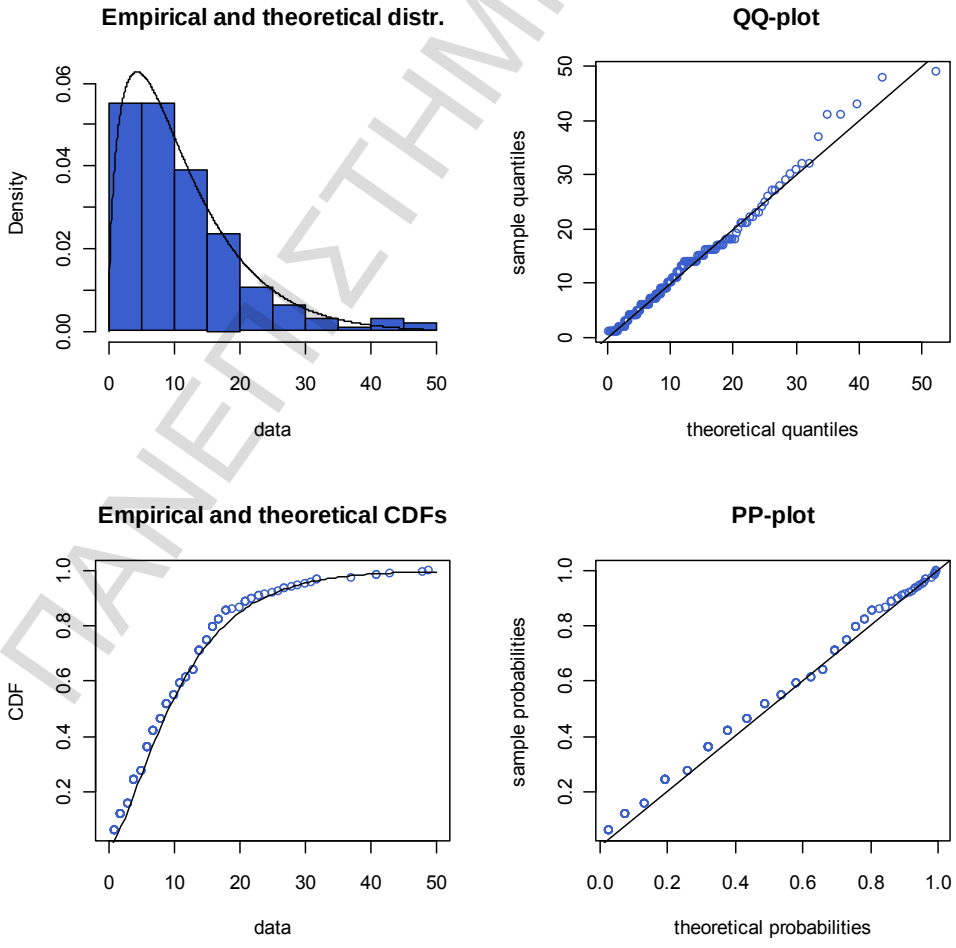
Fitted Distributions for |dif_Pts|	
Gamma	*Weibull*
shape = 1.60408	shape = 1.30482
scale = 0.13917	scale = 12.52666

Πίνακας 3.9 *Έλεγχοι καλής προσαρμογής της |dif_Pts|*

Goodness of Fit Tests for |dif_Pts|					
	KS	*CVM*	*AD*	*CHISQ*	*LLH*
Gamma	0.05935	**0.08209**	**0.61029**	**0.09485**	**-626.347**
Weibull	**0.05498**	0.08489	0.65319	0.05024	-627.451

Βάσει των παραπάνω ελέγχων, θα πρέπει να επιλέξουμε τη Γάμμα κατανομή. Το p-value για το χ^2 αν και δείχνει σημαντικότητα, είναι λίγο πάνω από το ε.σ. 5% , ενώ το αντίστοιχο για το KS τεστ υπολογίζεται σε 0.54672 (για την Weibull είναι 0.63192). Ακολουθούν τα διαγράμματα για την προσαρμογή της κατανομής.

Σχήμα 3.3 *Γραφήματα προσαρμογής Γάμμα κατανομής για την |dif_Pts|*



➢ Παρατηρούμε ότι στο ιστόγραμμα η εκτιμώμενη καμπύλη έχει κάπως υψηλή κορυφή και στη δεξιά άκρη της δεν προσαρμόζονται καλά οι τιμές.
➢ Στο Q-Q διάγραμμα οι πολύ υψηλές τιμές αποκλίνουν σημαντικά από την ευθεία.
➢ Η εμπειρική συνάρτηση κατανομής είναι πολύ κοντά με την θεωρητική.
➢ Στο P-P διάγραμμα αν και πολλά σημεία φαίνεται να μην είναι πάνω στην ευθεία, εντούτοις όμως είναι κοντά σε αυτήν.

Η κατανομή στο σύνολό της κρίνεται ικανοποιητική, τόσο από τους γραφικούς ελέγχους, όσο και από τους στατιστικούς. Όπως προαναφέρθηκε το κύριο πρόβλημα είναι η συχνότητα των πολύ μεγάλων διαφορών στο σκορ. Από τα 185 σε 9 παιχνίδια παρατηρήθηκε διαφορά άνω των 30 πόντων.

3.5 Έλεγχος για την «κανονικότητα» της διαφοράς πόντων των περιόδων

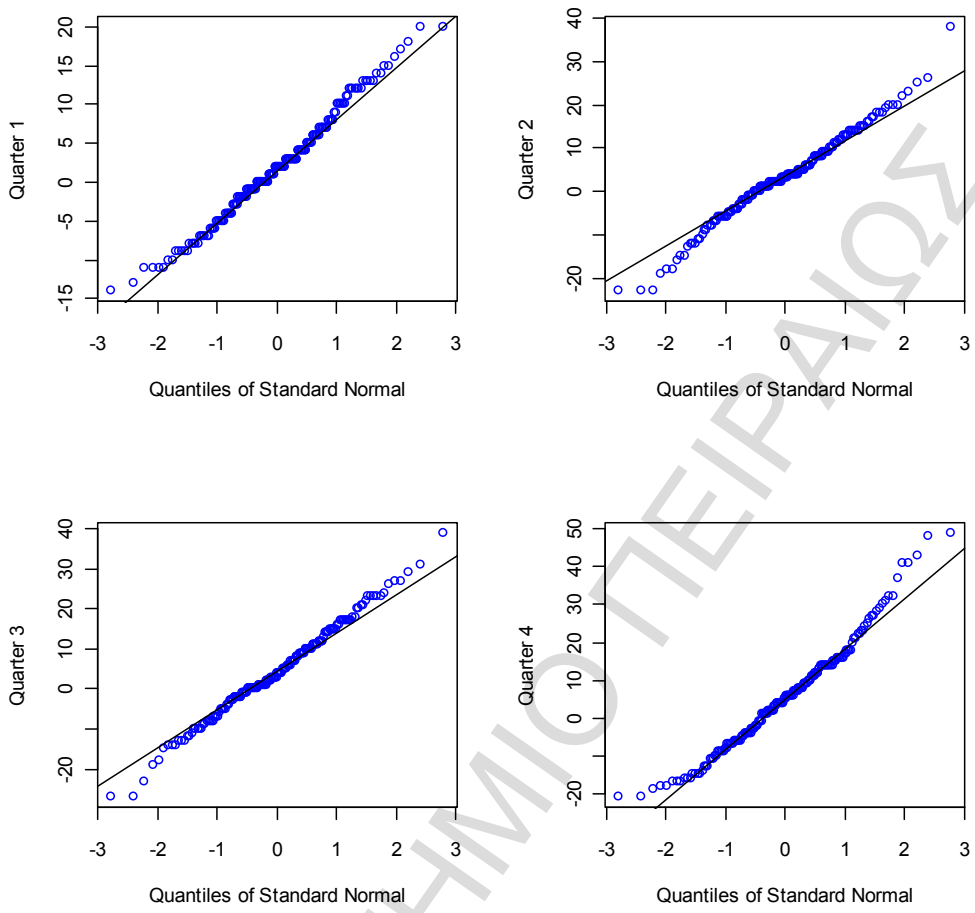
Παρότι είδαμε ότι, για την τελική διαφορά των πόντων υπάρχει καταλληλότερη κατανομή από την Κανονική που να την ερμηνεύει, εντούτοις η Κανονική κατανομή αποτελεί μια καλή προσέγγιση. Αναμένουμε το ίδιο να συμβεί και για τη διαφορά του σκορ στο τέλος κάθε περιόδου. Έτσι θα κάνουμε σύντομους στατιστικούς ελέγχους για να το επαληθεύσουμε. Παρακάτω αναφέρονται οι παράμετροι της κατανομής.

Πίνακας 3.10 *Προσαρμογή Κανονικής κατανομής στις περιόδους*

Normal Distribution				
	Q1	Q2	Q3	Q4
mean	1.77838	3.17297	4.25405	5.39460
sd	6.92739	9.61600	11.13287	13.74800

Τα Q-Q διαγράμματα των περιόδων είναι τα παρακάτω.

Σχήμα 3.4 *Κανονικά Q-Q διαγράμματα για κάθε περίοδο*



Πέρα από κάποιες πολύ μεγάλες / μικρές τιμές, τα σημεία δε φαίνεται να αποκλίνουν σημαντικά από την ευθεία γραμμή.

Στατιστικά θα επιβεβαιώσουμε την Κανονικότητα των περιόδων με τους παρακάτω ελέγχους χ^2 και Kolmogorov-Smirnov.

Πίνακας 3.11 *Έλεγχοι Κανονικότητας των περιόδων*

Normality tests				
	Q1	Q2	Q3	Q4
Chi-Squared	0.18053	0.08209	0.32334	0.30924
KS	0.36186	0.32186	0.74815	0.54705

Πράγματι, από τα p-values δε μπορούμε να απορρίψουμε σε επίπεδο σημαντικότητας 5% ότι, η διαφορά πόντων στο τέλος κάθε περιόδου ακολουθεί την Κανονική κατανομή.

ΚΕΦΑΛΑΙΟ 4

Θεωρία και μέθοδοι αξιολόγησης μοντέλων Logit & Probit

Στο παρόν κεφάλαιο, αποδίδεται η θεωρία που διέπει τα μοντέλα λογιστικής και probit παλινδόμησης και αναφέρονται διαγνωστικοί έλεγχοι για την αξιολόγηση και σύγκριση των μοντέλων αυτών.

4.1 Εισαγωγή στη Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση (Logistic Regression) ανήκει στην οικογένεια των Γενικευμένων Γραμμικών Μοντέλων (GLM) και χρησιμοποιείται αντί της κλασικής παλινδρόμησης στις περιπτώσεις όπου η μεταβλητή απόκρισης (response variable) είναι κατηγορική και θέλουμε να μελετήσουμε τη σχέση της με ένα σετ συνεχών και κατηγορικών μεταβλητών. Στην πράξη είναι ιδιαίτερα σύνηθες να κάνουμε προβλέψεις για κάποιο γεγονός που παίρνει κυρίως δύο τιμές, όπως επιτυχία / αποτυχία ενός αθλητικού αποτελέσματος, βελτίωση / μη βελτίωση της κατάστασης ενός ασθενή ή ύπαρξη / μη ύπαρξη ενός κοινωνικο-οικονομικού φαινομένου.

Τα GLM επιτρέπουν μεταβλητές απόκρισης που έχουν διαφορετική κατανομή από την Κανονική, να συνδέονται μέσω μιας συνάρτησης σύνδεσης (link function) με ένα γραμμικό μοντέλο. Ένα GLM έχει τη μορφή $g(\mu) = \beta X$, όπου μ είναι η μέση τιμή της κατανομής της μεταβλητής απόκρισης Y, X είναι οι ανεξάρτητες μεταβλητές, β είναι οι συντελεστές παλινδρόμησης, και g είναι η συνάρτηση σύνδεσης του μ. Όταν η Y είναι δίτιμη ακολουθεί την κατανομή Bernoulli και τη θέση του μ παίρνει το p, $g(p) = \beta X$, όπου p είναι η πιθανότητα να συμβεί το γεγονός $\{P(Y = 1)\}$.

4.2 Το λογιστικό μοντέλο

Στη λογιστική παλινδρόμηση η logit είναι η συνάρτηση σύνδεσης. Η εξίσωση του μοντέλου με μορφή πινάκων γράφεται:

$$g(\underset{\sim}{p}) = \ln\left(\frac{p}{1-p}\right) = \underset{\sim}{\beta} X$$

ή αναλυτικότερα:

$$logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \mathrm{X}_{1i} + \ldots + \beta_p \mathrm{X}_{pi} = \beta_0 + \sum_{j=1}^{k} \beta_j \mathrm{X}_{ji}$$

Η ανάπτυξη αυτή του μοντέλου δικαιολογείται από δύο βασικούς λόγους:

α. Διασφαλίζουμε ότι οι πιθανότητες p_i παίρνουν τιμές ανάμεσα στο διάστημα 0 και 1, δηλαδή $0 \le p_i \le 1$ και $\ln\left(\frac{p_i}{1-p_i}\right) \in (-\infty, +\infty)$.

β. Ο δεύτερος είναι ότι με έναν απλό εκθετικό μετασχηματισμό μετατρέπουμε τα log-odds σε πιθανότητες. Έτσι με αντιλογαρίθμηση προκύπτουν οι παρακάτω σχέσεις:

$$\frac{p_i}{1-p_i} = e^{\beta_0 + \sum_{j=1}^{k} \beta_j \mathrm{X}_j} \Leftrightarrow p_i = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j \mathrm{X}_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j \mathrm{X}_j}} = \frac{1}{1 + e^{-\beta_0 - \sum_{j=1}^{p} \beta_j \mathrm{X}_j}},$$

η οποία είναι και η αθροιστική συνάρτηση κατανομής της Λογιστικής κατανομής.

Η συνάρτηση logit προτιμάται σε σχέση με άλλες συναρτήσεις σύνδεσης, λόγω της εύκολης διαισθητικής ερμηνείας των αποτελεσμάτων με βάση τη σχετική πιθανότητα (odds), όπου αυτή είναι ο λόγος $odds = \frac{p_i}{1-p_i}$ και συγκρίνει την πιθανότητα να συμβεί το ενδεχόμενο (p_i) με την αντίστοιχη να μη συμβεί ($1\text{-}p_i$). Επίσης ισχύει ότι $p_i = \frac{odds}{1+odds}$

Ο μετασχηματισμός που χρησιμοποιείται στη λογιστική παλινδρόμηση λέγεται logit μετασχηματισμός, όπου αντί της *Y*, χρησιμοποιείται ο λογάριθμος των πιθανοτήτων. Αυτό συμβαίνει διότι η *Y* παίρνει μόνο δύο τιμές και τα κατάλοιπα δε μπορεί να ακολουθούν την κανονική κατανομή (τα κατάλοιπα έχουν μόνο δύο πιθανές τιμές για κάθε *X*). Η καλύτερη γραμμή που περιγράφει τη σχέση ανάμεσα στα *X* και στο *Y* δεν είναι ευθεία, αλλά έχει το σχήμα ενός S.

Υποθέσεις της λογιστικής παλινδρόμησης:

- Οι πραγματικές δεσμευμένες πιθανότητες είναι μια λογιστική συνάρτηση των ερμηνευτικών μεταβλητών
- Καμία σημαντική μεταβλητή δεν παραλείπεται
- Καμιά εξωγενής μεταβλητή δεν περιλαμβάνεται
- Η μέτρηση των ερμηνευτικών μεταβλητών γίνεται χωρίς σφάλμα
- Οι παρατηρήσεις είναι ανεξάρτητες
- Οι ερμηνευτικές μεταβλητές δεν είναι γραμμικός συνδυασμός η μία της άλλης
- Καμία υπόθεση δεν γίνεται για την κατανομή των ερμηνευτικών μεταβλητών

4.3 Το μοντέλο Probit

Σε ένα probit μοντέλο, η συνάρτηση σύνδεσης είναι η αντίστροφη συνάρτηση της τυπικής Κανονικής κατανομής, η οποία μοντελοποιείται ως ένας γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών.

$$\text{probit}(p_i) = \Phi^{-1}(p_i) = \beta_0 + \beta_1 \mathrm{X}_1 + \ldots + \beta_p X_p$$

Οπότε $p_i = \Phi(\beta_0 + \beta_1 \mathrm{X}_{1\mathrm{i}} + \ldots + \beta_\mathrm{p} \mathrm{X}_{\mathrm{pi}})$

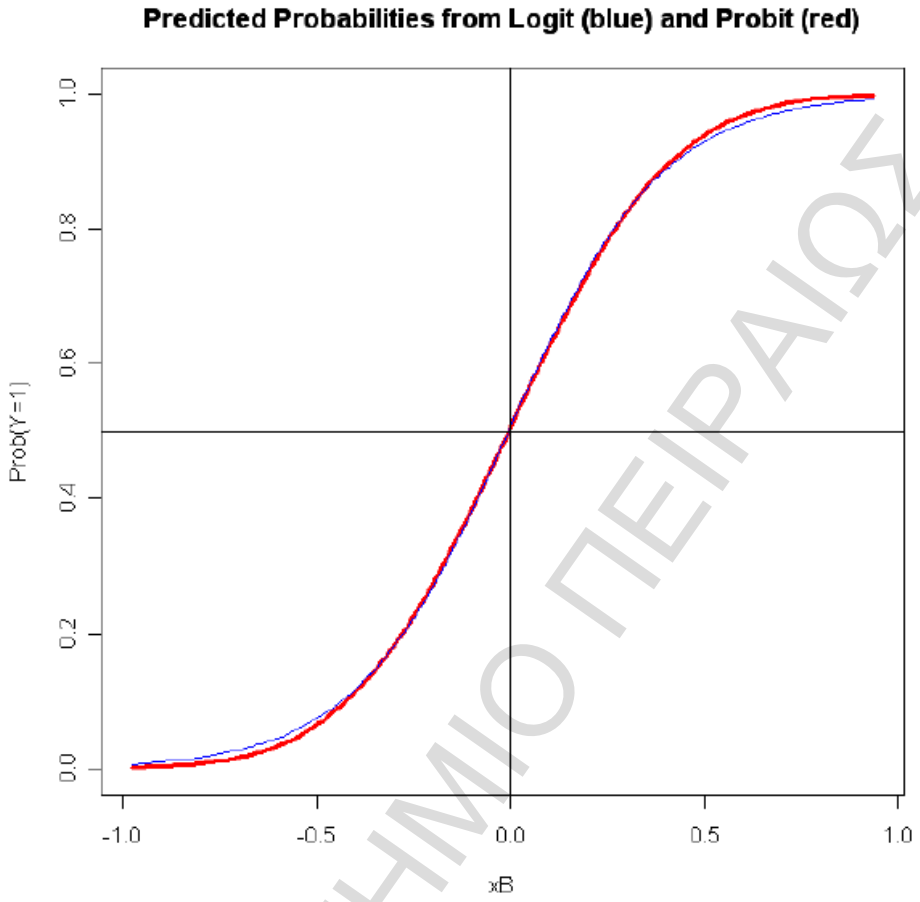
Όπου Φ συνάρτηση κατανομής της N(0,1) και Φ^{-1} η αντίστροφή της.

Οι συντελεστές της probit παλινδρόμησης έχουν διαφορετική ερμηνεία από την αντίστοιχη στη logit (όπως θα δούμε παρακάτω). Αύξηση μίας μονάδας στην ανεξάρτητη μεταβλητή, προκαλεί αύξηση του *Z-score* της *P(Y=1)* κατά την τιμή του συντελεστή.

4.4 Σύγκριση Logit και Probit

Οι συντελεστές των δύο μοντέλων δεν είναι άμεσα συγκρίσιμοι μεταξύ τους, διότι βρίσκονται σε διαφορετική κλίμακα. Επίσης διαφέρουν τα άκρα των καμπυλών τους και η Probit είναι πιο ευαίσθητη σε ακραίες τιμές. Παρόλα αυτά η σημαντικότητα των ερμηνευτικών μεταβλητών είναι ταυτόσημη και σε γενικές γραμμές τα μοντέλα logit και probit καταλήγουν σε πανομοιότυπα συμπεράσματα.

Σχήμα 4.1 *Συγκριτικό διάγραμμα Logit και Probit καμπυλών*



Η καμπύλη της probit προσεγγίζει τους άξονες λίγο πιο γρήγορα από τη logit καμπύλη (έχει πιο παχιές ουρές). Στο παραπάνω συγκριτικό διάγραμμα παρατηρούμε ότι οι πιθανότητες είναι συμμετρικές στο 0.5, κάτι αντίστοιχο που συμβαίνει στην πράξη.

4.5 Εκτίμηση των παραμέτρων

Για τα μη γραμμικά μοντέλα παλινδρόμησης, οι μέθοδοι OLS και WLS δεν είναι εφαρμόσιμες. Έτσι, αν και υπάρχει η επιλογή μεθόδου Μη-Γραμμικών Ελαχίστων Τετραγώνων (Non-Linear Least Squares), μία πολύ προτιμότερη είναι η Μέθοδος Μέγιστης Πιθανοφάνειας (MLE).

Συνάρτηση Πιθανοφάνειας

Αφού τα Y_i είναι τ.μ. Bernoulli, όπου $P(Y_i=1)=p_i$ και $P(Y_i=0)=1-p_i$,η συνάρτηση πυκνότητας πιθανότητας είναι:

$$f_i(Y_i) = p_i^{Y_i}(1-p_i)^{1-Y_i} \qquad Y_i=0,\ 1 \text{ και } i=1,\ldots,n$$

Οι παρατηρήσεις Y_i είναι ανεξάρτητες, οπότε η από κοινού συνάρτηση πιθανότητας θα είναι:

$$L = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}$$

Λογαριθμώντας καταλήγουμε στη σχέση:

$$\ln L = \ln \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i} = \sum_{i=1}^{n}\left[Y_i \ln\left(\frac{p_i}{1-p_i}\right)\right] + \sum_{i=1}^{n} \ln(1-p_i)$$

Αν αντικαταστήσουμε το $\ln\left(\frac{p_i}{1-p_i}\right)$ και το $1-p_i$, θα έχουμε τη λογαριθμική συνάρτηση πιθανοφάνειας των εκτιμώμενων παραμέτρων:

$$\ln L = \sum_{i=1}^{n} Y_i\left(\beta_0 + \sum_{j=1}^{k}\beta_j X_j\right) + \sum_{i=1}^{n} \ln\left(1 - \frac{e^{\beta_0+\sum_{j=1}^{k}\beta_j X_j}}{1+e^{\beta_0+\sum_{j=1}^{k}\beta_j X_j}}\right)$$

$$= \sum_{i=1}^{n} Y_i\left(\beta_0 + \sum_{j=1}^{k}\beta_j X_j\right) + \sum_{i=1}^{n} \ln\left(1+e^{\beta_0+\sum_{j=1}^{k}\beta_j X_j}\right)$$

Οι τιμές των β_0, β_j που μεγιστοποιούν την παραπάνω συνάρτηση καλούνται εκτιμητές μέγιστης πιθανοφάνειας (MLE). Εδώ να πούμε πως δεν μπορούμε να βρούμε τους εκτιμητές, όπως θα τους βρίσκαμε στα γραμμικά μοντέλα (να βρούμε τις τιμές των β_0, β_j που θα μεγιστοποιούσαν τη λογαριθμική συνάρτηση πιθανοφάνειας), γιατί πολύ απλά δεν υπάρχουν λύσεις κλειστής μορφής για τις τιμές των β_0, β_j που θα μεγιστοποιούσαν το lnL. Πρέπει να χρησιμοποιηθούν επαναληπτικές αριθμητικές μέθοδοι οι οποίες θα μας δίνουν τους εκτιμητές b_0 και b_j.

Η πιο διαδομένη μέθοδος για αυτήν την μεγιστοποίηση είναι η μέθοδος IRLS (Iteratively Reweighted Least Squares), όπου γίνονται κάποιες αρχικές εικασίες για

τις τιμές των αγνώστων παραμέτρων και με μια επαναληπτική διαδικασία προσαρμόζονται αυτές οι εκτιμήσεις, μέχρι να βρεθεί η μέγιστη τιμή του L. Οι βέλτιστες τιμές των β θα βρεθούν, είτε μέσω του αλγορίθμου Newton–Raphson, είτε μέσω μιας παραλλαγής αυτού, του σκορ του Fisher (Fisher scoring) που στη λογιστική παλινδρόμηση είναι ισοδύναμοι (Storvik, 2011). Μια άλλη διαφορετική μέθοδος για εκτίμηση των παραμέτρων είναι αυτή του Bayes.

4.6 Έλεγχοι των παραμέτρων του μοντέλου

Οι παρακάτω έλεγχοι για την σημαντικότητα μιας ερμηνευτικής μεταβλητής είναι ασυμπτωτικά ισοδύναμοι για μεγάλα δείγματα.

1. Ο έλεγχος του Wald (Wald test)

Για κάθε μία παράμετρο ξεχωριστά μπορεί να γίνει ο έλεγχος της υπόθεσης H_0: $\beta_j=0$ μέσω της ποσότητας $z=\beta_j/S_j$ (όπου S_j είναι η εκτίμηση του τυπικού σφάλματος για κάθε β_j) και σύγκριση των z^2 με το ποσοστημόριο της χ^2 κατανομής με 1 β.ε (π.χ. $x^2_{1,1-0.05} = 3.84$) .

2. Ο έλεγχος βάσει του σκορ (Score test)

Καλείται και Lagrange Multiplier test (LM) και χρησιμοποιεί την πρώτη παράγωγο της συνάρτησης πιθανοφάνειας για την υπόθεση $\beta_j=0$ (το Wald τεστ βασίζεται στη δεύτερη παράγωγο).

3. Ο έλεγχος λόγου πιθανοφανειών (Likelihood Ratio test)

Υπολογίζεται με την διαδικασία που αναφέρθηκε προηγουμένως, μόνο που τώρα συγκρίνεται ένα μοντέλο με και χωρίς μια ερμηνευτική μεταβλητή, για να εξεταστεί αν η διαφορά στην προσαρμογή από την προσθήκη αυτής της μεταβλητής είναι σημαντική.

Οι παραπάνω έλεγχοι ασυμπτωτικά είναι ισοδύναμοι, δηλαδή για μεγάλα δείγματα δίνουν το ίδιο αποτέλεσμα. Γενικότερα οι δύο τελευταίοι είναι πιο αξιόπιστοι, αλλά ο έλεγχος του Wald είναι αυτός που χρησιμοποιείται συχνότερα λόγω της ευκολίας στην εφαρμογή του.

4.7 Ερμηνεία των συντελεστών

Στη λογιστική παλινδρόμηση οι συντελεστές δίνουν την αλλαγή στις log odds του αποτελέσματος, για αύξηση μιας μονάδας της ανεξάρτητης μεταβλητής. Έτσι κάθε όρος στη σχέση, αντιπροσωπεύει τη συνεισφορά του στην εκτίμηση των log-odds. Για κάθε μοναδιαία αύξηση (μείωση) της X_j (όταν οι υπόλοιπες μεταβλητές είναι σταθερές), προβλέπεται να υπάρχει αύξηση (μείωση) κατά β_j μονάδες στο λογάριθμο των σχετικών πιθανοτήτων (του ενδεχομένου Y=1). Αν όλες οι παράμετροι τεθούν ίσες με 0, τότε τα προβλεπόμενα log-odds θα είναι ίσα με το σταθερό όρο β_0.

Για τον λόγο ότι δεν είναι τόσο φυσικό να σκεφτόμαστε σε όρους log-odds, η σχέση $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_\kappa X_\kappa$ μπορεί με αντιλογαρίθμηση να διαμορφωθεί σε:

$$\frac{\hat{p}}{1-\hat{p}} = e^{\beta_0+\beta_1 X_1+...+\beta_\kappa X_\kappa} = e^{\beta_0+\beta_1 X_1+...+\beta_k X_k} = e^{\beta_0} \times e^{\beta_1 X_1} \times ... \times e^{\beta_k X_k}$$

Τώρα οι συντελεστές βρίσκονται ως δύναμη του e. Σε όρους odds η επίδραση της κάθε ανεξάρτητης μεταβλητής είναι πολλαπλασιαστική. Έτσι κάθε μοναδιαία αύξηση της X_j, προκαλεί πολλαπλασιαστική αύξηση της σχετικής πιθανότητας επιτυχίας κατά e^{β_j} μονάδες, όταν οι υπόλοιπες είναι σταθερές. Αν η X_j μειωθεί κατά μία μονάδα, τότε ο πολλαπλασιαστικός παράγοντας είναι $e^{-\beta_j}$. Αν όλοι οι συντελεστές τεθούν ίσοι με 0, οι εκτιμώμενες σχετικές πιθανότητες είναι ίσες με e^{β_0}.

4.8 Κατάλοιπα (residuals)

Στη λογιστική παλινδρόμηση, οι τύποι των καταλοίπων που χρησιμοποιούνται στην πράξη για διαγνωστικούς σκοπούς είναι τα κατάλοιπα απόκλισης (Deviance residuals), τα κατάλοιπα του Pearson και τα τυποποιημένα κατάλοιπα (Standardized residuals). Αυτά είναι χρήσιμα για την ανίχνευση «ακραίων» τιμών και για να προσδιοριστεί η επιρροή των παρατηρήσεων στο μοντέλο. Διαγραμματικά, τα κατάλοιπα ταξινομημένα με τις προσαρμοσμένες τιμές του μοντέλου, βρίσκονται σε καμπύλες τόσες, όσες και τα επίπεδα του γεγονότος (δηλαδή δύο για δίτιμο γεγονός).

α. Deviance residuals

Ορίζεται ως η τετραγωνική ρίζα της συνεισφοράς της i παρατήρησης στην απόκλιση, με το πρόσημο (sign) του καταλοίπου απόκρισης $(y_i - \hat{p}_i)$.

$$r_{di} = \sqrt{d_i}\,(sign(y_i - \hat{p}_i))$$

$$d_i = -2 \cdot \left[y_i \log\left(\frac{\hat{p}_i}{y_i}\right) + (1 - y_i) \log\left(\frac{1-\hat{p}_i}{1-y_i}\right) \right]$$

$\hat{p}_i$: εκτιμώμενη πιθανότητα επιτυχίας

y_i: παρατηρούμενη τιμή

β. Pearson residuals

Ορίζονται ως η διαφορά ανάμεσα στα παρατηρούμενα και αναμενόμενα αποτελέσματα για μια παρατήρηση, διαιρεμένη με την τετραγωνική ρίζα της διακύμανσης του αναμενόμενου αποτελέσματος. Μετράνε τις σχετικές αποκλίσεις ανάμεσα στις παρατηρούμενες και τις προσαρμοσμένες τιμές.

$$e_i = \frac{(y_i - \hat{p}_i)}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$$

γ. Standardized Pearson residuals

Σε γενικές γραμμές, το τυποποιημένο κατάλοιπο είναι το κατάλοιπο διαιρεμένο με μια εκτίμηση της τυπικής απόκλισής του. Στη λογιστική παλινδρόμηση ορίζεται ως:

$$\frac{e_i}{\sqrt{1-h_i}}$$

h_i: μόχλευση (hat value) της i παρατήρησης

Αναλόγως υπολογίζεται και το τυποποιημένο κατάλοιπο της απόκλισης (*Standardized Deviance residual*), ενώ υπάρχει και μια παραλλαγή τυποποίησης το *Studentized residual.* Το studentized κατάλοιπο εκφράζει την αλλαγή της απόκλισης στο μοντέλο, εάν η περίπτωση εξαιρεθεί.

4.9 Επιλογή μεταβλητών

α. Για να αποφύγουμε υπερπροσαρμογή (overfitting) του μοντέλου, ο αριθμός των ερμηνευτικών μεταβλητών που θα εισάγουμε στο μοντέλο παίζει σπουδαίο ρόλο. Αν η μια τιμή της μεταβλητής απόκρισης (π.χ. *Y*=1) υπερτερεί ή υστερεί σημαντικά στη συχνότητα εμφάνισής της, σε σχέση με την άλλη τιμή της (π.χ. *Y*=0), τότε είναι λάθος να χρησιμοποιήσουμε πολλές επεξηγηματικές μεταβλητές, διότι οι εκτιμητές των συντελεστών μπορεί να είναι αρκετά μεροληπτικοί και η εκτίμηση των τυπικών σφαλμάτων τους κακή.

Ένας πρακτικός κανόνας (Peduzzi et al, 1996) είναι να χρησιμοποιούμε μία επεξηγηματική μεταβλητή για τουλάχιστον 10 αποκρίσεις από κάθε κατηγορία. Για παράδειγμα αν έχουμε n = 200 παρατηρήσεις και από αυτές οι 50 είναι Y = 0, τότε δεν είναι καλό να χρησιμοποιήσουμε πάνω από 5 επεξηγηματικές μεταβλητές. Ο κανόνας είναι:

$$N = 10k / p$$

Όπου

N: μέγεθος δείγματος

k: αριθμός ανεξάρτητων μεταβλητών

p: το μικρότερο ποσοστό θετικών ή αρνητικών περιπτώσεων

Αν υπάρχουν και όροι αλληλεπίδρασης, ο αριθμός αυτός αυξάνεται. Ενώ σε άλλες έρευνες (Harrell, 2001) υποστηρίζεται ότι είναι ιδανικότερος ένας αριθμός 15-20 παρατηρήσεις για κάθε παράμετρο του μοντέλου (η σταθερά λογίζεται ως μία παράμετρος).

β. Οι μεταβλητές που θα επιλέξουμε δεν θα πρέπει να παραβιάζουν τις υποθέσεις του λογιστικού μοντέλου, όπως ότι πρέπει να είναι ανεξάρτητες και καμιά από αυτές δεν θα πρέπει να έχει ισχυρή συσχέτιση με κάποια άλλη.

γ. Γενικότερα, θα πρέπει να συγκρίνουμε πολλά μοντέλα, ώστε να δούμε ποιες μεταβλητές σχετίζονται περισσότερο με την απόκριση. Κάποιες μέθοδοι για την καλύτερη επιλογή των ανεξάρτητων μεταβλητών είναι οι παρακάτω:

- ***Forward*** *(Προς τα εμπρός επιλογή)*

Η διαδικασία ξεκινά με την «καλύτερη» μεταβλητή, στη συνέχεια προσθέτει την καλύτερη από τις υπόλοιπες, μέχρις ότου προσθέτοντας μια νέα μεταβλητή η αύξηση της λογαριθμικής συνάρτησης πιθανοφάνειας να μην είναι στατιστικά σημαντική.

- ***Backward*** *(Προς τα πίσω εξάλειψη)*

Ξεκινά με όλο το σύνολο των μεταβλητών (πλήρες μοντέλο) και απορρίπτει διαδοχικά τη «χειρότερη» από τις εναπομείνασες.

- ***Stepwise*** *(Σταδιακή επιλογή)*

Σε κάθε βήμα ελέγχονται ποιες μεταβλητές πρέπει να προστεθούν ή να αφαιρεθούν με βάση κάποιο κριτήριο (συνήθως p-value του συντελεστή ή ελέγχου πιθανοφάνειας), ώστε στο νέο σύνολο όλες οι μεταβλητές να είναι στατιστικώς σημαντικές. Η διαδικασία θεωρείται η καλύτερη, διότι κάνει διπλούς ελέγχους.

4.10 Διαγνωστικοί έλεγχοι

4.10.1 Έλεγχος πολυσυγγραμμικότητας

Το φαινόμενο της πολυσυγγραμικότητας εμφανίζεται με την ύπαρξη ισχυρών γραμμικών σχέσεων, μεταξύ των επεξηγηματικών μεταβλητών του μοντέλου. Η πολυσυγγραμικότητα δεν επηρεάζει την πρόβλεψη της μεταβλητής απόκρισης, προκαλεί όμως σύγχυση στην εκτίμηση των συντελεστών του μοντέλου (προκύπτουν μεγάλα τυπικά σφάλματα), δηλαδή δε μπορούν να καθοριστούν οι επιδράσεις των επεξηγηματικών μεταβλητών στη μεταβλητή απόκρισης.

Οι πιο διαδεδομένοι (Belsley et al, 1980) έλεγχοι για το φαινόμενο της πολυσυγγραμμικότητας (multicollinearity) είναι μέσω των συντελεστών διόγκωσης της διακύμανσης VIF (Variance Inflation Factors) και του ελέγχου με ιδιοτιμές και ιδιοδιανύσματα (Condition Number Test).

α. VIF

Ένα VIF είναι πάντα >= 1 και υπολογίζει πόσες φορές η διακύμανση του αντίστοιχου συντελεστή παλινδρόμησης, αυξάνεται λόγω της πολυσυγγραμμικότητας. Το $\sqrt{\text{VIF}}$ δείχνει πόσο μεγαλύτερο το τυπικό σφάλμα του συντελεστή είναι, συγκρινόμενο με το τι θα ήταν, αν η μεταβλητή ήταν ασυσχέτιστη με τις υπόλοιπες ανεξάρτητες μεταβλητές του μοντέλου.

$$VIF = \frac{1}{tolerance} = \frac{1}{1 - R_i^2}$$

R_i^2 : συντελεστής προσδιορισμού που προκύπτει από την παλινδρόμηση των άλλων μεταβλητών πάνω στην i μεταβλητή

Tolerance (Ανοχή): δείκτης που δείχνει το ποσοστό της διακύμανσης που δεν εξηγείται από τις υπόλοιπες συμμεταβλητές.

Αν μια μεταβλητή συσχετίζεται έντονα με άλλη μεταβλητή, η ανοχή τείνει στο 0 και το VIF γίνεται πολύ μεγάλο.

Ένας γενικότερος κανόνας είναι ότι σε μικρά δείγματα ένα VIF ≥ 5 παρέχει ισχυρή ένδειξη εμφάνισης του φαινομένου, ενώ για μεγαλύτερα δείγματα το κριτήριο εφαρμόζεται σε μεγαλύτερες τιμές του VIF. Ορισμένοι (Belsley et al 1980 & Kutner, 2004), θέτουν πιο ελαστικές τιμές του κανόνα με VIF ≥ 10.

β. Condition Number Test

Σύμφωνα με ορισμένους συγγραφείς, είναι το καλύτερο διαγνωστικό για ανίχνευση του φαινομένου (Belsley et al, 1980 & Gujarati, 1988). Υπολογίζεται ένας «δείκτης κατάστασης» (Condition Index) για κάθε ιδιοτιμή (eigenvalue) και αν ο μεγαλύτερος εξ αυτών είναι ≥ 30, τότε υπάρχουν πολύ ισχυρές ενδείξεις για εμφάνιση πολυσυγγραμμικότητας. Τιμές > 10 & < 30 παρέχουν ένδειξη πολυσ/τητας, ενώ αν ο δείκτης είναι ≤ 10 τότε δεν υπάρχει τέτοιο πρόβλημα.

$$\text{Condition Index} = \sqrt{\max\left(eigenvalues\right) / eigenvalue}$$

Το τεστ παρέχει επιπλέον πληροφορίες που μπορούν να βοηθήσουν στην εξεύρεση της πηγής του προβλήματος, τις «αναλογίες αποσύνθεσης της διακύμανσης» (Variance Decomposition Proportions) που συνδέονται με κάθε condition index. Ορίζονται ως αναλογία (ποσοστό) του VIF, που προκύπτει από τη γραμμική σχέση που απεικονίζει η αντίστοιχη ιδιοτιμή (και ιδιοδιάνυσμα). Αν ένας μεγάλος condition

index συνδέεται με 2 ή περισσότερες μεταβλητές με μεγάλα variance decomposition proportions (≥ 0.5), τότε αυτές οι μεταβλητές μπορεί να προκαλούν προβλήματα συγγραμμικότητας.

4.10.2 Έλεγχοι ανίχνευσης «ακραίων» τιμών

Αυτές οι τιμές όταν υπάρχουν δημιουργούν προβλήματα, καθώς επηρεάζουν την εκτίμηση των συντελεστών του μοντέλου και συνεπώς την προβλεπτική του ικανότητα.

α. Outliers

Οι παρατηρήσεις των οποίων οι τιμές διαφέρουν σημαντικά από τις άλλες παρατηρήσεις του δείγματος καλούνται έκτροπες (outliers). Ή αλλιώς ορίζονται ως οι παρατηρήσεις, που εμφανίζουν μεγάλα κατάλοιπα σε ένα μοντέλο.

Ένας εμπειρικός κανόνας, για να θεωρήσουμε παρατηρήσεις ως έκτροπες είναι ότι όταν τα τυποποιημένα κατάλοιπα (standardized residuals) είναι πάνω από 3 ή κάτω από -3. Για μικρά δείγματα (κυρίως με n<80), πέρα από τα όρια ±2 ή ±2.5 είναι πιθανά outliers (Schwab, 2003). Οι έκτροπες παρατηρήσεις διακρίνονται γραφικά, με ένα διάγραμμα διασποράς των standardized residuals.

Να αναφέρουμε ότι υπάρχουν περιπτώσεις όπου, τα outliers να μην είναι τυχαίες ή λανθασμένες τιμές, αλλά ενδεχομένως να περιέχουν σημαντικές πληροφορίες για την σχέση μεταξύ των μεταβλητών και πιθανόν θα μπορούσαν να εξηγηθούν, με την προσθήκη επιπλέον ερμηνευτικών μεταβλητών.

β. Influential observations

Είναι παρατηρήσεις με μεγάλη επίδραση στα αποτελέσματα του μοντέλου. Η απόσταση του Cook (Cook's distance) είναι ένα ευρέως χρησιμοποιούμενο μέτρο της επίδρασης (μεταβολής), που προκαλεί στους συντελεστές του μοντέλου η αφαίρεση της i παρατήρησης. Μετρά την επίδραση μιας παρατήρησης στη συνολική ικανότητα του μοντέλου να προβλέπει όλες τις περιπτώσεις.

$$D_i = s_i^2 \frac{h_{ii}}{p(1-h_{ii})},$$

Όπου:

h_{ii}: μόχλευση

s_i: τυποποιημένο κατάλοιπο

p: αριθμός παραμέτρων του μοντέλου

Όσον αφορά τη μόχλευση (hat value), ισχύει ότι $0 \leq h_{ii} \leq 1$ και το άθροισμα αυτών είναι ίσο με $h_{ii}=m+1$ (όπου m ο αριθμός των ανεξάρτητων μεταβλητών). Ένας γενικός κανόνας για να χαρακτηριστεί μια παρατήρηση ως «influential» είναι $h_{ii} > 2\frac{m+1}{n}$ (Habing, 2004).

Υπάρχουν διάφορες απόψεις, σχετικά με το ποια τιμή cutoff είναι κατάλληλη για τον εντοπισμό των παρατηρήσεων υψηλής επιρροής. Ένας πρακτικός κανόνας για τον εντοπισμό είναι $D_i > 1$ (Cook & Weisberg, 1982), ενώ άλλοι ανέφεραν ότι το $D_i > 4/n$ (n: αριθμός των παρατηρήσεων) θα μπορούσε να χρησιμοποιηθεί (Bollen & Jackman, 1990). Άλλος ένας πρακτικός κανόνας είναι ότι για σετ δεδομένων με n>15 παρατηρήσεις, $D_i>0.7$ για μοντέλο με p=2, δηλ. με 1 ανεξάρτητη μεταβλητή, $D_i>0.8$ για p=3, δηλ. μοντέλο με 2 ανεξάρτητες μεταβλητές και $D_i>0.85$ για p>3, δηλ. με περισσότερες μεταβλητές από 2 (McDonald, 2002).

Ακόμα όμως και αν δεν υπάρχουν παρατηρήσεις που να υπακούουν τους παραπάνω κανόνες, αν υπάρχουν κάποιες που έχουν πολύ μεγαλύτερες τιμές D_i από όλες τις υπόλοιπες θα μπορούσαν να χαρακτηριστούν ως «influential» (Chatterjee et al, 2000). Μια διαδομένη μέθοδος για την ανίχνευση ασυνήθιστων ως προς την επιρροή τους παρατηρήσεων, είναι η κατασκευή ενός διαγράμματος διασποράς των αποστάσεων του Cook, έναντι του αριθμού των περιπτώσεων.

Άλλα γνωστά μέτρα, για την ανίχνευση παρατηρήσεων υψηλής επιρροής είναι τα Mahalanobis distance, dfbetas και dffits.

Τρόπος χειρισμού των ακραίων παρατηρήσεων

Για να υπολογιστεί η επίπτωση που έχουν οι παραπάνω παρατηρήσεις, ακολουθείται η εξής στρατηγική:

Αρχικά, προσαρμόζουμε ένα μοντέλο (αναφοράς) με όλες τις περιπτώσεις. Ύστερα, προσαρμόζουμε ένα (αναθεωρημένο) μοντέλο έχοντας εξαιρέσει τις «ύποπτες» παρατηρήσεις και ελέγχουμε αν το μοντέλο έχει βελτιωθεί. Επίσης εξετάζουμε τους συντελεστές των μεταβλητών, καθώς και τις συσχετίσεις αυτών.

Ένας πρακτικός κανόνας είναι, ότι εάν το αναθεωρημένο μοντέλο έχει σωστή κατηγοριοποίηση μικρότερη από 2% παραπάνω ακρίβεια από το μοντέλο αναφοράς, τότε θα μένουμε στο μοντέλο αναφοράς (Schwab, 2003).

4.11 Έλεγχοι καλής προσαρμογής (Goodness of fit tests)

Οι έλεγχοι της απόκλισης και του Pearson είναι οι πιο γνωστοί για την αξιολόγηση της προσαρμογής ενός λογιστικού μοντέλου. Οι υπόλοιποι έλεγχοι είναι συμπληρωματικοί και ορισμένοι δίνουν ιδιαίτερα χρήσιμες πληροφορίες για την προβλεπτική ικανότητα του μοντέλου. Επίσης, όλα τα κριτήρια είναι συγκρίσιμα μεταξύ διαφορετικών μοντέλων, όπου προβλέπουν το ίδιο αποτέλεσμα στο ίδιο σετ δεδομένων.

α. Η απόκλιση (deviance)

Η απόκλιση ενός γενικευμένου γραμμικού μοντέλου M είναι η ποσότητα:

$$Deviance_M = -2\log\frac{L_M}{L_S} = -2(\log\ L_M - \log\ L_S) = -2\log L_M$$

και είναι ένας θετικός αριθμός. Όπου:

L_M: η μέγιστη πιθανοφάνεια υπό το μοντέλο M.

L_S: η μέγιστη πιθανοφάνεια υπό το κορεσμένο μοντέλο (δηλαδή εκείνο που έχει μία παράμετρο για κάθε μία παρατήρηση και ταιριάζει τέλεια στα δεδομένα), η οποία ισούται με 1 και επομένως $\log L_S=0$.

Στην ανάλυση δίτιμων μεταβλητών, η απόκλιση έχει ασυμπτωτικά κατανομή χι-τετράγωνο με βαθμούς ελευθερίας df = n – k (n: πλήθος παρατηρήσεων και k: πλήθος παραμέτρων μοντέλου M).

Η καλή προσαρμογή του μοντέλου εκφράζεται με μικρές τιμές της απόκλισης, ότι δηλαδή το μοντέλο είναι κοντά στο κορεσμένο που έχει τέλεια προσαρμογή. Στη σύγκριση μοντέλων όμως, θα πρέπει να λάβουμε υπόψιν και τους βαθμούς ελευθερίας κάθε μοντέλου, έτσι χρησιμοποιείται το Deviance/df, μεγάλες τιμές του οποίου υποδεικνύουν κακή προσαρμογή.

Likelihood Ratio Test (LRT)

Για δύο μοντέλα M_0, M_1 με το δεύτερο να είναι ειδική περίπτωση του πρώτου (περιέχει ένα υποσύνολο των επεξηγηματικών μεταβλητών του M_0), ο έλεγχος του λόγου πιθανοφανειών για την υπόθεση H_0: ισχύει το M_0, κατά της H_1: ισχύει το M_1 είναι ουσιαστικά η διαφορά των αποκλίσεων:

$$G = -2\log\frac{L_1}{L_0} = -2(\log\ L_1 - \log\ L_0) = -2(\log\ L_0 - \log\ L_S) - \left(-2(\log\ L_1 - \log\ L_S)\right) =$$
$$= Deviance_0 - Deviance_1$$

β. Χι-τετράγωνο του Pearson

Ένα άλλο μέτρο προσαρμογής του μοντέλου είναι η στατιστική συνάρτηση χι-τετράγωνο του Pearson. Ισούται με το άθροισμα των τετραγώνων των καταλοίπων Pearson και μεγάλες τιμές του υποδεικνύουν κακή προσαρμογή.

$$X^2 = \sum_i \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)}$$

df = n – k, όπως και της απόκλισης

n: πλήθος παρατηρήσεων

k: πλήθος παραμέτρων μοντέλου

γ. pseudo R-square

Τα ψευδο-R^2 είναι κάποια ενδεικτικά μέτρα που χρησιμεύουν στο να έχουμε μια πρώτη εικόνα για την ποιότητα του μοντέλου. Μια μικρή τιμή κοντά στο 0 θα δώσει ισχυρές ενδείξεις για την κακή επεξηγηματική ικανότητα του μοντέλου. Το αντίστροφο συμβαίνει για μια μεγάλη τιμή κοντά στο 1 (κάτι όμως που δε θα πρέπει να μας ενθουσιάσει). Τα ψευδο-R^2 όμως είναι έγκυρα και χρήσιμα για την σύγκριση μοντέλων, αρκεί αυτή να γίνεται μεταξύ ίδιου τύπου R^2. Παρακάτω, παρατίθενται τα πιο δημοφιλή, όπου με M_{Full} συμβολίζεται το πλήρες προσαρμοσμένο μοντέλο, με $M_{Intercept}$ το μοντέλο μόνο με τη σταθερά και με N ο αριθμός των παρατηρήσεων.

γ.1. McFadden's R²

Η μορφή του είναι η εξής:

$$R^2 = 1 - \frac{\ln L(M_{Full})}{\ln L(M_{Intercept})}$$

Αν συγκρίνουμε δυο μοντέλα από το ίδιο σετ δεδομένων το R^2 του McFadden θα είναι μεγαλύτερο για το μοντέλο με τη μεγαλύτερη πιθανοφάνεια. Μια μεγάλη τιμή δείχνει ότι το πλήρες μοντέλο έχει πολύ καλύτερη προσαρμογή από το αντίστοιχο με μόνο τη σταθερά.

γ.2. Cox & Snell R²

Έχει την ακόλουθη μορφή:

$$R^2 = 1 - \left(\frac{L(M_{Intercept})}{L(M_{Full})} \right)^{2/N}$$

Ο λόγος των πιθανοφανειών αντανακλάει τη βελτίωση του πλήρους μοντέλου, από αυτό με τη σταθερά (μικρότερος λόγος, μεγαλύτερη η βελτίωση). Η μέγιστη τιμή που μπορεί να πάρει το Cox & Snell R^2 είναι μικρότερη από 1 (αν το πλήρες μοντέλο προβλέψει άριστα το αποτέλεσμα με πιθανοφάνεια 1, τότε $R^2=1-(L(M_{Intercept}))^{2/N}$).

γ.3. Nagelkerke / Cragg & Uhler's R²

Η μορφή του είναι η εξής:

$$R^2 = \frac{1 - \left(\frac{L(M_{Intercept})}{L(M_{Full})} \right)^{2/N}}{1 - L(M_{Intercept})^{2/N}}$$

Το Nagelkerke R^2 προσαρμόζει το Cox & Snell, έτσι ώστε το εύρος των πιθανών τιμών του να φτάνει το 1 (όπου θα αντιστοιχεί στην τέλεια πρόβλεψη της απόκρισης). Έτσι αυτό το μέτρο είναι πιο χρήσιμο διαισθητικά.

δ. Hosmer-Lemeshow και le Cessie and Houwelingen έλεγχοι

Παρακάτω παρουσιάζονται δύο έλεγχοι για την προσαρμογή του μοντέλου, με τον δεύτερο να είναι μια εξέλιξη του πρώτου.

δ.1. Ο έλεγχος των Hosmer & Lemeshow

Ο έλεγχος ομαδοποιεί τις παρατηρήσεις βάσει των εκτιμώμενων πιθανοτήτων σε συγκεκριμένο αριθμό κλάσεων και υπολογίζει τα στατιστικά $\hat{C}$ και $\hat{H}$ μέσω του χ^2 του Pearson. Η μηδενική υπόθεση είναι ότι δεν υπάρχει διαφορά, ανάμεσα στις παρατηρούμενες και αναμενόμενες τιμές της μεταβλητής απόκρισης σε καμία ομάδα.

Το Hosmer-Lemeshow τεστ είναι εξαρτημένο από τον αριθμό των ομάδων που θα σχηματιστούν (απαιτεί προκαθορισμό και συνήθως η προεπιλογή είναι cutpoints=10) και απαιτεί συνεχείς ερμηνευτικές μεταβλητές. Επίσης είναι ασταθές στην παρουσία δεσμών (ties) ανάμεσα στις εκτιμώμενες πιθανότητες, δεν «τιμωρεί» την υπερπροσαρμογή και δε συστήνεται για μικρά δείγματα (n<400). Έτσι το τεστ αυτό αν και γνωστό, δεν έχει τόσο σημαντική ισχύ (Harrell, 2001).

δ.2. Ο έλεγχος των Le Cessie & Van Houwelingen

Πρόκειται για ένα τεστ νεότερο, χωρίς τα παραπάνω μειονεκτήματα, με μεγαλύτερη και πιο γενική ισχύ (Harrell, 2001). Είναι ένα μη παραμετρικό τεστ για το μη σταθμισμένο άθροισμα των τετραγώνων των σφαλμάτων. Η μηδενική υπόθεση είναι ότι οι πραγματικές πιθανότητες είναι αυτές που προκύπτουν από το μοντέλο.

ε. Πίνακες ταξινόμησης

Οι πίνακες ταξινόμησης (Classification Tables) είναι ένας καλός τρόπος να αξιολογήσουμε την προβλεπτική ικανότητα του μοντέλου. Αν η εκτιμώμενη πιθανότητα είναι ίση ή ξεπερνάει μία τιμή cutoff (0.5 η πλέον συνηθισμένη), τότε η περίπτωση προβλέπεται να έχει το γεγονός (Y=1), διαφορετικά προβλέπεται να μην το έχει (Y=0). Επίσης μια επιλογή ως τιμή cutoff που ακολουθείται, είναι το ποσοστό των παρατηρήσεων όπου Y=1.

Αν η $\hat{Y}$ είναι η πρόβλεψη της Y, τότε κάποια χρήσιμα μέτρα είναι τα παρακάτω:

Ευαισθησία (Sensitivity), όπου είναι το ποσοστό των αληθώς θετικών $(\hat{Y}=Y=1)$.

Ειδικότητα (Specificity), όπου είναι το ποσοστό των αληθώς αρνητικών $(\hat{Y}=Y=0)$.

Συγκεκριμένα:

$$Sensitivity = P(\hat{Y} = 1 \mid Y = 1) = \frac{\#(\hat{Y} = 1 \,\&\, Y = 1)}{\#(Y = 1)}$$

$$Specificity = P(\hat{Y} = 0 \mid Y = 0) = \frac{\#(\hat{Y} = 0 \,\&\, Y = 0)}{\#(Y = 0)}$$

$$P\left(correct\ clasiffication\right) = \frac{\#(\hat{Y} = Y)}{n} = P(\hat{Y} = 1 \ \&\ Y = 1) + P(\hat{Y} = 0 \ \&\ Y = 0) =$$

$$\left(Sensitivity\right)P\left(Y = 1\right) + \left(Specificity\right)P\left(Y = 0\right)$$

Η τελευταία πιθανότητα είναι η προβλεπτική ακρίβεια του μοντέλου, δηλαδή όλες οι περιπτώσεις που προβλέφθηκαν σωστά στο σύνολο των παρατηρήσεων.

Το ποσοστό ορθής ταξινόμησης του μοντέλου, θα πρέπει να είναι σημαντικά υψηλότερο από την ακρίβεια πρόβλεψης που θα επιτυγχάναμε από τύχη και μόνο. Ένας πρακτικός κανόνας είναι ότι, το ποσοστό ορθής ταξινόμησης θα πρέπει να είναι τουλάχιστον 25% πιο υψηλό από το ανάλογο ποσοστό κατά τύχη (σταθμισμένο με τις κατηγορίες της απόκρισης), για να έχει το μοντέλο υψηλή προβλεπτική ικανότητα (Schwab, 2003).

στ. Καμπύλες ROC

Μια τέτοια καμπύλη ορίζεται από τα σημεία του άξονα x "1 − Specificity" και y "Sensitivity", για όλες τις τιμές cutoff από 0 έως 1 και δίνει περισσότερη πληροφορία απ' ό,τι ένας πίνακας ταξινόμησης (που είναι για μία μόνο τιμή cutoff). Η καμπύλη ROC είναι η γραφική παράσταση της πιθανότητας των 'σωστά θετικών' $P(Y > c)$, ως προς την πιθανότητα των 'εσφαλμένα θετικών' $P(X > c)$ για ένα εύρος cutoff τιμών c.

Η επάνω αριστερή γωνία θεωρείται η βέλτιστη θέση σε ένα γράφημα ROC, δείχνοντας ένα υψηλό ποσοστό αληθώς θετικών και ένα χαμηλό ποσοστό ψευδώς θετικών.

Το εμβαδόν κάτω από την καμπύλη ROC ταυτίζεται με ένα μέτρο προβλεπτικής ισχύος του μοντέλου που καλείται «δείκτης συμφωνίας» C (Concordance Index ή ROC area ή AUC) και εκτιμάται με παραμετρικές και μη μεθόδους. Ο δείκτης C εκτιμά την πιθανότητα να είναι ίσες οι προβλέψεις με τις πραγματικές παρατηρήσεις.

Όσο μεγαλύτερος είναι ο δείκτης τόσο μεγαλύτερη θεωρείται η προβλεπτική ικανότητα του μοντέλου ($C \geq 0.8$ για ένα καλό μοντέλο). Αν ισούται με 0.50 τότε το μοντέλο προβλέπει τελείως τυχαία. Το μέτρο AUC είναι ιδιαίτερα χρήσιμο για τα

σύνολα δεδομένων με μη ισορροπημένη κατανομή της μεταβλητής απόκρισης (η μία κλάση κυριαρχεί έναντι της άλλης).

ζ. Brier & Skill *score*

Παρακάτω αναφέρονται δύο χρήσιμα μέτρα για το «σκοράρισμα» των μοντέλων.

ζ.1. Brier score

Είναι ένα μέτρο αξιολόγησης μοντέλων, όσον αφορά την ακρίβεια των προβλέψεων πιθανότητας, όπου υπολογίζει τη μέση τετραγωνική απόκλιση των εκτιμώμενων πιθανοτήτων κάποιων γεγονότων και του αποτελέσματός τους. Έτσι ένα χαμηλό σκορ εκφράζει υψηλή ακρίβεια (το τέλειο σκορ είναι 0). Ορίζεται ως:

$$BS = \frac{1}{N}\sum_{n=1}^{N}(p_n - o_n)^2$$

p_n: η πιθανότητα πρόβλεψης / o_n: το αποτέλεσμα / N: σύνολο περιπτώσεων

ζ.2. Skill score

Είναι μια αναπαράσταση του σφάλματος πρόβλεψης που σχετίζεται με την ακρίβεια πρόβλεψης ενός μοντέλου που μας ενδιαφέρει, με την αντίστοιχη ενός μοντέλου αναφοράς. Το score παίρνει τιμές στο (-∞, 1]. Η τέλεια πρόβλεψη επιτυγχάνεται με σκορ ίσο με 1. Παίρνει τιμή 0 αν τα δυο μοντέλα που συγκρίνονται έχουν την ίδια ικανότητα και αρνητικές τιμές αν οι προβλέψεις του μοντέλου είναι λιγότερο ικανές από τις αντίστοιχες του μοντέλου αναφοράς.

$$SS = 1 - \frac{BS}{BS_{ref}}$$

η. Πληροφοριακά κριτήρια

Δημοφιλή κριτήρια για τη σύγκριση μοντέλων είναι τα AIC (Akaike information criterion) και BIC (Bayesian information criterion ή Schwarz criterion). Το καλύτερο μοντέλο, θα είναι αυτό που τα ελαχιστοποιεί. Ορίζονται ως:

$$AIC = -2 \cdot lnL + 2 \cdot k$$
$$BIC = -2 \cdot lnL + k \cdot ln(n)$$

Όπου k: βαθμοί ελευθερίας του μοντέλου / n: αριθμός παρατηρήσεων / L: τιμή μέγιστης πιθανοφάνειας

Προσανατολίζονται στη μεγιστοποίηση της εκτιμώμενης πιθανοφάνειας, αλλά το BIC εφαρμόζει μεγαλύτερη ποινή στις ελεύθερες παραμέτρους για τα συνθετότερα μοντέλα και οδηγεί σε απλούστερα απ' ότι το AIC. Σε γενικές γραμμές το AIC αντανακλά τον κίνδυνο ένα μοντέλο να υπερπροσαρμοστεί (overfit, όπως το να περιέχει πολλές παραμέτρους συγκριτικά με το μέγεθος του δείγματος), ενώ το BIC να υποπροσαρμοστεί (underfit). Ένα μοντέλο που εμφανίζει ένα από τα δύο αυτά φαινόμενα, δεν έχει καλή προβλεπτική ικανότητα. Το AIC είναι αποδοτικότερο όταν στόχος είναι η μεγιστοποίηση της ορθής πρόβλεψης (Harrell, 2011), παρόλα αυτά οι απόψεις των ερευνητών για την χρήση αυτών των κριτηρίων διαφέρουν.

Επίσης γνωστό κριτήριο είναι το AICc (corrected AIC), που είναι το AIC με διόρθωση για μικρού μεγέθους δείγματα. Υπολογίζεται ως ακολούθως:

$$AICc = AIC + 2 \cdot k \cdot (k+1) / (n-k-1)$$

Προτείνεται να χρησιμοποιείται όταν το n είναι μικρό ή το k είναι μεγάλο, με έναν γενικό κανόνα n/k<40 (Burnham & Anderson, 2002). Στην επιλογή μεταξύ μοντέλων με το ίδιο k, τα AIC και AICc δίνουν πανομοιότυπα αποτελέσματα, όπως και όταν το n γίνεται μεγάλο.

ι. Μη παραμετρικές συσχετίσεις

Αναφέρονται τα σημαντικότερα μέτρα σχέσης, ανάμεσα στις εκτιμώμενες πιθανότητες και τις παρατηρούμενες αποκρίσεις.

ι.1. Kendall's tau-a

Αυτός ο μη παραμετρικός συντελεστής συσχέτισης, αποτελεί ένα εναλλακτικό μέτρο σχέσης του συντελεστή συσχέτισης του Spearman. Ορίζεται ως:

$$\tau_a = \frac{C - D}{\frac{1}{2} n(n-1)}$$

Όπου

C: αριθμός των «σύμφωνων» ζευγών (concordant pairs) / D: αριθμός των «ασύμφωνων» ζευγών (discordant pairs) / n: αριθμός παρατηρήσεων

Τα ζεύγη παρατηρήσεων (x_i, y_i) και (x_j, y_j) των τ.μ. X και Y αντίστοιχα, χαρακτηρίζονται ως «σύμφωνα» αν για τις τάξεις (ranks) των δυο στοιχείων ισχύει ότι {$x_i < x_j$ και $y_i < y_j$} ή {$x_i > x_j$ και $y_i > y_j$} και «ασύμφωνα» αν {$x_i < x_j$ και $y_i > y_j$} ή {$x_i > x_j$ και $y_i < y_j$}.

Όπως και σε άλλα μέτρα συσχετίσεων το tau-a παίρνει τιμές στο [-1, 1], με θετική συσχέτιση να σημαίνει ότι οι τάξεις και των δυο μεταβλητών αυξάνονται ή μειώνονται μαζί και αρνητική ότι οι τάξεις κινούνται αντίρροπα μεταξύ τους. Όταν είναι 0, οι τάξεις είναι ανεξάρτητες.

ι.2. Goodman-Kruskal Gamma

Είναι άλλο ένα μέτρο για την σχέση δυο μεταβλητών και προτιμάται σε σχέση με τα Kendall's tau-a και Spearman rank όταν υπάρχουν πολλές «ισοπαλίες» στις τάξεις των ζευγών, δηλαδή όταν $x_i = x_j$ ή $y_i = y_j$. Οι τιμές που παίρνει και η ερμηνεία του είναι όπως του Kendall. Ορίζεται ως:

$$G = \frac{C - D}{C + D}$$

ι.3. Κριτήριο Somers' D

Είναι ένα ασύμμετρο μέτρο της σχέσης μεταξύ δυο μεταβλητών, όπου μπορούμε να εκτιμήσουμε το D_{YX} σαν ένα μέτρο της επίδρασης της X πάνω στη Y ή μπορούμε να εκτιμήσουμε το D_{XY} σαν έναν δείκτη της απόδοσης ενός μοντέλου με ερμηνευτική μεταβλητή την X πάνω στην απόκριση Y. Το Somers' D στην λογιστική παλινδρόμηση παρέχει μια εκτίμηση της συσχέτισης της τάξης της παρατηρούμενης δίτιμης μεταβλητής απόκρισης και των προβλεπόμενων πιθανοτήτων και επομένως αποτελεί ένα δείκτη προσαρμογής του μοντέλου.

Το D ορίζεται ως $d_{XY} = \frac{C - D}{C + D + T}$,

όπου T είναι ο αριθμός δεσμών (tied pairs). Επίσης συνδέεται με το δείκτη συμφωνίας "C" μέσω της σχέσης $D_{XY} = 2(C - 0.5)$, δηλαδή στην ουσία πρόκειται για

μια έκδοση του C σε άλλη κλίμακα, όπου μπορεί να πάρει τιμές μεταξύ -1 και 1, σαν ένας συνηθισμένος συντελεστής συσχέτισης (αντί για 0 και 1).

4.12 Επικύρωση του μοντέλου (Cross-Validation)

Στο τελευταίο στάδιο αξιολόγησης του μοντέλου, εξετάζεται η ισχύς του μοντέλου ή αλλιώς, η ικανότητα γενίκευσης αυτού, μέσω της cross-validation ανάλυσης, χρησιμοποιώντας ένα τυχαίο δείγμα από τα συνολικά δεδομένα.

Μια συνήθης στρατηγική που ακολουθείται, είναι οι περιπτώσεις να χωρίζονται τυχαία σε δύο υποσύνολα: ένα δείγμα «εκπαίδευσης» (training sample) που περιέχει τα 2/3 των περιπτώσεων και ένα δείγμα «επικύρωσης» (validation sample) που περιέχει το υπόλοιπο 1/3 των περιπτώσεων. Στη συνέχεια το μοντέλο που προσαρμόστηκε στο πλήρες σετ δεδομένων, προσαρμόζεται στο δείγμα εκπαίδευσης και έτσι προκύπτει ένα νέο μοντέλο με τις ίδιες ερμηνευτικές μεταβλητές, αλλά διαφορετικούς συντελεστές. Αυτό με τη σειρά του, χρησιμοποιείται για να προβλέψει τη μεταβλητή απόκρισης του δείγματος επικύρωσης.

Η ακρίβεια της ταξινόμησης για το δείγμα επικύρωσης, χρησιμοποιείται για να εκτιμηθεί πόσο καλά το μοντέλο με βάση το δείγμα εκπαίδευσης, θα αποδώσει στον πληθυσμό (που αναπαρίσταται από το σύνολο των δεδομένων). Ένας πρακτικός κανόνας είναι ότι, εάν το ποσοστό της ορθής κατηγοριοποίησης του δείγματος επικύρωσης, έχει ίση ή μικρότερη από 10% διαφορά της αντίστοιχης του δείγματος εκπαίδευσης, αυτό είναι μια επαρκής απόδειξη της χρησιμότητας του λογιστικού μοντέλου (Schwab, 2003).

Εκτός όμως από την ικανοποίηση του κριτηρίου για την ακρίβεια κατηγοριοποίησης, απαιτούμε ότι η σημαντικότητα της συνολικής σχέσης των μεταβλητών του δείγματος εκπαίδευσης, είναι σύμφωνη με τα αποτελέσματα για το μοντέλο που προσαρμόστηκε στο πλήρες σύνολο δεδομένων.

ΚΕΦΑΛΑΙΟ 5

Προσαρμογή Logit & Probit μοντέλων στα δεδομένα

Θα προσαρμόσουμε Logit και Probit μοντέλα για να εξετάσουμε ποιες μεταβλητές από τα δεδομένα και σε τι βαθμό η κάθε μία, μπορούν να χρησιμοποιηθούν για να προβλέψουν τη νίκη ή μη της γηπεδούχου ομάδας.

5.1 Σημαντικότητα μεταβλητών

Στα GLM μοντέλα που θα χρησιμοποιήσουμε, πολύ σημαντική κρίνεται η χρήση της καταλληλότερης συνάρτησης σύνδεσης που συνδέει το αποτέλεσμα με τις ερμηνευτικές μεταβλητές. Στη διεθνή βιβλιογραφία (Bouler & Stekler, 1999) για το άθλημα του μπάσκετ, προτείνεται η χρήση probit μοντέλων με μεταβλητή απόκρισης το δίτιμο αποτέλεσμα, κάτι το οποίο καλούμαστε να εξετάσουμε αν επιβεβαιώνεται στη συνέχεια.

Στόχος μας, είναι μέσω των logit και probit μοντέλων να βρούμε ποιος συνδυασμός μεταβλητών, παρέχει την μεγαλύτερη ποσότητα πληροφορίας για τον προσδιορισμό του τελικού αποτελέσματος ενός αγώνα μπάσκετ, δηλαδή ποιες μεταβλητές μαζί προβλέπουν καλύτερα τη μεταβλητή *Win*.

Στην περιγραφική ανάλυση (Κεφ. 2) κάναμε κάποιες εκτιμήσεις για τους παράγοντες που επηρεάζουν σημαντικά το τελικό αποτέλεσμα του αγώνα. Είδαμε ότι η ευστοχία, οι ασίστ, τα λάθη και τα ριμπάουντ διαδραματίζουν σημαντικό ρόλο στο δρόμο για τη νίκη. Ενώ είδαμε και άλλες μεταβλητές, όπως τα φάουλ και τα κοψίματα, με εξαιρετικά ασθενή σχέση με το τελικό αποτέλεσμα και το τελικό σκορ.

Πριν προχωρήσουμε στη σύνθεση ενός ολοκληρωμένου λογιστικού μοντέλου, θα ήταν ενδιαφέρον να εξετάσουμε πόσο η κάθε ανεξάρτητη μεταβλητή συνεισφέρει στην απόκλιση, όταν αυτή βρίσκεται μόνη της ως ερμηνευτική μεταβλητή, σε ένα

λογιστικό μοντέλο με μεταβλητή απόκρισης τη *Win*. Έτσι, έπειτα από προσαρμογή 31 λογιστικών μοντέλων, προκύπτει ο παρακάτω συγκεντρωτικός πίνακας.

Πίνακας 5.1 Σημαντικότητα μεταβλητών στην πρόβλεψη του αποτελέσματος

A/A	*Win ~*	*Change in Deviance*	*Contribution %*	**P-Value**
1	dif_Pts	239.86	100	–
2	dif_Rkg	171.69	71.58	***
3	Rkg	112.68	46.98	***
4	dif_Q3	95.279	39.72	***
5	Pts	70.460	29.38	***
6	dif_Q2	67.577	28.17	***
7	dif_Q1	50.709	21.14	***
8	dif_As	42.642	17.78	***
9	dif_2Fg	38.679	16.13	***
10	As	27.141	11.32	***
11	dif_St	26.477	11.04	***
12	Reb_d	26.234	10.94	***
13	dif_3Fg	24.675	10.29	***
14	dif_To	22.713	9.47	***
15	dif_Reb	21.084	8.79	***
16	St	20.269	8.45	***
17	dif_Group	18.509	7.72	*
18	Group	17.339	7.23	**
19	X2Fg	16.229	6.77	***
20	Ft_at	13.335	5.56	***
21	X3Fg	10.899	4.54	***
22	X2Fg_at	9.9065	4.13	**
23	To	7.9881	3.33	**
24	Bl_ag	5.8428	2.44	*
25	Fl_rv	5.6776	2.37	*
26	*X3Fg_at*	*2.4907*	*1.04*	*0.1145*
27	*Bl_fv*	*0.6701*	*0.28*	*0.4130*
28	*Ft*	*0.3357*	*0.14*	*0.5623*
29	*Fl_cm*	*0.1688*	*0.07*	*0.6812*
30	*Reb_o*	*0.1543*	*0.06*	*0.6944*
31	*dif_Ft*	*0.0778*	*0.03*	*0.7803*

*** σημαντική στο α=0.001 ** σημαντική στο 0.01 * σημαντική στο 0.05

Κάποιες παρατηρήσεις δίνονται στη συνέχεια:

❖ Όπως είναι λογικό, η μεταβλητή *dif_Pt*s που είναι η διαφορά του τελικού σκορ κάθε αγώνα, προσδιορίζει άριστα τον τελικό νικητή.

❖ Ασφαλώς και οι μεταβλητές *Rkg*, *dif_Rkg* από μόνες τους σε ένα λογιστικό μοντέλο, προβλέπουν καλύτερα από κάθε άλλη μεταβλητή τον τελικό νικητή, αφού ενσωματώνουν συνδυαστική πληροφορία από το παιχνίδι των ομάδων.

❖ Οι μεταβλητές εκείνες που εκφράζουν διαφορές, περιέχουν μεγάλη ποσότητα πληροφορίας και έτσι επηρεάζουν σημαντικά το τελικό αποτέλεσμα.

❖ Παρατηρούμε ότι, από μόνοι τους οι καθοριστικότεροι παράγοντες για τον προσδιορισμό του νικητή, εμφανίζονται όσοι έχουν άμεση σχέση με τους πόντους και το σκοράρισμα, όπως οι τελικοί πόντοι, οι διαφορές του σκορ ανά περίοδο και οι διαφορές στην ευστοχία στα σουτ 2 πόντων (16.13% συνεισφορά) κατά κύριο λόγο και κατά δεύτερο στα σουτ 3 πόντων (10.29% συνεισφορά).

❖ Εντυπωσιακή είναι η συνεισφορά στην απόκλιση, από τη διαφορά στις ασίστ με 17.78%, ενώ δεν είναι αμελητέα η περίπου 9% συνεισφορά των διαφορών στα λάθη και στα ριμπάουντ.

❖ Η συνεισφορά των μεταβλητών 26-31 στην απόκλιση δεν κρίνεται σημαντική. Οι προσπάθειες για τρίποντα, τα φάουλ (κατά) και τα κοψίματα (υπέρ) δε φαίνεται να επηρεάζουν σημαντικά το τελικό αποτέλεσμα. Μάλιστα, η ευστοχία στις ελ. βολές είχε μηδενική συνεισφορά, ενώ οι προσπάθειες που έγιναν για αυτές έχουν σημαντική συνεισφορά (p-value < 0.001) με 5.56% στην απόκλιση.

❖ Οι μεταβλητές που είναι σημαντικές μόνο στο ε.σ. 1% και 5%, όπως τα φάουλ υπέρ, τα κοψίματα κατά, ο αριθμός των λαθών και οι προσπάθειες για δίποντα, πολύ δύσκολα θα καταφέρουν να ενσωματωθούν στο βέλτιστο μοντέλο.

❖ Επίσης, οι κατηγορικές μεταβλητές *Group* και *dif_Group* έχουν μικρότερο βαθμό σημαντικότητας, από τις μεταβλητές με ανάλογη συνεισφορά και αυτό συμβαίνει διότι απαιτούν 5 και 10 βαθμούς ελευθερίας αντίστοιχα, ενώ οι συνεχείς μόνο έναν.

❖ Τα ευρήματα των παραπάνω μοντέλων, για την πλειοψηφία των μεταβλητών ήρθαν σε συμφωνία με τα αντίστοιχα της περιγραφικής ανάλυσης, όσον αφορά τη σχέση τους με τη *Win*.

Πριν συνεχίσουμε, αξίζει να αναφέρουμε ότι σε πρόσφατη έρευνα (Witkos, 2010) με προσαρμογή OLS μοντέλων παλινδρόμησης, οι σημαντικότερες ερμηνευτικές μεταβλητές ήταν οι τελικοί πόντοι και ο αριθμός των λαθών ανά παιχνίδι, ενώ σε μια άλλη (Ibáñez et al, 2009) με χρήση διαχωριστικής ανάλυσης, οι σημαντικότεροι παράγοντες ήταν τα εύστοχα δίποντα σουτ, τα αμυντικά ριμπάουντ και οι ασίστ.

5.2 Επιλογή και προσαρμογή μοντέλων

Θα προσαρμόσουμε λογιστικά μοντέλα, με τις ερμηνευτικές μεταβλητές που βρήκαμε ότι είναι στατιστικά σημαντικές στην πρόβλεψη της *Win*. Το βέλτιστο μοντέλο που θα κατασκευαστεί, θα πρέπει να περιέχει μόνο σημαντικές μεταβλητές.

Από τις 185 περιπτώσεις, στις 65 συνέβη το ενδεχόμενο της ήττας (*Win*=0) για τη γηπεδούχο ομάδα και επομένως, τα μοντέλα που θα κατασκευάσουμε, δε θα πρέπει να περιέχουν περισσότερες από 6 ερμηνευτικές μεταβλητές, βάσει του εμπειρικού κανόνα του Peduzzi ότι θα πρέπει να αντιστοιχούν τουλάχιστον 10 παρατηρήσεις ανά ερμηνευτική μεταβλητή (Ενότητα 4.9).

Εκτός όμως από τις μεταβλητές που βρέθηκαν μη σημαντικές και δε θα συμπεριληφθούν στο λογιστικό μοντέλο που θα κατασκευάσουμε, υπάρχουν και κάποιες ακόμη που δεν είναι δόκιμο ή δε μπορούν να προσαρμοστούν στο τελικό μοντέλο. Οι *dif_Pts* και *dif_Rkg* ασφαλώς και δε μπορούν να εισαχθούν στο μοντέλο, διότι προσδιορίζουν τον τελικό νικητή. Η ειδική βαθμολογία *Rkg* επίσης δε θα συμπεριληφθεί στο τελικό μοντέλο μας, διότι πρόκειται για μια μεταβλητή με συνδυαστική πληροφορία και το ενδιαφέρον μας εστιάζεται στους επί μέρους παράγοντες, για το πως επηρεάζουν το τελικό αποτέλεσμα. Παρόλα αυτά, τα αποτελέσματα του βέλτιστου μοντέλου με αυτήν τη μεταβλητή, παρατίθενται στο Παράρτημα. Επίσης η *dif_Group* περιέχει μεγάλο αριθμό κατηγοριών (10) σε σύγκριση με το μέγεθος του δείγματος και δε μπορεί να χρησιμοποιηθεί για την κατασκευή του βέλτιστου μοντέλου.

Θα προσαρμόσουμε λογιστικά μοντέλα, με όλες τις μεθόδους επιλογής μεταβλητών που αναφέρθηκαν στη θεωρία.

Η διαδικασία έχει ως εξής:

Αρχικά εξαιρούμε τις μη σημαντικές μεταβλητές που αναφέρθηκαν και έπειτα προσαρμόζουμε ένα μοντέλο με όλες τις υπόλοιπες (πλήρες μοντέλο). Λόγω του πλήθους των μεταβλητών, η stepwise μέθοδος θα λειτουργήσει ως backward μέθοδος για την επιλογή των σημαντικών μεταβλητών. Δηλαδή, η διαδικασία θα εκκινήσει από το πλήρες μοντέλο και ύστερα κατά βήμα, θα αποκλείονται οι μεταβλητές που έχουν τη μικρότερη (μη σημαντική) συμμετοχή στην απόκλιση (p-value ≥ 0.05) όταν συνυπάρχουν με όλες τις άλλες, έως ότου καταλήξουμε σε ένα μοντέλο, όπου στο σύνολό του οι μεταβλητές θα είναι στατιστικώς σημαντικές (p-value < 0.05).

Έτσι, τα αποτελέσματα με τη μέθοδο stepwise είναι τα παρακάτω:

Πίνακας 5.2 *Επιλογή μεταβλητών μέσω της μεθόδου Stepwise*

```
> glm<-
glm(Win~Pts+X2Fg+X2Fg_at+X3Fg+Ft_at+Reb_d+As+St+To+Bl_ag+Fl_rv+dif_2F
g+dif_3Fg+dif_Reb+dif_As+dif_St+dif_To+dif_Q1+dif_Q2+dif_Q3+Group,fam
ily=binomial)
Warning messages:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> step<-step(glm)
Start:  AIC=44
Win ~ Pts + X2Fg + X2Fg_at + X3Fg + Ft_at + Reb_d + As + St +
    To + Bl_ag + Fl_rv + dif_2Fg + dif_3Fg + dif_Reb + dif_As +
    dif_St + dif_To + dif_Q1 + dif_Q2 + dif_Q3 + Group

          Df Deviance    AIC
- Fl_rv    1    0.000 42.000
- dif_Q2   1    0.000 42.000
- dif_Q3   1    0.000 42.000
- Group    1    0.000 42.000
- X3Fg     1    0.000 42.000
- dif_As   1    0.000 42.000
- As       1    0.000 42.000
- Reb_d    1    0.000 42.000
- X2Fg     1    0.000 42.000
- dif_St   1    0.000 42.000
- Pts      1    0.000 42.000
- To       1    0.000 42.000
- St       1    0.000 42.000
- dif_Q1   1    0.000 42.000
- Ft_at    1    0.000 42.000
- Bl_ag    1    0.000 42.000
- X2Fg_at  1    0.000 42.000
- dif_To   1    0.000 42.000
- dif_2Fg  1    0.000 42.000
<none>           0.000 44.000
- dif_3Fg  1   37.529 79.529
- dif_Reb  1   39.088 81.088

.....

Step:  AIC=20
Win ~ X2Fg_at + Ft_at + St + Bl_ag + dif_2Fg + dif_3Fg + dif_Reb +
    dif_To + dif_Q1

          Df Deviance     AIC
<none>         0.000  20.000
- St       1   20.602  38.602
- X2Fg_at  1   31.342  49.342
- dif_Q1   1   32.870  50.870
- Bl_ag    1   34.695  52.695
- Ft_at    1   53.809  71.809
- dif_Reb  1   65.476  83.476
- dif_2Fg  1   76.083  94.083
- dif_To   1   77.420  95.420
- dif_3Fg  1  106.569 124.569
There were 12 warnings (use warnings() to see them)
```

Με αυτήν τη μέθοδο επελέγησαν 9 ερμηνευτικές μεταβλητές, με το κριτήριο AIC να βελτιώνεται αισθητά, δίνοντας τιμή 20.

Εντύπωση προκαλεί ότι, στις επιλεχθείσες μεταβλητές, ως αντιπροσωπευτική μεταβλητή του σκορ του παιχνιδιού εισήχθη η *dif_Q1* (που εκφράζει τη διαφορά πόντων μετά το πέρας της πρώτης περιόδου) και όχι μία εκ των *Pts*, *dif_Q3* ή *dif_Q2,* οι οποίες εμφανίζονταν να δίνουν μεγαλύτερη ποσότητα πληροφορίας. Επίσης, έκπληξη αποτελεί το γεγονός, ότι δε συμπεριλήφθηκε κάποια εκ των δύο μεταβλητών που αντιπροσωπεύουν τις ασίστ ενός αγώνα, η διαφορά των οποίων έχει αναφερθεί επανειλημμένα σε διάφορα σημεία αυτής της μελέτης.

Ασφαλώς ένα τέτοιο μοντέλο με τόσες μεταβλητές και τόσο λίγα δεδομένα, δε μπορεί να αξιοποιηθεί για να προβλέψει το γεγονός, λόγω του ότι προκύπτουν σοβαρά αριθμητικά προβλήματα στην εκτίμηση των συντελεστών του μοντέλου. Υπάρχει έντονη υπερπροσαρμογή, καθώς το τελικό αποτέλεσμα προβλέπεται σχεδόν τέλεια.

Έτσι εργαστήκαμε ως εξής, απομακρύναμε κατά βήμα την πιο ασήμαντη μεταβλητή (με το μεγαλύτερο p-value), έως ότου να καταλήξουμε σε ένα μοντέλο, όπου όλες οι μεταβλητές να είναι σημαντικές σε ε.σ. 5%. Αυτό ήταν το εξής:

$$\log it(p_i) = -5.44 + 0.41 \times dif\_2Fg + 0.32 \times dif\_3Fg - 1.04 \times dif\_To + 0.46 \times dif\_\mathrm{Re}b + 0.31 \times Ft\_at$$

Ένα μοντέλο με 5 μεταβλητές (τα αποτελέσματα παρατίθενται στον πίνακα Π.10), με εξαιρετικά χαμηλή τελική απόκλιση (Residual deviance) ίση με 39.511 και το AIC κριτήριο ίσο με 51.511. Όμως, το προειδοποιητικό μήνυμα «Warning message: glm.fit: fitted probabilities numerically 0 or 1 occurred» μας πληροφορεί ότι, προκύπτουν εκτιμώμενες τιμές που παίρνουν ακραίες τιμές στο διάστημα [0,1] και συγκεκριμένα 17.

Ναι μεν θέλουμε ένα μοντέλο που να ερμηνεύει, όσο γίνεται περισσότερο την απόκλιση, χρειαζόμαστε όμως ένα ευσταθές μοντέλο, λιγότερο ευαίσθητο, το οποίο να μπορεί εν δυνάμει να γενικευτεί και σε άλλα σετ δεδομένων, για την πρόβλεψη του τελικού αποτελέσματος. Έτσι, από το προηγούμενο μοντέλο με όλους τους δυνατούς συνδυασμούς των 5 μεταβλητών, καταλήξαμε σε ένα μοντέλο με 4 ερμηνευτικές μεταβλητές (απομακρύνθηκε η dif_Reb). Σε αυτό, αντιστοιχούν σε κάθε μία ερμηνευτική μεταβλητή περίπου 16 (65/4) παρατηρήσεις από την κατηγορία

Win=0 (όπου είναι και η μικρότερη), εκπληρώνοντας έτσι και τον κανόνα του Harrell για τουλάχιστον 15 αποκρίσεις για κάθε μεταβλητή (βλ. 4.9).

Στη συνέχεια όμως, χρησιμοποιήσαμε και τη μέθοδο forward για να δούμε και να συγκρίνουμε τα αποτελέσματα. Για το λόγο ότι, η σειρά εισαγωγής των ανεξάρτητων μεταβλητών παίζει ρόλο στην τελική διαμόρφωση του μοντέλου, η βέλτιστη εισαγωγή αυτών έγινε βάσει της υπεροχής κάποιων μεταβλητών, έναντι άλλων. Έτσι, προσαρμόστηκε ένα μοντέλο με την dif_Q3 που έχει τη μεγαλύτερη συνεισφορά από τις διαθέσιμες και ύστερα ανά βήμα, προσαρμόζαμε μία-μία όλες τις υπόλοιπες. Εκείνη που είχε τη μεγαλύτερη συνεισφορά, εισήχθη στο μοντέλο. Με την ίδια λογική προσαρμόζονται και οι υπόλοιπες μέχρι να καταλήξουμε σε ένα μοντέλο με μέγιστο αριθμό 5-6 ερμηνευτικών μεταβλητών.

Μετά από διαδοχικές προσαρμογές μοντέλων, στην προσπάθεια να μεγιστοποιηθεί το κριτήριο AIC και παράλληλα να μειωθεί η τελική απόκλιση (Residual deviance), το μοντέλο που προέκυψε περιείχε τις εξής μεταβλητές: dif_Q3, dif_2Fg, dif_3Fg, dif_Reb, dif_To, Ft_at. Το αποτέλεσμα μας χαροποίησε, καθώς μέσα στις επιλεγόμενες μεταβλητές της μεθόδου συμπεριλήφθηκαν και οι 5 μεταβλητές που είχαμε δει λίγο πριν.

Επίσης σε δοκιμές που έγιναν, αποφύγαμε τη συνύπαρξη μεταβλητών με υψηλές συσχετίσεις. Να αναφερθεί ότι, η επιλογή των κατάλληλων μεταβλητών σε ένα μοντέλο, συχνά έγκειται σε υποκειμενικά χαρακτηριστικά, αναλόγως τον ερευνητή και τη φύση του προβλήματος. Τέλος, η κατηγορική μεταβλητή Group, θεωρήθηκε μη σημαντική σε ε.σ. 5% για εισαγωγή σε ένα βέλτιστο μοντέλο και το ίδιο συνέβη και με τις αλληλεπιδράσεις αυτής με τις συνεχείς μεταβλητές.

5.3 Το επιλεχθέν logit μοντέλο και η ερμηνεία του

Αρχικά, παραθέτουμε τον πίνακα ανάλυσης της απόκλισης για το καταληκτικό μοντέλο με τις 4 ερμηνευτικές μεταβλητές που αναφέρθηκε.

Πίνακας 5.3 *Πίνακας ανάλυσης της απόκλισης*

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Win

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                      184     239.86
dif_2Fg  1   38.679       183     201.18 4.996e-10 ***
dif_3Fg  1   38.712       182     162.47 4.912e-10 ***
dif_To   1   37.601       181     124.87 8.680e-10 ***
Ft_at    1   27.001       180      97.87 2.033e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ο πίνακας αυτός μας πληροφορεί, πόσο σημαντική είναι μια μεταβλητή για να εισαχθεί στο μοντέλο. Τα εξαιρετικά μικρά p-values των στατιστικών χ^2 υποδεικνύουν ότι, κάθε μια από τις 4 μεταβλητές βελτιώνει σημαντικά την προσαρμογή του μοντέλου. Βλέπουμε ότι, οι 4 συνδυαστικοί παράγοντες παίζουν σπουδαίο ρόλο στη διαμόρφωση του νικητή.

Τη μεγαλύτερη αλλά και ισάξια συνεισφορά στην απόκλιση φαίνεται να έχουν οι dif_2Fg και dif_3Fg, αφού συνεισφέρουν περίπου 39 μονάδες η κάθε μία σε σύνολο 239.86 μονάδων της απόκλισης (με αποτέλεσμα αυτή να μειώνεται σε 162.47 μονάδες), με τις dif_To και Ft_at να ακολουθούν.

Στη συνέχεια, παραθέτονται τα αναλυτικά αποτελέσματα με τους συντελεστές του μοντέλου και τη σημαντικότητα αυτών.

Πίνακας 5.4 *Ανάλυση των MLE εκτιμήσεων του μοντέλου*

```
Call:
glm(formula = Win ~ dif_2Fg + dif_3Fg + dif_To + Ft_at, family =
binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.47999  -0.22119   0.09179   0.33758   2.21612

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.07986    0.84462  -3.646 0.000266 ***
dif_2Fg      0.19682    0.03624   5.431 5.62e-08 ***
dif_3Fg      0.14582    0.02554   5.710 1.13e-08 ***
dif_To      -0.31745    0.06338  -5.009 5.47e-07 ***
Ft_at        0.18157    0.04096   4.432 9.32e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 239.86  on 184  degrees of freedom
Residual deviance:  97.87  on 180  degrees of freedom
AIC: 107.87

Number of Fisher Scoring iterations: 7
```

Ο παραπάνω πίνακας παρουσιάζει τους συντελεστές (στήλη Estimate) του μοντέλου και παρατηρούμε ότι:

- Τα p-values των στατιστικών Wald για τις παραπάνω παραμέτρους είναι <0.001, κατά πολύ μικρότερα από το επίπεδο σημαντικότητας 0.05. Η αρχική υπόθεση ότι οι συντελεστές β είναι 0, απορρίπτεται εμφατικά. Όλες οι παράμετροι είναι σημαντικές και τα μεγέθη των εκτιμήσεων δεν είναι μεγάλα.
- Τα τυπικά σφάλματα είναι μικρά και δεν εμφανίζεται κανένα > 2 (Schwab, 2003), που να υποδεικνύει κάποιο αριθμητικό πρόβλημα.
- Αν συγκρίνουμε τα p-values του πίνακα της τυπικής απόκλισης και της ανάλυσης των συντελεστών, θα δούμε ότι αυτά είναι πολύ κοντά για κάθε μεταβλητή, πράγμα που υποστηρίζει την προηγούμενη διαπίστωση.
- Η τελική απόκλιση του μοντέλου είναι -2LL=97.87. Ένα καλό μοντέλο θα πρέπει να έχει μικρή τιμή (τέλεια προσαρμογή όταν -2LL=0).
- Ο αλγόριθμος μεγιστοποίησε το λογάριθμο της πιθανοφάνειας (συνέκλινε) ύστερα από 7 επαναλήψεις. Πρακτικά λίγες επαναλήψεις, δείχνουν ευστάθεια του μοντέλου.

Ο παρακάτω πίνακας αναφέρει τη μέγιστη και ελάχιστη προσαρμοσμένη τιμή.

Πίνακας 5.5 *Προσαρμοσμένες τιμές του μοντέλου*

```
Fitted Values
   Case        Value
Min 121 0.0002382935
Max 172 0.9999627126
```

Το μοντέλο δε δίνει καμία προσαρμοσμένη τιμή ακριβώς 0 ή 1, κάτι που θα οδηγούσε σε πιθανή υπερπροσαρμογή και αστάθεια του μοντέλου.

Ερμηνεία συντελεστών

Το εκτιμώμενο μοντέλο είναι το παρακάτω:

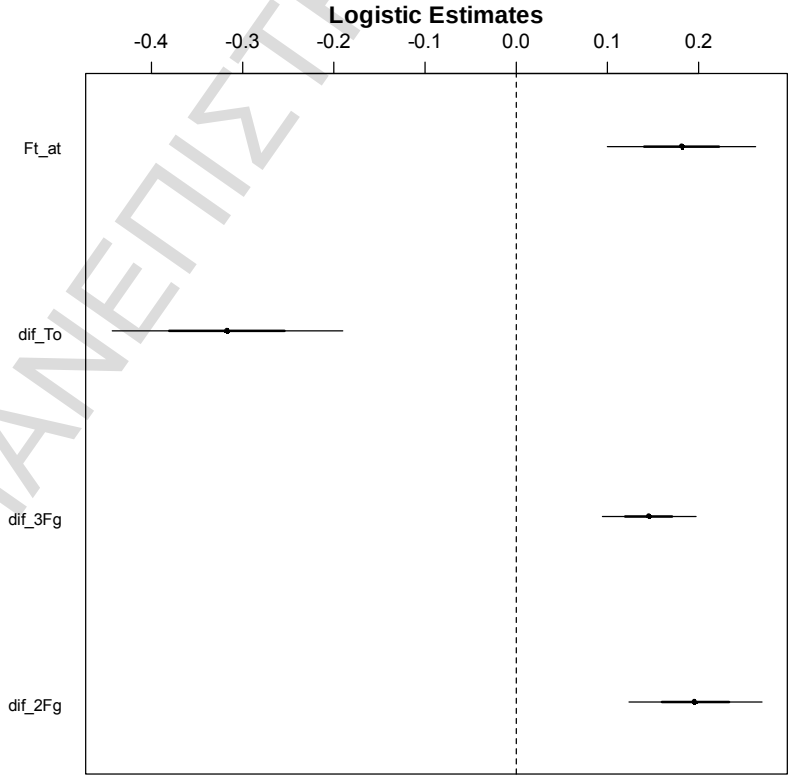
$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -3.08 + 0.20 \times dif\_2Fg + 0.15 \times dif\_3Fg - 0.32 \times dif\_To + 0.18 \times Ft\_at$$

και η εκτιμώμενη σχετική πιθανότητα είναι:

$$\frac{\hat{p}_i}{1-\hat{p}_i} = e^{-3.08} \times e^{0.20 \times dif\_2Fg} \times e^{0.15 \times dif\_3Fg} \times e^{-0.32 \times dif\_To} \times e^{0.18 \times Ft\_at}$$

Το ακόλουθο γράφημα, απεικονίζει τα μεγέθη των παραμέτρων:

Σχήμα 5.1 *Γράφημα εκτιμήσεων του μοντέλου με 95% δ.ε.*



Η παράμετρος της dif_To προκαλεί τη μεγαλύτερη μεταβολή στα log-odds της Win, σε σχέση με τις υπόλοιπες και μάλιστα αυτή είναι αρνητική (-0.32). Αυτή η εκτίμηση έχει και το μεγαλύτερο διάστημα εμπιστοσύνης. Έπειτα δεύτερη σε μέγεθος κατ' απόλυτη τιμή είναι η dif_2Fg, που μάλιστα έχει μικρότερο δ.ε. από τη λίγο μικρότερη σε μέγεθος εκτίμηση Ft_at. Ο συντελεστής της dif_3Fg φαίνεται να έχει τη μικρότερη επίδραση, με στενά όμως όρια δ.ε. λόγω του μικρού τυπικού σφάλματος της εκτίμησης.

Ένα παράδειγμα ερμηνείας παραμέτρου του μοντέλου είναι ότι, για κάθε μοναδιαία αύξηση της Ft_at, ο λογάριθμος της σχετικής πιθανότητας (log odds) της νίκης (έναντι της ήττας) αυξάνεται κατά 0.18 και όμοια για τις υπόλοιπες. Όμως, μια πιο διαισθητική ερμηνεία των συντελεστών, θα γίνει με τη βοήθεια του παρακάτω πίνακα.

Πίνακας 5.6 *Λόγοι σχετικών πιθανοτήτων*

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits		Exp(- Coef)
dif_2Fg	1.21752409	1.142148806	1.3181681	0.8213390
dif_3Fg	1.15698448	1.105941794	1.2234874	0.8643158
dif_To	0.72800371	0.634501486	0.8154997	1.3736194
Ft_at	1.19910168	1.112962127	1.3091527	0.8339576

Αυτός ο πίνακας δείχνει τους συντελεστές, ως odds ratios. Τα Exp(Coef) εκφράζουν πόσο πιθανό είναι να νικήσουν οι γηπεδούχοι, έναντι του να μη νικήσουν.

Παρατηρήσεις:

- Οι 3 από τους 4 συντελεστές των ανεξάρτητων μεταβλητών είναι θετικοί, δηλαδή υψηλές τιμές αυτών συνδέονται με αυξημένες πιθανότητες νίκης, ενώ η μεταβλητή που αναπαριστά τα λάθη των ομάδων και έχει αρνητικό συντελεστή, δηλώνει ότι λιγότερα λάθη αυξάνουν την πιθανότητα νίκης.
- Η σταθερά έχει αρνητικό συντελεστή και εκφράζει ότι όταν οι συντελεστές των ανεξάρτητων μεταβλητών είναι 0, η γηπεδούχος ομάδα έχει σχετική πιθανότητα να νικήσει ($e^{-3.08}$=) 0.046.

- Η σχετική πιθανότητα της dif_2Fg είναι 1.22, δηλαδή αν η γηπεδούχος ομάδα αυξήσει από τον αντίπαλο την ευστοχία στη διαφορά διπόντων κατά 1%, αυτό αυξάνει την σχετική πιθανότητα να νικήσει κατά περίπου 22% (δηλαδή η πιθανότητα να νικήσει είναι 1.22 φορές την πιθανότητα να χάσει), με την προυπόθεση ότι οι υπόλοιπες μεταβλητές παραμένουν σταθερές.
- Για τη διαφορά τριπόντων, η σχετική πιθανότητα να νικήσει είναι μόλις 1.16, με 95% δ.ε. της εκτίμησης {1.11, 1.22}.
- Αν οι γηπεδούχοι κάνουν ένα λάθος περισσότερο, μειώνουν τη σχετική πιθανότητα της νίκης κατά (1-0.73 =) 27% και αν κάνουν ένα λάθος λιγότερο, την αυξάνουν κατά (Exp(–Coef) = 1.37) 37%.
- Μια επιπλέον προσπάθεια στις ελεύθερες βολές, αυξάνει την σχετική πιθανότητα να νικήσει κατά 20% (έναντι του να μη νικήσει), ενώ μια προσπάθεια λιγότερη μειώνει αυτήν τη πιθανότητα κατά 17% (= 1-0.83).
- Τα διαστήματα εμπιστοσύνης είναι στενά και όχι μεγάλα, ακόμη μία ένδειξη για αξιόπιστες εκτιμήσεις του μοντέλου.

5.4 Διαγνωστικοί έλεγχοι

5.4.1 Έλεγχοι Καλής Προσαρμογής

1. Θα χρησιμοποιήσουμε το χ^2 τεστ του λόγου της πιθανοφάνειας (likelihood ratio chi-square test).

Πίνακας 5.7 *Έλεγχος λόγου πιθανοφάνειας του μοντέλου*

```
Likelihood ratio test for MLE method
Chi-squared 4 d.f. =  141.9931 , P value =  1.056595e-29
```

Το p-value του ελέγχου χ-τετράγωνο (141.9931) είναι <0.0001, κατά πολύ μικρότερο από το επίπεδο σημαντικότητας 0.05. Η αρχική υπόθεση ότι δεν υπάρχει διαφορά ανάμεσα στο μοντέλο με μόνο την σταθερά και στο προσαρμοσμένο μοντέλο, απορρίπτεται εμφατικά. Έτσι, υπάρχει σημαντική σχέση ανάμεσα στις ερμηνευτικές μεταβλητές και το τελικό αποτέλεσμα (Win).

2. Ακολουθεί ο πίνακας ταξινόμησης, ένα πολύ σημαντικό μέτρο αξιολόγησης της προβλεπτικής ικανότητας του μοντέλου.

Πίνακας 5.8 *Πίνακας ταξινόμησης*

Classification table					
			Predicted		
	Win		0	1	**% Correct**
Observed	0	65	51	14	**78.46**
	1	120	9	111	**92.50**
Overall	185				**87.57**

Το 87.57% των αποτελεσμάτων προβλέφθηκε σωστά από το μοντέλο, δηλαδή 23 από τα 185 παιχνίδια δεν κατηγοριοποιήθηκαν σωστά. Πιο συγκεκριμένα από τα 65 παιχνίδια που έληξαν με ήττα της γηπεδούχου ομάδας, εκτιμήθηκαν ορθά τα 51 σε ποσοστό 78.46% (ειδικότητα – specificity), ενώ από τα 120 παιχνίδια που έληξαν με νίκη, εκτιμήθηκαν σωστά τα 111 σε ποσοστό της τάξης του 92.50% (ευαισθησία – sensitivity).

Για να είναι επαρκές το συνολικό ποσοστό σωστής κατηγοριοποίησης, αυτό θα πρέπει να είναι μεγαλύτερο του 25% από την κατά τύχη πιθανότητα. Η κατά τύχη πιθανότητα σταθμισμένη με την αναλογία των δύο κατηγοριών της Win είναι 0.544 ($= 0.351^2 + 0.649^2$), δηλαδή χωρίς κάποιο μοντέλο, απλά και μόνο από τύχη η πιθανότητα επιλογής του νικητή ή ηττημένου είναι 54.4%. Σύμφωνα με το κριτήριο ισχύει ότι 87.57% > 68% (1.25 x 54.4%), πράγμα που σημαίνει ότι το κριτήριο ικανοποιείται απόλυτα.

Στον προηγούμενο πίνακα ως σημείο cut-off θεωρήθηκε το 0.5, όπου αν η προσαρμοσμένη τιμή της Win, είναι ίση ή μεγαλύτερη από αυτό, τότε η περίπτωση κατηγοριοποιείται ως νίκη και αν είναι μικρότερη ως ήττα. Θα ήταν ενδιαφέρον να δούμε τι συμβαίνει σε διαφορετικά cut-off σημεία. Θυμίζουμε ότι το παρατηρούμενο ποσοστό νίκης είναι 64.9%.

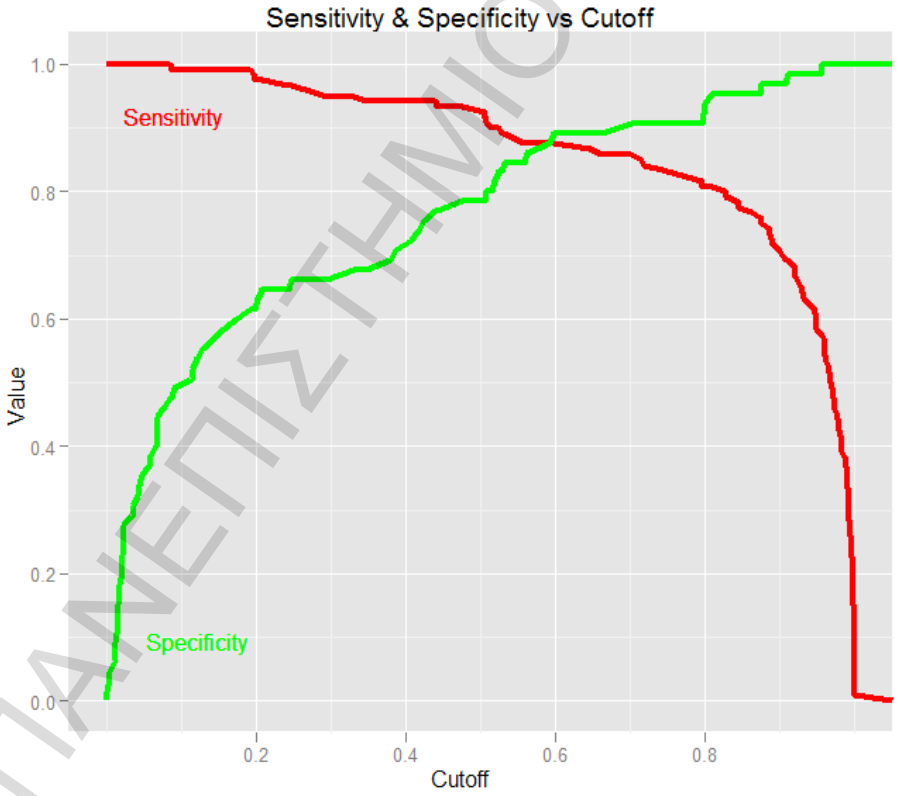
Ακολουθεί πίνακας με τα στοιχεία της ευαισθησίας, ειδικότητας και συνολικής ακρίβειας, συναρτήσει των cutoff σημείων ανά 0.1.

Πίνακας 5.9 *Ακρίβεια ταξινόμησης για διάφορες τιμές του cutoff*

Classification accuracy									
Cut-off	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Sensitivity	0.992	0.975	0.950	0.942	0.925	0.867	0.858	0.808	0.700
Specificity	0.508	0.631	0.677	0.723	0.846	0.892	0.908	0.938	0.969
% Correct	82.16	85.41	85.41	86.49	87.57	87.57	87.57	85.41	79.46

Βάσει του πίνακα, μπορούμε να αποφασίσουμε για την τιμή του cutoff, ανάλογα αν ο στόχος μας είναι απλά η συνολικά σωστή κατηγοριοποίηση ή η αύξηση της ακρίβειας από τη μία ή την άλλη κατηγορία. Παρατηρούμε ότι η συνολική ακρίβεια είναι η μέγιστη στα σημεία από το 0.5 έως και το 0.7, όπου συμπεριλαμβάνεται και το 0.65, η παρατηρούμενη αναλογία της νίκης. Όπως και το cutoff σημείο που προσδιορίζεται από το επόμενο γράφημα.

Σχήμα 5.2 *Διάγραμμα ευαισθησίας και ειδικότητας ανά cut-off σημείο*

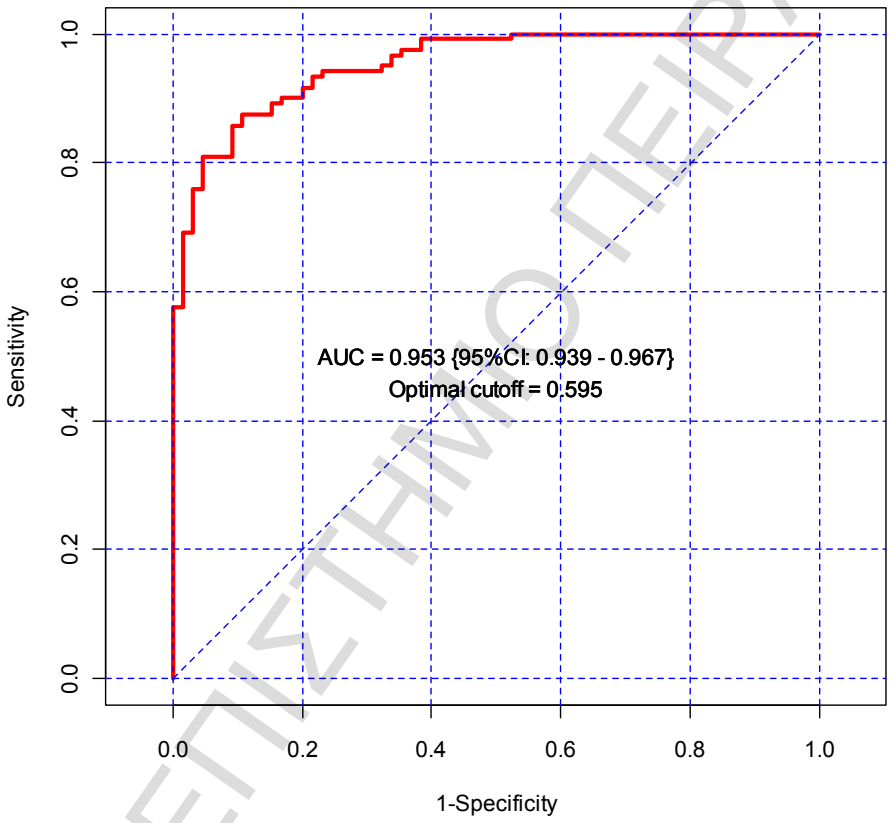


Παρατηρούμε ότι, όσο αυξάνεται η τιμή του cutoff η ευαισθησία μειώνεται, ενώ αντίθετα η ειδικότητα αυξάνεται. Στο σημείο 0 κατηγοριοποιούνται όλοι ως νικητές, ενώ στο 1, ως ηττημένοι. Όσο αυξάνει το cutoff, η τιμή πρόβλεψης της νίκης αυξάνει (πράσινη γραμμή), ενώ η τιμή πρόβλεψης της ήττας μειώνεται (κόκκινη γραμμή). Η ευαισθησία φαίνεται να έχει απότομες αλλαγές από το 0.6 σημείο, ενώ οι τιμές της

ειδικότητας κινούνται πιο ομαλά. Οι τιμές είναι περίπου ίσες για τιμή 0.59 του κατωφλιού, όπου η ακρίβεια κατηγοριοποίησης παίρνει και τη μέγιστή της τιμή.

4. Ένα ακόμη μέτρο της προβλεπτικής ισχύος του μοντέλου είναι η καμπύλη ROC.

Σχήμα 5.3 *Καμπύλη ROC*



Η χαρακτηριστική καμπύλη ROC, ορίζεται από την αναλογία των ψευδώς θετικών (οριζόντιος άξονας) και την αντίστοιχη των αληθώς θετικών (κάθετος άξονας) για όλες τις τιμές cutoff.

Το μοντέλο κρίνεται ιδιαίτερα επιτυχημένο, καθώς η καμπύλη ROC απέχει πολύ από την διαγώνιο και το εμβαδόν κάτω από την καμπύλη είναι 0.953. Η τιμή είναι κοντά στο 1, που σημαίνει ότι η συνολική ακρίβεια του μοντέλου είναι εξαιρετικά υψηλή και κατά συνέπεια, υψηλές θα είναι και οι τιμές της ευαισθησίας και της ειδικότητας. Έτσι είναι κατά πολύ καλύτερο από ένα τυχαίο μοντέλο (με εμβαδόν ίσο

με 0.5) και επομένως οι προβλέψεις του μοντέλου με τις πραγματικές παρατηρήσεις είναι πολύ κοντά.

Τέλος, το cutoff σημείο που είναι πιο κοντά στην επάνω αριστερή γωνία έχει τιμή 0.595 και θεωρείται το βέλτιστο (closest topleft και max accuracy methods), όσον αφορά τη διάκριση μεταξύ των νικητών και των ηττημένων.

5. Στη συνέχεια, θα εφαρμόσουμε 2 ελέγχους για το αν υπάρχει έλλειψη προσαρμογής, το Hosmer-Lemeshow τεστ και le Cessie and Houwelingen τεστ.

Πίνακας 5.10 *Έλεγχος Hosmer-Lemeshow*

```
$C
        Hosmer-Lemeshow C statistic

data:  fitted(glm) and Win
X-squared = 1.7149, df = 8, p-value = 0.9885

$H
        Hosmer-Lemeshow H statistic

data:  fitted(glm) and Win
X-squared = 4.0036, df = 8, p-value = 0.8568
```

Τα πολύ υψηλά p-values των παραπάνω στατιστικών, μας οδηγούν στο να αποδεχθούμε την αρχική υπόθεση ότι οι παρατηρούμενες με τις εκτιμώμενες συχνότητες από το μοντέλο δε διαφέρουν σημαντικά, κάτι που σημαίνει καλή προσαρμογή του μοντέλου.

Πίνακας 5.11 *Έλεγχος le Cessie and Houwelingen*

```
le Cessie-van Houwelingen-Copas-Hosmer global goodness of fit test

data:  fitted(glm) and Win
z = 1.3754, p-value = 0.169
```

P-value=0.169 > 0.05 του ελέγχου και έτσι, δεν απορρίπτουμε την υπόθεση ότι οι πραγματικές πιθανότητες είναι αυτές που προκύπτουν από το μοντέλο.

6. Για την ακρίβεια των προβλέψεων του μοντέλου, θα παρατεθεί το Brier και το Skill score.

Πίνακας 5.12 *Brier & Skill score*

$bs
0.08660811
$ss
0.6199792

Το Brier score είναι πολύ κοντά στο 0 και εκφράζει υψηλή ακρίβεια, ενώ το skill score θεωρείται υψηλό λόγω της «αυστηρότητας» του δείκτη, όπου εκφράζει υψηλή προβλεπτική ακρίβεια του μοντέλου.

7. Παρακάτω αναφέρονται οι τιμές των σημαντικότερων τύπων ψευδο-R^2, που αποτελούν ενδεικτικά μέτρα καλής προσαρμογής.

Πίνακας 5.13 *Pseudo-R^2*

r2McFadden	r2Cox&Snell	r2Nagelkerke
0.5919750	0.5358419	0.7375372

Οι τιμές των ψευδο-R^2 χαρακτηρίζονται ως αρκετά καλές και δείχνουν καλή προσαρμογή του μοντέλου. Διαισθητικά, αυτό το αντιλαμβανόμαστε καλύτερα με το τελευταίο ψευδο-R^2, όπου είναι προσαρμοσμένο για να παίρνει την τιμή 1 με την τέλεια προσαρμογή.

8. Οι σημαντικότεροι μη παραμετρικοί συντελεστές συσχέτισης των εκτιμώμενων και των παρατηρούμενων τιμών είναι οι ακόλουθοι.

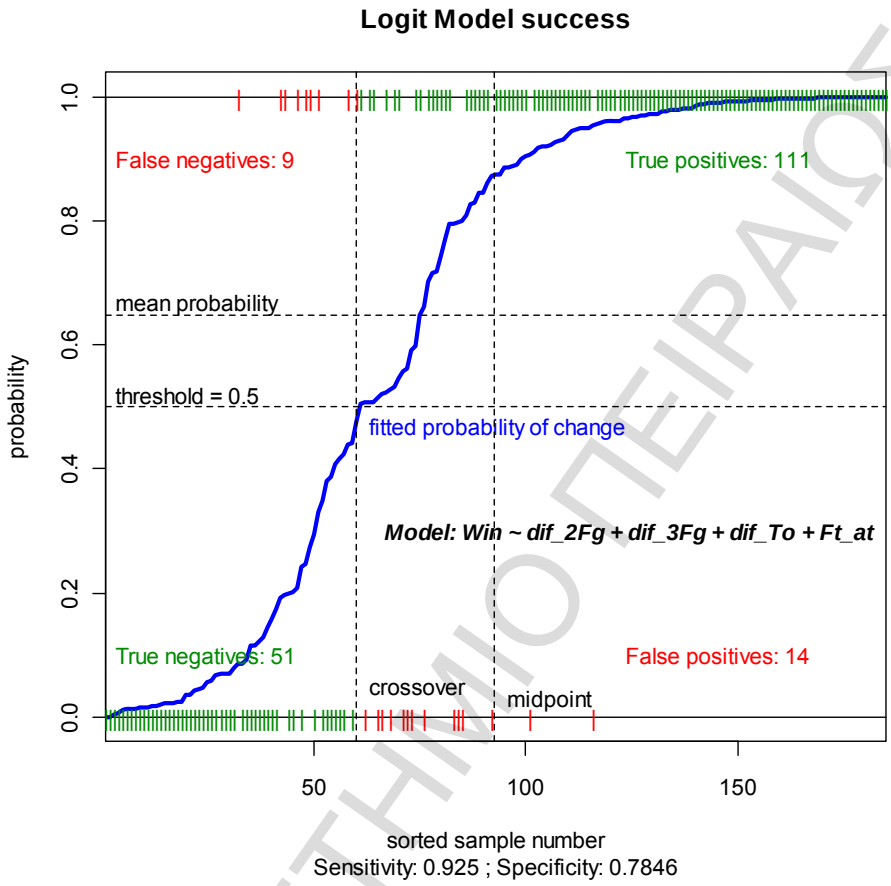
Πίνακας 5.14 *Μη παραμετρικοί συντελεστές*

Rank Discrimination Indexes	
Dxy	0.905
gamma	0.906
tau-a	0.415

Το Somers' D και το gamma έχουν πολύ υψηλές τιμές και είναι κοντά στο 1, το tau-a έχει μια πιο μετριοπαθής τιμή.

9. Μια απεικόνιση της καλής προσαρμογής του μοντέλου, θα δοθεί με το ακόλουθο διάγραμμα.

Σχήμα 5.4 *Διάγραμμα επιτυχίας του λογιστικού μοντέλου*



Ταξινομώντας της προσαρμοσμένες τιμές, προκύπτει η λογιστική καμπύλη με τη χαρακτηριστική της μορφή και τα άκρα της σχεδόν αγγίζουν τα σημεία 0 και 1. Αυτό από μόνο του μας δείχνει ότι το μοντέλο αποδίδει καλά, οι παρατηρήσεις όπου Win=0, είναι όλες συγκεντρωμένες κάτω αριστερά, ενώ οι αντίστοιχες όπου Win=1, είναι όλες συγκεντρωμένες πάνω δεξιά. Επίσης, οι προσαρμοσμένες τιμές δεν είναι συγκεντρωμένες στη μέση πιθανότητα, κάτι που θα έδειχνε κακή προσαρμογή.

Τα κόκκινα σημεία αντιπροσωπεύουν τις περιπτώσεις που το μοντέλο απέτυχε να προβλέψει σωστά το αποτέλεσμα. Παρατηρούμε ότι υπάρχουν αρκετά "ψευδώς θετικά", κάτι που οδηγεί σε χαμηλή ειδικότητα (specificity), ενώ τα λιγότερα "ψευδώς αρνητικά" οδηγούν σε υψηλή ευαισθησία (sensitivity). Στο διάγραμμα απεικονίζεται το 0.5 cutoff σημείο, πάνω στην οποία στηρίχθηκε η κατηγοριοποίηση.

5.4.2 Έλεγχοι για αριθμητικά προβλήματα των συντελεστών του μοντέλου

α. Έλεγχοι Πολυσυγγραμμικότητας

Θα εξετάσουμε εάν οι μεταβλητές συνδέονται με τέτοιο τρόπο μεταξύ τους, ώστε να προκαλούν πρόβλημα στις εκτιμήσεις του μοντέλου.

Αρχικά, θα παρατεθούν οι συντελεστές συσχέτισης του Pearson για τις ερμηνευτικές μεταβλητές.

Πίνακας 5.15 *Συσχετίσεις των ερμηνευτικών μεταβλητών*

Pearson Correlations

	dif_2Fg%	dif_3Fg%	dif_To	Ft_at
dif_2Fg%	1	-0.071	-0.091	-0.001
dif_3Fg%	-0.071	1	0.104	-0.115
dif_To	-0.091	0.104	1	-0.033
Ft_at	-0.001	-0.115	-0.033	1

Οι συντελεστές συσχέτισης ανάμεσα στις ανεξάρτητες μεταβλητές, δείχνουν ότι αυτές είτε δεν έχουν καμία σχέση μεταξύ τους, είτε αυτή είναι αμυδρή, κάτι που είναι και το ιδανικό.

Στον επόμενο πίνακα, βρίσκονται τα αποτελέσματα από τους ελέγχους.

Πίνακας 5.16 *Μέτρα πολυσυγγραμμικότητας*

```
vif(glm)
 dif_2Fg  dif_3Fg   dif_To    Ft_at
1.766412 2.014243 1.865728 1.289190

Condition
Index   Variance Decomposition Proportions
         intercept dif_2Fg dif_3Fg dif_To Ft_at
1  1.000 0.022      0.019   0.007   0.028  0.022
2  1.361 0.001      0.202   0.399   0.226  0.001
3  1.524 0.001      0.777   0.134   0.140  0.001
4  1.580 0.004      0.000   0.438   0.606  0.006
5  6.176 0.971      0.001   0.022   0.000  0.969
```

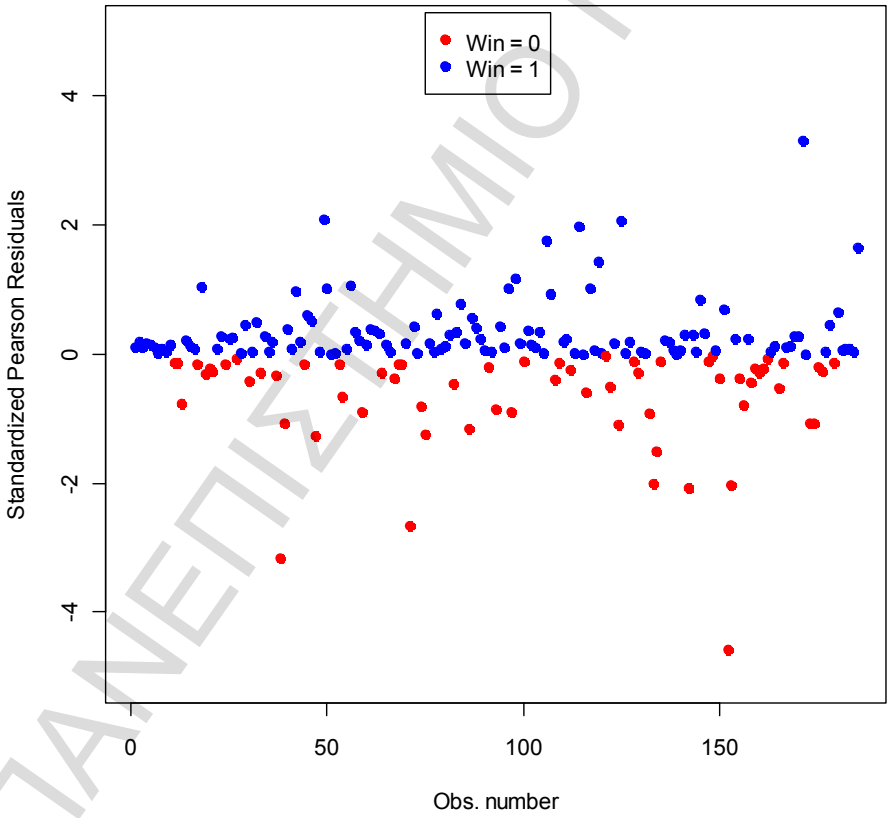
Οι δείκτες vif είναι αρκετά μικροί και < 5. Ο μεγαλύτερος εξ αυτών είναι της dif_3Fg, με τιμή 2.014 και εκφράζει ότι το τυπικό σφάλμα του συντελεστή αυτής της μεταβλητής είναι $1.4\,(\sqrt{2.014})$ φορές τόσο μεγάλο, όσο θα ήταν αν αυτή η μεταβλητή ήταν ασυσχέτιστη με τις υπόλοιπες.

Επίσης, μικρός είναι και ο Condition Index (6.176) < 10, επιβεβαιώνοντας ότι δεν εντοπίζεται πρόβλημα πολυσυγγραμμικότητας στο μοντέλο. Ακόμη, κανένας condition index δε συνδέεται με 2 ή περισσότερες μεταβλητές με variance decomposition proportions ≥ 0.5.

β. Έλεγχοι Ακραίων Τιμών

Θα προσπαθήσουμε να βρούμε έκτροπες τιμές και τις περιπτώσεις που έχουν μεγάλη επιρροή, στη διαμόρφωση της εκτίμησης των συντελεστών του μοντέλου. Ένα χρήσιμο διάγραμμα είναι το παρακάτω, με τα τυποποιημένα κατάλοιπα του Pearson.

Σχήμα 5.5 *Διάγραμμα διασποράς τυποποιημένων καταλοίπων*

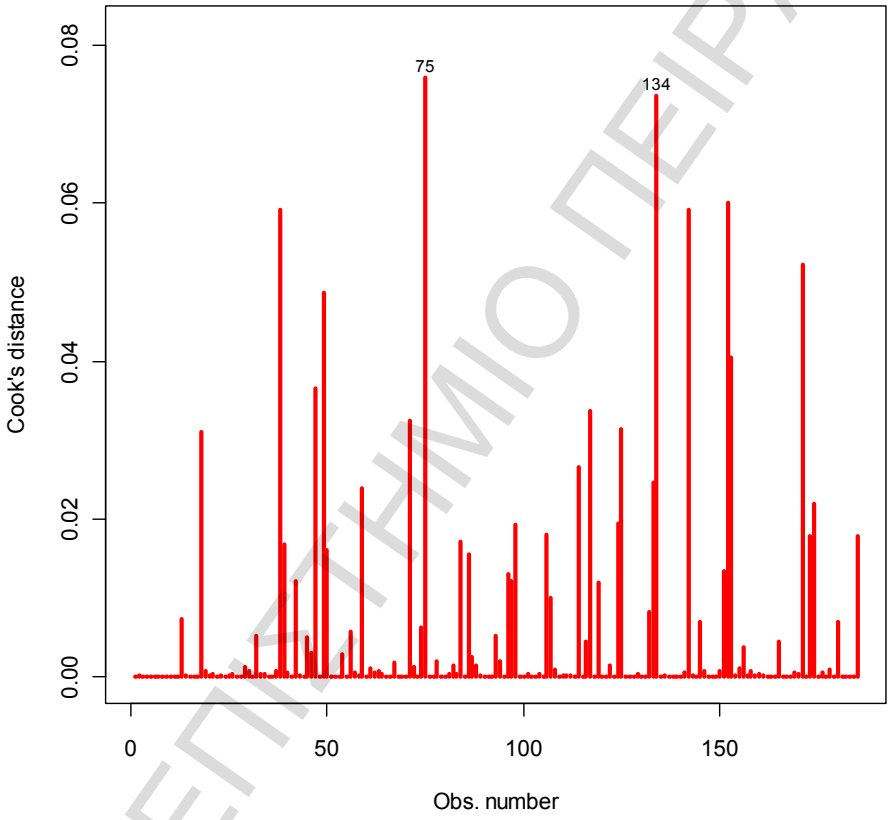


Τα αρνητικά (θετικά) τυποποιημένα κατάλοιπα του Pearson, αντιστοιχούν στις παρατηρήσεις όπου το τελικό αποτέλεσμα ήταν η ήττα (νίκη) και μεγάλα μεγέθη αυτών, δείχνουν ότι οι διαφορές των πραγματικών αποτελεσμάτων με τις αντίστοιχες πιθανότητες του μοντέλου είναι μεγάλες. Έτσι, αναμένουμε ότι σε αυτές τις περιπτώσεις, το μοντέλο θα προβλέπει εσφαλμένα το αποτέλεσμα.

Ανιχνεύονται «έκτροπες τιμές», με τυποποιημένα κατάλοιπα που υπερβαίνουν τον πρακτικό κανόνα του ±3 (βλ. 4.10.2). Όμως, δε μπορούμε να παραβλέψουμε και τα κατάλοιπα που είναι πάνω από 2.5 κατ' απόλυτη τιμή, καθώς απέχουν πολύ από όλα τα υπόλοιπα.

Με σκοπό να εξετάσουμε τις παρατηρήσεις υψηλής επιρροής, θα κάνουμε ένα διάγραμμα με τις αποστάσεις του Cook.

Σχήμα 5.6 *Γράφημα απόστασης του Cook*



Η μέση απόσταση του Cook είναι πολύ μικρή (0.0058), όμως παρατηρείται μεγάλη διαφορά αυτής της τιμής με τη μέγιστη τιμή της. Αν χρησιμοποιήσουμε τον αυστηρό κανόνα 4/n (Bollen & Jackman, 1990), μια παρατήρηση με μεγάλη επιρροή θα θεωρείται αν ξεπεράσει το όριο (4/185=) 0.0216. Ανιχνεύονται 17 τιμές, αλλά κρίνεται σκόπιμο να εξετάσουμε μόνο τις περιπτώσεις 75 και 134, που έχουν σημαντικά τη μεγαλύτερη επιρροή από όλες τις υπόλοιπες.

Παρακάτω, παρατίθενται οι περιπτώσεις των |standardized residuals| ≥ 2.5 και αυτών με τις δύο μεγαλύτερες αποστάσεις του Cook.

Πίνακας 5.17 *Έκτροπες παρατηρήσεις*

Outliers				
Case	Observed	Fitted values	Predicted	Stand/zed Residuals
171	1	0.086	0	3.303
152	0	0.954	1	-4.577
38	0	0.907	1	-3.168
71	0	0.874	1	-2.660

Η περίπτωση 152, έχει το υψηλότερο τυποποιημένο κατάλοιπο κατ' απόλυτη τιμή με 4.6 και ακολουθεί η περίπτωση 171 με 3.3.

Πίνακας 5.18 *Υψηλής επιρροής παρατηρήσεις*

Influential obs.					
Case	Observed	Fitted values	Predicted	Cook's distance	Hat values
75	0	0.558	1	0.076	0.195
134	0	0.662	1	0.074	0.139

Η μέγιστη απόσταση του Cook είναι 0.076. Οι τιμές της μόχλευσης σε αυτές τις περιπτώσεις, παίρνουν επίσης τις μέγιστες τιμές τους.

Το μοντέλο σε όλες τις παραπάνω περιπτώσεις, απέτυχε να προβλέψει ορθά το τελικό αποτέλεσμα. Επίσης, σε καμία από αυτές δεν παρατηρείται παράταση του αγώνα.

Θα είχε ενδιαφέρον να δούμε, γιατί το μοντέλο απέτυχε να προβλέψει σωστά τουλάχιστον 2 από τα παιχνίδια (περίπτωση με τη μεγαλύτερη επιρροή και εκείνη με το μεγαλύτερο κατάλοιπο).

Ύστερα από εξέταση της 75^{ης} περίπτωσης διαπιστώθηκε ότι, η ομάδα είχε αρκετά μεγαλύτερα ποσοστά ευστοχίας στα σουτ 2 (12.9%) και 3 (22%) πόντων από τον αντίπαλο, θα εκτιμούσαμε ότι η ομάδα θα φτάσει στη νίκη, όμως ο κατά πολύ μεγαλύτερος αριθμών λαθών (17) από την αντίπαλη ομάδα την οδήγησε στην ήττα με 3 πόντους διαφορά. Στο 152^{ο} παιχνίδι, με μηδενική διαφορά λαθών, μεγαλύτερο ποσοστό στα τρίποντα (11%) και αρκετές ελ. βολές (28), ο γηπεδούχος δεν κατάφερε να νικήσει. Είχε 17 ριμπάουντ λιγότερα από τον αντίπαλο, μεταβλητή που δεν περιέχει το μοντέλο μας.

Στη συνέχεια θα εξαιρέσουμε αυτές τις 6 «προβληματικές» τιμές και θα ελέγξουμε την ακρίβεια ταξινόμησης που θα δώσει το νέο μοντέλο.

Πίνακας 5.19 *Πίνακας ταξινόμησης αναθεωρημένου μοντέλου*

Classification table					
			Predicted		
	Win		0	1	**% Correct**
Observed	0	60	51	9	**85.00**
	1	119	6	113	**94.96**
Overall	185				**91.62**

Τα αποτελέσματα είναι μάλλον εντυπωσιακά. Όπως παρατηρούμε το ποσοστό σωστής κατηγοριοποίησης είναι μεγαλύτερο από 2% και συγκεκριμένα 4.05%, από το μοντέλο με όλα τα δεδομένα. Ωστόσο, για τις ανάγκες αυτής της μελέτης, θα συμπεριλάβουμε όλες τις παρατηρήσεις στην ανάλυσή μας. Το αναθεωρημένο μοντέλο παρατίθεται στο Παράρτημα.

5.4.3 Ανάλυση cross-validation

Χωρίσαμε τυχαία τις 185 παρατηρήσεις, βάσει γεννήτριας τυχαίων αριθμών (με τιμή seed: 12345), στα 2/3 (65%) που αντιστοιχεί στο δείγμα εκπαίδευσης και στο υπόλοιπο 1/3 (35%) που αντιστοιχεί στο δείγμα επικύρωσης. Έτσι, θα εξετάσουμε πως ένα logit μοντέλο με τις ίδιες ερμηνευτικές μεταβλητές προσαρμόζεται στο 65% των δεδομένων και αυτό θα εξετασθεί πόσο καλά μπορεί να προβλέψει το υπόλοιπο 35%. Τα αποτελέσματα της ανάλυσης έχουν ως εξής:

Πίνακας 5.20 *Ανάλυση διασταυρούμενης ισχύος*

Logit Model	***Full data***	***Split***
Model Chi-Square	141.993 $p < 0.0001$	88.200 $p < 0.0001$
Nagelkerke R-square	0.737	0.721
Accuracy Rate for Training Sample	87.57%	85.47%
Accuracy Rate for Validation Sample	–	88.24%
Significant Coefficients ($p < 0.05$)	ALL	ALL

Παρατηρούμε ότι, επιτυγχάνεται στατιστική σημαντικότητα για την τιμή του χ^2 (p-value < 0.001) και όλοι οι συντελεστές του μοντέλου είναι επίσης σημαντικοί.

Το κριτήριο που υποστηρίζει την ακρίβεια κατηγοριοποίησης του μοντέλου είναι ότι, η ακρίβεια για το δείγμα επικύρωσης δεν πρέπει να είναι μικρότερη πάνω από 10% (βλ. 4.12), από την αντίστοιχη του δείγματος εκπαίδευσης (85.47%). Αυτό όμως, όχι απλώς δε συμβαίνει, αλλά το μοντέλο προβλέπει καλύτερα τα δεδομένα επικύρωσης με 88.24%.

Επίσης, το Nagelkerke R^2 και η ακρίβεια ταξινόμησης είναι σε σχεδόν τα ίδια επίπεδα, με το μοντέλο που προσαρμόστηκε σε όλο το σετ δεδομένων. Συνεπώς, επιβεβαιώνεται η εγκυρότητα του μοντέλου. Η συνολική στατιστική ισχύς του, δεν αφήνει περιθώρια αμφισβήτησης πως δε μπορεί να γενικευτεί σε έναν μεγαλύτερο πληθυσμό αγώνων μπάσκετ.

5.5 Προσαρμογή Probit μοντέλου

Θα προσαρμόσουμε ένα μοντέλο probit με τις ίδιες μεταβλητές που χρησιμοποιήσαμε στο logit μοντέλο. Τα αποτελέσματα της προσαρμογής, βρίσκονται στον παρακάτω πίνακα.

Πίνακας 5.21 *Αποτελέσματα Probit μοντέλου*

```
> glm<-glm(Win ~ dif_2Fg + dif_3Fg + dif_To + Ft_at,
family=binomial(link=probit))
> anova(glm,test="Chisq")

Analysis of Deviance Table

Model: binomial, link: probit

Response: Win

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                      184     239.86
dif_2Fg  1   38.871       183     200.99 4.527e-10 ***
dif_3Fg  1   37.452       182     163.54 9.367e-10 ***
dif_To   1   40.200       181     123.34 2.293e-10 ***
Ft_at    1   26.590       180      96.75 2.516e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary(glm)

Call:
glm(formula = Win ~ dif_2Fg + dif_3Fg + dif_To + Ft_at, family =
binomial(link = probit))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.50966  -0.18584   0.03984   0.33245   2.20462

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.72386    0.46187  -3.732  0.00019 ***
dif_2Fg      0.11354    0.01934   5.871 4.33e-09 ***
dif_3Fg      0.08404    0.01346   6.243 4.30e-10 ***
dif_To      -0.18185    0.03391  -5.363 8.19e-08 ***
Ft_at        0.10187    0.02206   4.618 3.87e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 239.86  on 184  degrees of freedom
Residual deviance:  96.75  on 180  degrees of freedom
AIC: 106.75

Number of Fisher Scoring iterations: 8
```

Οι εισαχθείσες μεταβλητές και όλες οι παράμετροι είναι σημαντικές σε ε.σ. 5%, κάτι που αναμέναμε. Παρατηρούμε ότι, οι συντελεστές έχουν διαφορετική τιμή για το probit μοντέλο και αυτό διότι βρίσκονται σε διαφορετική κλίμακα από τη logit.

Το μοντέλο είναι το ακόλουθο:

$$probit = -1.72 + 0.11 \times dif\_2Fg + 0.08 \times dif\_3Fg - 0.18 \times dif\_To + 0.10 \times Ft\_at$$

και η εκτιμώμενη πιθανότητα για νίκη της γηπεδούχου ομάδας είναι:

$$\hat{p}_i = \Phi(-1.72 + 0.11 \times dif\_2Fg + 0.08 \times dif\_3Fg - 0.18 \times dif\_To + 0.10 \times Ft\_at)$$

Οι θετικοί συντελεστές, δηλώνουν ότι μία αύξηση στην ερμηνευτική μεταβλητή οδηγεί σε αύξηση της εκτιμώμενης πιθανότητας (και μείωση για αρνητικό συντελεστή). Για την ακρίβεια, μια μοναδιαία αύξηση της dif_To, προκαλεί μείωση του Z-score της πιθανότητας νίκης κατά -0.18 και παρόμοια για τις υπόλοιπες.

Η αύξηση της πιθανότητας που αποδίδεται από μια μοναδιαία αύξηση μιας ερμηνευτικής μεταβλητής, εξαρτάται τόσο από τις τιμές των άλλων ερμηνευτικών, όσο και από τις αρχικές τους τιμές. Για μια λιγότερο περίπλοκη ερμηνεία των συντελεστών, θα χρησιμοποιηθεί ένα παράδειγμα. Αν θεωρήσουμε ότι και οι 3

μεταβλητές που εκφράζουν διαφορές παίρνουν την τιμή 0 και η Ft_at έχει αρχική τιμή 20 (όση και η διάμεσός της), τότε κάποιες εκτιμώμενες πιθανότητες είναι:

$$\Phi(-1.72+0.10\times 20)=\Phi(0.28)=0.6102612$$
$$\Phi(-1.72+0.10\times 21)=\Phi(0.38)=0.6480273$$
$$\Phi(-1.72+0.10\times 22)=\Phi(0.48)=0.6843863$$

Παρατηρούμε ότι, οι πιθανότητες δε μεταβάλλονται κατά έναν κοινό παράγοντα. Η αύξηση της Ft_at από τις 20 προσπάθειες στις 21, έχει διαφορετική επίδραση στην πιθανότητα, σε σχέση με την αύξηση από τις 21 στις 22 προσπάθειες.

5.6 Σύγκριση αποτελεσμάτων Logit και Probit μοντέλων

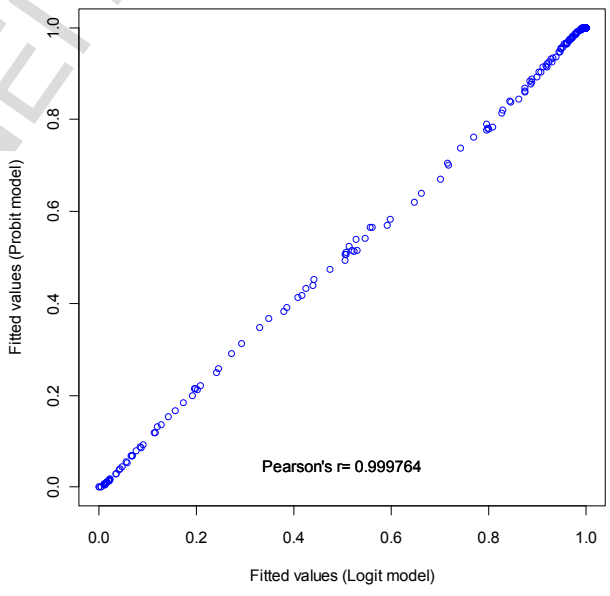
Για να δείξουμε ότι τα αποτελέσματα των logit και probit μοντέλων είναι συγκρίσιμα, θα αντικαταστήσουμε τις τιμές του παραδείγματος στην εξίσωση του logit μοντέλου που είχαμε υπολογίσει. Έτσι για μηδενικές διαφορές και Ft_at=20:

$$\frac{\hat{p}_i}{1-\hat{p}_i}=e^{-3.08}\times e^{0.18\times 20}\Leftrightarrow\frac{\hat{p}_i}{1-\hat{p}_i}=1.682\Leftrightarrow\hat{p}_i=\frac{1.682}{1+1.682}\Leftrightarrow\hat{p}_i=0.6271439$$

Η διαφορά στις δύο πιθανότητες είναι σχετικά μικρή (0.017).

Θα είχε μάλλον αρκετό ενδιαφέρον να συγκρίνουμε τις προσαρμοσμένες τιμές από τα δύο μοντέλα:

Σχήμα 5.7 *Συγκριτικό γράφημα προσαρμοσμένων τιμών logit και probit μοντέλων*



Όπως παρατηρούμε, οι προσαρμοσμένες τιμές και των δύο μοντέλων βρίσκονται πάνω σε μία τέλεια ευθεία και ο συντελεστής συσχέτισης του Pearson είναι σχεδόν 1. Οι διαφορές των εκτιμώμενων πιθανοτήτων είναι μηδαμινές.

Για να πραγματοποιηθεί η σύγκριση των μοντέλων, στον παρακάτω πίνακα συνοψίζονται τα κυριότερα μέτρα για την αξιολόγησή τους.

Πίνακας 5.22 *Συγκριτικός πίνακας προσαρμογής του Logit & Probit μοντέλου*

	AIC	**BIC**	**-2LL**	**Correct %**	**AUC**	**r2CU**	**BS**
Logit	107.870	123.972	97.87	87.57	0.9526	0.738	0.0866
Probit	106.750	122.852	96.75	88.11	0.9528	0.741	0.0865

Το probit μοντέλο βάσει των παραπάνω δεικτών αξιολόγησης έχει ελαφρώς καλύτερη προσαρμογή από το αντίστοιχο logit.

Η απόκλιση για το μοντέλο probit είναι ελαφρώς μικρότερη, πράγμα που μας οδηγεί σε καλύτερη προσαρμογή του μοντέλου. Τα κριτήρια AIC και BIC δίνουν μικρότερες και επομένως «καλύτερες» τιμές για τα probit μοντέλα. Επίσης ο δείκτης AUC και το Nagelkerke R^2 είναι ελαφρώς μεγαλύτερα και το μοντέλο κατηγοριοποιεί σωστά μία περίπτωση παραπάνω.

Οι διαφορές στις 2 συναρτήσεις σύνδεσης δεν είναι μεγάλες, όμως το probit μοντέλο εμφάνισε ελαφρώς καλύτερη προσαρμογή στα δεδομένα, από το αντίστοιχο logit, στην πρόβλεψη του τελικού αποτελέσματος. Έτσι επιβεβαιώνουμε τις μελέτες (Bouler & Stekler, 1999) που έχουν γίνει, ότι τα probit μοντέλα προσαρμόζονται καλύτερα σε δεδομένα του μπάσκετ.

ΚΕΦΑΛΑΙΟ 6

Τελικά συμπεράσματα

Στα πλαίσια αυτής της μελέτης, επεξεργαστήκαμε 185 παιχνίδια από το κορυφαίο Ευρωπαϊκό πρωτάθλημα μπάσκετ και μελετήσαμε τα κύρια στατιστικά στοιχεία των γηπεδούχων όπως πόντους, ευστοχία, λάθη κλπ, αλλά και έμμεσα στοιχεία του αντιπάλου παίρνοντας τις διαφορές των ομάδων στα ριμπάουν, τις ασίστ κλπ και όλα αυτά σε ένα σύνολο 32 μεταβλητών.

Όπως είδαμε στο 2^{ο} Κεφάλαιο, οι γηπεδούχοι νίκησαν στο 64.9% και έχασαν στο 35.1% των παιχνιδιών. Οι ομάδες που επέτυχαν τουλάχιστον 90 πόντους ήταν και νικήτριες, ενώ από όσες επέτυχαν 80-89 πόντους, μόνο το 9.7% από αυτές ηττήθηκαν. Επίσης, τα παιχνίδια στο ημίχρονο και τη λήξη του αγώνα είχαν μια μέση διαφορά 3 και 5 πόντων αντίστοιχα, υπέρ του γηπεδούχου, ενώ παράταση, σημειώθηκε στο 4.3% των παιχνιδιών και οι γηπεδούχοι κέρδισαν στο 75% αυτών.

Βρήκαμε ότι, οι τελικοί πόντοι που επιτυγχάνονται έχουν ισχυρή σχέση με τις Ασίστ και το ποσοστό ευστοχίας σουτ 3 πόντων. Το τελικό αποτέλεσμα και η διαφορά του τελικού σκορ, έδειξαν να σχετίζονται σημαντικά με ότι έχει να κάνει με πόντους και σκορ, τη διαφορά ευστοχίας στα σουτ 2 και 3 πόντων, τη διαφορά στις ασίστ, στα κλεψίματα, στα ριμπάουντ και στα λάθη.

Είδαμε επίσης ότι, ανατροπές του σκορ είναι δύσκολο να επιτευχθούν. Οι γηπεδούχοι που είχαν το προβάδισμα στην 1^{η} περίοδο, κέρδισαν στο 82.4% αυτών, ενώ αντίστοιχα ήταν τα ποσοστά για το ημίχρονο και την 3^{η} περίοδο. Αντίθετα, όσοι γηπεδούχοι ήταν πίσω στο σκορ σε κάθε μία από τις 3 πρώτες περιόδους, κέρδισαν μόλις στο 38.2%, 30.4% και 25.9% αυτών αντίστοιχα, ενώ τα αντίστοιχα ποσοστά για τους φιλοξενούμενους ήταν κοντά στο 18%. Στο ημίχρονο το 30.3% των γηπεδούχων ήταν πίσω στο σκορ, ενώ το 65.9% ήταν μπροστά. Να σημειωθεί ότι στην περίπτωση ισοπαλίας στο τέλος της 3^{ης} περιόδου, οι γηπεδούχοι αναδείχθηκαν νικητές σε

ποσοστό 83.3%. Η γηπεδούχος (φιλοξενούμενη) ομάδα που προηγούνταν και στις 3 πρώτες περιόδους κέρδισε το 87.8% (91.2%) των παιχνιδιών, ενώ το γεγονός αυτό συνέβη στο 40% (18.4%) του συνόλου των παιχνιδιών.

Επιβεβαιώσαμε απόλυτα την έρευνα (Cooper et al, 1992), ότι η ομάδα που προηγούνταν στο σκορ στο ημίχρονο κέρδισε κατά προσέγγιση το 75% (77.5% στο δείγμα μας) των παιχνιδιών, ενώ όταν αυτό συνέβαινε μετά το πέρας της $3^{ης}$ περιόδου το ποσοστό ήταν περίπου 80% (79.8% στο δείγμα μας). Μετά το πέρας της $1^{ης}$ περιόδου το εμπειρικό ποσοστό ήταν 74.1%, αρκετά υψηλό θα λέγαμε.

Όταν οι γηπεδούχοι ανήκαν σε υψηλότερο γκρουπ δυναμικότητας, κέρδισαν το 76.1% των αγώνων, ενώ όσοι ήταν σε μικρότερο το ποσοστό ήταν 53.9%. Όσοι ανήκαν στο ίδιο γκρουπ είχαν μοιρασμένα ποσοστά 50%. Οι γηπεδούχοι που είχαν περισσότερες ασίστ από τους αντιπάλους, κέρδισαν το 80.7% των παιχνιδιών, περισσότερα ριμπάουντ το 77.7%, περισσότερα κλεψίματα το 82.1% και λιγότερα λάθη το 76.9%. Ενδιαφέρον παρουσιάζει ότι, όταν είχαν και οι δυο ομάδες ίσα κλεψίματα το ποσοστό για νίκη ήταν στο 50-50, ενώ όταν είχαν ίσα ριμπάουντ οι γηπεδούχοι ηττήθηκαν σε 2 παιχνίδια παραπάνω από τους φιλοξενούμενους. Οι γηπεδούχοι είναι φανερό ότι υπερτερούν στα στατιστικά σε όλους τους τομείς και ο παράγοντας έδρα, φαίνεται να παίζει σπουδαίο ρόλο στην εξέλιξη ενός αγώνα.

Στο Κεφάλαιο 3 είδαμε ότι, μέσω στατιστικών κριτηρίων (όπως χ^2 και KS) και γραφημάτων, η κατανομή Γάμμα με παραμέτρους Gamma(54.9, 0.71) προσαρμόζεται καλύτερα στους τελικούς πόντους, ενώ για τη διαφορά του τελικού σκορ η καταλληλότερη κατανομή είναι η Λογιστική με παραμέτρους Logistic(4.84, 7.72). Για την απόλυτη διαφορά πόντων, προσαρμόστηκε και πάλι η Γάμμα κατανομή, ενώ για τα σκορ των τριών πρώτων περιόδων, καταλήξαμε ότι η Κανονική κατανομή είναι μια καλή προσέγγιση.

Στο Κεφάλαιο 5, εκτιμήσαμε τη σημαντικότητα των μεταβλητών στην πρόβλεψη του αποτελέσματος, μέσω logit μοντέλων. Καθοριστικότεροι παράγοντες, εμφανίζονται όσοι έχουν άμεση σχέση με το σκοράρισμα, όπως οι τελικοί πόντοι, οι διαφορές του σκορ ανά περίοδο και οι διαφορές στην ευστοχία στα σουτ 2 πόντων κατά κύριο λόγο και κατά δεύτερο στα σουτ 3 πόντων. Έπειτα, οι πιο σημαντικοί παράγοντες εμφανίζονται οι διαφορές στις ασίστ, στα λάθη και στα ριμπάουντ.

Τόσο από την περιγραφική ανάλυση, όσο και από την προσαρμογή λογιστικών μοντέλων, η ευστοχία στις βολές, τα επιθετικά ριμπάουντ, τα κοψίματα, τα φάουλ και οι προσπάθειες στα τρίποντα (8 μεταβλητές) δε φάνηκαν να επηρεάζουν

σημαντικά το τελικό αποτέλεσμα. Ένα ενδιαφέρον στοιχείο είναι ότι, αν και η ευστοχία στις ελ. βολές έχει μηδενική συνεισφορά, οι προσπάθειες που έγιναν για αυτές έχουν σημαντική.

Στη συνέχεια, μετά από επαναληπτικές διαδικασίες stepwise, καταλήξαμε σε ένα λογιστικό μοντέλο με 4 ερμηνευτικές μεταβλητές, όπου καμία μεταβλητή που να εκφράζει τους πόντους και το σκορ δεν εισήχθη μέσα σε αυτό:

$$\log it(\hat{p}) = -3.08 + 0.20 \times dif\_2Fg + 0.15 \times dif\_3Fg - 0.32 \times dif\_To + 0.18 \times Ft\_at$$

Το μοντέλο πέρασε όλους τους ελέγχους με επιτυχία και αποδείχθηκε ιδιαίτερα ευσταθές, με υψηλή προβλεπτική ικανότητα, αλλά και αρκετά αξιόπιστο, διασφαλίζοντας ότι οι συντελεστές του μοντέλου εκτιμήθηκαν σωστά. Οι διαφορές στην ευστοχία των αντίπαλων ομάδων στα σουτ διπόντων και τριπόντων, μαζί με τις διαφορές στα λάθη των ομάδων και σε συνδυασμό με τον αριθμό των προσπαθειών για ελεύθερες βολές, μας έκαναν να προβλέψουμε σωστά την έκβαση των αγώνων στο 87.57% των περιπτώσεων, ενώ μετά από ανίχνευση και εξαίρεση κάποιων «προβληματικών» παρατηρήσεων που επηρέαζαν τα αποτελέσματα, η ακρίβεια αυξήθηκε στο 91.62%. Επίσης, οι τιμές των ελέγχων καλής προσαρμογής ήταν ιδιαίτερα υψηλές (AUC=0.953, Nagelkerke R^2=0.876 κλπ) και με τη μέθοδο cross-validation, διασφαλίσαμε ότι το μοντέλο μπορεί να γενικευτεί σε ένα ευρύτερο σετ δεδομένων του μπάσκετ.

Η μεταβλητή που αναπαριστά τα λάθη των ομάδων, έχει τη μεγαλύτερη επίδραση στη σχετική πιθανότητα νίκης του μοντέλου, έτσι αν η γηπεδούχος ομάδα κάνει ένα λάθος περισσότερο, μειώνεται η (σχετική) πιθανότητα να νικήσει (έναντι του να μη νικήσει) κατά 27% και αν κάνει ένα λάθος λιγότερο αυτή αυξάνεται κατά 37%. Μια αύξηση στη διαφορά διπόντων κατά 1% σε σχέση με τον φιλοξενούμενο, αυξάνει τη σχετική πιθανότητα να νικήσει κατά περίπου 22%, ενώ το αντίστοιχο για τα τρίποντα είναι 16%. Μια επιπλέον προσπάθεια στις ελεύθερες βολές, αυξάνει τη σχετική πιθανότητα να νικήσει κατά 20%, ενώ μια προσπάθεια λιγότερη μειώνει αυτήν την πιθανότητα κατά 17%.

Τέλος, το probit μοντέλο που εφαρμόστηκε είχε ελαφρώς καλύτερη προσαρμογή από το αντίστοιχο logit, επιβεβαιώνοντας τις έρευνες (Bouler & Stekler, 1999), εντούτοις η ερμηνεία των αποτελεσμάτων του είναι λιγότερο διαισθητική.

Παράρτημα

Τα γκρουπ δυναμικότητας στους ομίλους για την περίοδο 2010-11:

Πίνακας Π.1 *Πίνακας ομάδων που απαρτίζουν τα γκρουπ δυναμικότητας*

1ο γκρουπ: ΤΣΣΚΑ Μόσχας, Μπαρτσελόνα, Ολυμπιακός, Κάχα Λαμποράλ
2ο γκρουπ: Σιένα, Παναθηναϊκός, Μακάμπι Τελ Αβίβ, Ρεάλ Μαδρίτης
3ο γκρουπ: Παρτιζάν Βελιγραδίου, Βαλένθια, Ουνικάχα, Λιέτουβος Ρίτας
4ο γκρουπ: Φενερμπαχτσέ, Πρόκομ, Εφές Πίλσεν, Ρόμα
5ο γκρουπ: Ζαλγκίρις, Τσιμπόνα, Αρμάνι Μιλάνο, Μπάμπεργκ
6ο γκρουπ: Ολίμπια Λιουμπλιάνα, Σολέ, Κίμκι, Σαρλερουά

Ακολουθεί πίνακας με αύξοντα αριθμό για όλα τα παιχνίδια:

Πίνακας Π.2 *Πίνακας με τα παιχνίδια*

Games	
1	BC Khimki vs. Asseco Prokom
2	Zalgiris vs. Partizan mt:s
3	Caja Laboral vs. Maccabi Electra
4	Olympiacos vs. Real Madrid
5	Lottomatica Roma vs. Brose Baskets
6	Unicaja vs. Spirou Charleroi
7	Fenerbahce Ulker vs. Lietuvos Rytas
8	Regal Barcelona vs. Cibona Zagreb
9	Montepaschi Siena vs. Cholet Basket
10	Union Olimpija vs. Efes Pilsen
11	CSKA Moscow vs. Armani Jeans Milano
12	Power E. Valencia vs. Panathinaikos
13	Asseco Prokom vs. Caja Laboral
14	Partizan mt:s vs. BC Khimki
15	Maccabi Electra vs. Zalgiris
16	Brose Baskets vs. Olympiacos
17	Spirou Charleroi vs. Lottomatica Roma
18	Real Madrid vs. Unicaja
19	Lietuvos Rytas vs. Montepaschi Siena
20	Cholet Basket vs. Regal Barcelona
21	Cibona Zagreb vs. Fenerbahce Ulker
22	Efes Pilsen vs. Power E. Valencia
23	Panathinaikos vs. CSKA Moscow
24	Armani Jeans Milano vs. Union Olimpija
25	BC Khimki vs. Caja Laboral
26	Zalgiris vs. Asseco Prokom

27	Partizan mt:s vs. Maccabi Electra
28	Olympiacos vs. Unicaja
29	Brose Baskets vs. Spirou Charleroi
30	Lottomatica Roma vs. Real Madrid
31	Montepaschi Siena vs. Cibona Zagreb
32	Cholet Basket vs. Lietuvos Rytas
33	Regal Barcelona vs. Fenerbahce Ulker
34	Efes Pilsen vs. Armani Jeans Milano
35	Power E. Valencia vs. CSKA Moscow
36	Union Olimpija vs. Panathinaikos
37	Asseco Prokom vs. Partizan mt:s
38	Caja Laboral vs. Zalgiris
39	BC Khimki vs. Maccabi Electra
40	Olympiacos vs. Spirou Charleroi
41	Unicaja vs. Lottomatica Roma
42	Real Madrid vs. Brose Baskets
43	Fenerbahce Ulker vs. Montepaschi Siena
44	Cibona Zagreb vs. Cholet Basket
45	Regal Barcelona vs. Lietuvos Rytas
46	CSKA Moscow vs. Union Olimpija
47	Power E. Valencia vs. Armani Jeans Milano
48	Panathinaikos vs. Efes Pilsen
49	Partizan mt:s vs. Caja Laboral
50	Zalgiris vs. BC Khimki
51	Maccabi Electra vs. Asseco Prokom
52	Spirou Charleroi vs. Real Madrid
53	Lottomatica Roma vs. Olympiacos
54	Brose Baskets vs. Unicaja
55	Lietuvos Rytas vs. Cibona Zagreb
56	Cholet Basket vs. Fenerbahce Ulker
57	Montepaschi Siena vs. Regal Barcelona
58	Union Olimpija vs. Power E. Valencia
59	Armani Jeans Milano vs. Panathinaikos
60	Efes Pilsen vs. CSKA Moscow
61	Asseco Prokom vs. BC Khimki
62	Partizan mt:s vs. Zalgiris
63	Maccabi Electra vs. Caja Laboral
64	Brose Baskets vs. Lottomatica Roma
65	Spirou Charleroi vs. Unicaja
66	Real Madrid vs. Olympiacos
67	Lietuvos Rytas vs. Fenerbahce Ulker
68	Cholet Basket vs. Montepaschi Siena
69	Cibona Zagreb vs. Regal Barcelona
70	Armani Jeans Milano vs. CSKA Moscow
71	Panathinaikos vs. Power E. Valencia
72	Efes Pilsen vs. Union Olimpija
73	BC Khimki vs. Partizan mt:s
74	Caja Laboral vs. Asseco Prokom
75	Zalgiris vs. Maccabi Electra
76	Olympiacos vs. Brose Baskets
77	Lottomatica Roma vs. Spirou Charleroi
78	Unicaja vs. Real Madrid
79	Fenerbahce Ulker vs. Cibona Zagreb
80	Montepaschi Siena vs. Lietuvos Rytas
81	Regal Barcelona vs. Cholet Basket

82	CSKA Moscow vs. Panathinaikos
83	Power E. Valencia vs. Efes Pilsen
84	Union Olimpija vs. Armani Jeans Milano
85	Maccabi Electra vs. Partizan mt:s
86	Asseco Prokom vs. Zalgiris
87	Caja Laboral vs. BC Khimki
88	Spirou Charleroi vs. Brose Baskets
89	Unicaja vs. Olympiacos
90	Real Madrid vs. Lottomatica Roma
91	Cibona Zagreb vs. Montepaschi Siena
92	Lietuvos Rytas vs. Cholet Basket
93	Fenerbahce Ulker vs. Regal Barcelona
94	CSKA Moscow vs. Power E. Valencia
95	Armani Jeans Milano vs. Efes Pilsen
96	Panathinaikos vs. Union Olimpija
97	Zalgiris vs. Caja Laboral
98	Partizan mt:s vs. Asseco Prokom
99	Maccabi Electra vs. BC Khimki
100	Spirou Charleroi vs. Olympiacos
101	Lottomatica Roma vs. Unicaja
102	Brose Baskets vs. Real Madrid
103	Cholet Basket vs. Cibona Zagreb
104	Lietuvos Rytas vs. Regal Barcelona
105	Montepaschi Siena vs. Fenerbahce Ulker
106	Efes Pilsen vs. Panathinaikos
107	Union Olimpija vs. CSKA Moscow
108	Armani Jeans Milano vs. Power E. Valencia
109	Asseco Prokom vs. Maccabi Electra
110	BC Khimki vs. Zalgiris
111	Caja Laboral vs. Partizan mt:s
112	Unicaja vs. Brose Baskets
113	Real Madrid vs. Spirou Charleroi
114	Olympiacos vs. Lottomatica Roma
115	Fenerbahce Ulker vs. Cholet Basket
116	Cibona Zagreb vs. Lietuvos Rytas
117	Regal Barcelona vs. Montepaschi Siena
118	CSKA Moscow vs. Efes Pilsen
119	Power E. Valencia vs. Union Olimpija
120	Panathinaikos vs. Armani Jeans Milano
121	Lietuvos Rytas vs. Panathinaikos
122	Unicaja vs. Caja Laboral
123	Regal Barcelona vs. Maccabi Electra
124	Lottomatica Roma vs. Union Olimpija
125	Efes Pilsen vs. Montepaschi Siena
126	Real Madrid vs. Partizan mt:s
127	Power E. Valencia vs. Zalgiris
128	Olympiacos vs. Fenerbahce Ulker
129	Caja Laboral vs. Lietuvos Rytas
130	Panathinaikos vs. Unicaja
131	Maccabi Electra vs. Lottomatica Roma
132	Union Olimpija vs. Regal Barcelona
133	Partizan mt:s vs. Efes Pilsen
134	Montepaschi Siena vs. Real Madrid
135	Zalgiris vs. Olympiacos
136	Fenerbahce Ulker vs. Power E. Valencia

137	Caja Laboral vs. Panathinaikos
138	Unicaja vs. Lietuvos Rytas
139	Maccabi Electra vs. Union Olimpija
140	Regal Barcelona vs. Lottomatica Roma
141	Real Madrid vs. Efes Pilsen
142	Partizan mt:s vs. Montepaschi Siena
143	Fenerbahce Ulker vs. Zalgiris
144	Olympiacos vs. Power E. Valencia
145	Panathinaikos vs. Caja Laboral
146	Lietuvos Rytas vs. Unicaja
147	Lottomatica Roma vs. Regal Barcelona
148	Union Olimpija vs. Maccabi Electra
149	Montepaschi Siena vs. Partizan mt:s
150	Efes Pilsen vs. Real Madrid
151	Zalgiris vs. Fenerbahce Ulker
152	Power E. Valencia vs. Olympiacos
153	Panathinaikos vs. Lietuvos Rytas
154	Caja Laboral vs. Unicaja
155	Union Olimpija vs. Lottomatica Roma
156	Maccabi Electra vs. Regal Barcelona
157	Montepaschi Siena vs. Efes Pilsen
158	Partizan mt:s vs. Real Madrid
159	Zalgiris vs. Power E. Valencia
160	Fenerbahce Ulker vs. Olympiacos
161	Lietuvos Rytas vs. Caja Laboral
162	Unicaja vs. Panathinaikos
163	Regal Barcelona vs. Union Olimpija
164	Lottomatica Roma vs. Maccabi Electra
165	Efes Pilsen vs. Partizan mt:s
166	Real Madrid vs. Montepaschi Siena
167	Power E. Valencia vs. Fenerbahce Ulker
168	Olympiacos vs. Zalgiris
169	Caja Laboral vs. Maccabi Electra
170	Regal Barcelona vs. Panathinaikos
171	Real Madrid vs. Power E. Valencia
172	Olympiacos vs. Montepaschi Siena
173	Caja Laboral vs. Maccabi Electra
174	Regal Barcelona vs. Panathinaikos
175	Real Madrid vs. Power E. Valencia
176	Olympiacos vs. Montepaschi Siena
177	Maccabi Electra vs. Caja Laboral
178	Panathinaikos vs. Regal Barcelona
179	Power E. Valencia vs. Real Madrid
180	Montepaschi Siena vs. Olympiacos
181	Maccabi Electra vs. Caja Laboral
182	Panathinaikos vs. Regal Barcelona
183	Power E. Valencia vs. Real Madrid
184	Montepaschi Siena vs. Olympiacos
185	Real Madrid vs. Power E. Valencia

Πίνακας Π.3 *Πλήρης πίνακας με τα κύρια στατιστικά περιγραφικά μέτρα για όλες τις συνεχείς μεταβλητές*

Win		*Pts*	*X2Fg*	*X2Fg_at*	*X3Fg*	*X3Fg_at*	*Ft*	*Ft_at*	*Reb_o*	*Reb_d*	*As*	*St*	*To*	*Bl_fv*	*Bl_ag*	*Fl_cm*	*Fl_rv*	*Rkg*	*dif_2Fg*	*dif_3Fg*	*dif_Ft*	*dif_Reb*	*dif_As*	*dif_St*	*dif_To*	*dif_Rkg*	*dif_Q1*	*dif_Q2*	*dif_Q3*	*dif_Pts*
No	N																65													
	Mean	69.08	48.68	37.66	31.13	20.45	75.79	17.78	10.40	21.63	12.54	5.94	14.51	2.71	2.88	21.11	20.40	66.35	-5.04	-5.10	2.25	-2.34	-2.00	-1.49	1.46	-19.89	-2.78	-3.92	-5.11	-8.72
	Median	69.00	48.70	39.00	31.50	20.00	78.50	17.00	10.00	22.00	12.00	6.00	15.00	2.00	3.00	21.00	20.00	67.00	-6.70	-5.00	.900	-2.00	-2.00	-2.00	1.00	-21.00	-3.00	-3.00	-5.00	-7.00
	Std. Dev	7.85	7.83	6.21	9.75	4.44	11.00	5.81	3.32	4.02	3.37	2.78	4.60	1.69	1.69	3.24	4.0	12.60	9.47	14.66	18.07	8.08	5.14	4.17	5.98	16.13	5.63	8.54	8.29	5.62
	Min	54	31.8	22	5.0	11	33.3	7	5	11	6	0	5	0	0	14	9	29	-23.6	-51	-50.0	-22	-14	-12	-14	-56	-14	-23	-27	-21
	Max	89	66.6	51	52.9	30	100.0	33	20	32	20	14	28	8	7	30	29	98	15.6	22	43.7	17	9	7	17	15	10	12	10	-1
Yes	N																120													
	Mean	81.13	53.59	40.70	36.33	19.36	74.83	21.65	10.63	25.23	15.66	7.98	12.75	2.94	2.27	20.88	21.91	93.02	5.58	6.27	2.95	3.58	3.63	1.82	-2.45	30.27	4.25	7.02	9.33	13.04
	Median	81.00	53.50	40.00	35.20	19.00	76.40	21.00	10.00	25.00	15.00	8.00	12.00	3.00	2.00	20.50	21.00	93.00	5.05	7.35	1.90	2.00	3.00	1.00	-3.00	25.00	3.00	6.00	9.00	11.00
	Std. Dev	9.06	7.76	6.26	10.37	4.49	10.68	7.39	3.96	4.72	4.05	3.01	3.59	1.97	1.58	3.73	4.18	15.82	11.65	14.11	15.19	8.46	5.45	4.02	4.65	25.61	6.33	7.88	9.05	10.44
	Min	60	37.2	29	10.0	9	40.9	9	4	12	6	2	6	0	0	13	14	61	-16.0	-33	-36.8	-15	-10	-8	-12	-20	-8	-7	-8	1
	Max	104	74.2	57	66.6	32	95.4	39	24	38	28	16	24	11	7	32	33	137	41.7	39	50.0	28	20	15	10	128	20	38	39	49
Total	N																185													
	Mean	76.89	51.86	39.63	34.50	19.74	75.17	20.29	10.55	23.97	14.56	7.26	13.37	2.86	2.48	20.96	21.38	83.65	1.85	2.27	2.70	1.50	1.65	.65	-1.08	12.65	1.78	3.17	4.25	5.39
	Median	76.00	52.10	39.00	34.70	19.00	76.90	20.00	10.00	24.00	14.00	7.00	13.00	3.00	2.00	21.00	21.00	82.00	2.20	3.30	1.50	1.00	2.00	1.00	-1.00	11.00	2.00	3.00	3.00	5.00
	Std. Dev	10.38	8.11	6.39	10.43	4.49	10.77	7.10	3.74	4.795	4.10	3.08	4.05	1.87	1.64	3.55	4.17	19.49	12.03	15.27	16.21	8.78	5.98	4.36	5.47	33.04	6.95	9.64	11.16	13.79
	Min	54	31.8	22	5.0	9	33.3	7	4	11	6	0	5	0	0	13	9	29	-23.6	-51	-50.0	-22	-14	-12	-14	-56	-14	-23	-27	-21
	Max	104	74.2	57	66.6	32	100.0	39	24	38	28	16	28	11	7	32	33	137	41.7	39	50.0	28	20	15	17	128	20	38	39	49

Πίνακας Π.4 *Περιγραφικά μέτρα για τις απόλυτες διαφορές των μεταβλητών*

Win		dif_2Fg	dif_3Fg	dif_Ft	dif_Reb	dif_As	dif_St	dif_To	dif_Rkg	dif_Q1	dif_Q2	dif_Q3	dif_Pts
No	N						65						
	Mean	9.15	12.21	13.72	6.22	4.31	3.46	4.78	21.15	5.09	7.25	7.45	8.72
	Median	8.80	10.00	11.80	4.00	3.00	3.00	4.00	21.00	4.00	6.00	7.00	7.00
	Std. Dev	5.52	9.49	11.85	5.62	3.41	2.74	3.83	14.40	3.64	5.94	6.24	5.62
	Min	0.8	0	0	0	0	0	0	1	0	0	0	1
	Max	23.6	51	50.0	22	14	12	17	56	14	23	27	21
Yes	N						120						
	Mean	9.86	12.26	12.20	6.86	5.15	3.43	4.23	31.04	5.90	8.10	10.16	13.04
	Median	7.65	9.75	10.00	5.00	4.00	3.00	4.00	25.00	5.00	6.00	9.00	11.00
	Std. Dev	8.31	9.33	9.45	6.09	4.04	2.76	3.11	24.67	4.81	6.75	8.10	10.44
	Min	0	0	0	0	0	0	0	0	0	0	0	1
	Max	41.7	39	50.0	28	20	15	12	128	20	38	39	49
Total	N						185						
	Mean	9.61	12.24	12.73	6.63	4.85	3.44	4.43	27.57	5.62	7.80	9.21	11.52
	Median	7.90	9.90	10.30	5.00	4.00	3.00	4.00	23.00	5.00	6.00	8.00	9.00
	Std. Dev	7.44	9.36	10.35	5.92	3.84	2.74	3.38	22.09	4.44	6.48	7.59	9.26
	Min	0	0	0	0	0	0	0	0	0	0	0	1
	Max	41.7	51	50.0	28	20	15	17	128	20	38	39	49

Πίνακας Π.5 *Περιγραφικά μέτρα για τη νικήτρια ομάδα*

	Pts	*X2Fg*	*X2Fg_at*	*X3Fg*	*X3Fg_at*	*Ft*	*Ft_at*	*Reb_o*	*Reb_d*	*As*	*St*	*To*	*Bl_fv*	*Bl_ag*	*Fl_cm*	*Fl_rv*	*Rkg*	*dif_2Fg*	*dif_3Fg*	*dif_Ft*	*dif_Reb*	*dif_As*	*dif_St*	*dif_To*	*dif_Rkg*	*dif_Q1*	*dif_Q2*	*dif_Q3*	*dif_Pts*
N																10													
Mean	79.40	51.85	37.80	33.32	19.20	77.16	26.60	10.40	24.00	14.40	8.40	12.30	3.50	2.30	20.80	25.40	93.20	2.49	.37	6.37	3.20	1.20	2.40	-3.30	33.30	3.90	6.30	11.60	11.10
Median	77.00	53.00	39.00	33.30	17.00	78.80	26.00	9.50	23.00	14.00	9.00	12.00	3.00	2.50	20.00	25.50	97.00	4.50	-1.70	.15	2.00	2.00	3.00	-3.00	25.00	3.00	4.00	9.00	9.00
Std. Dev	9.26	5.89	3.68	10.05	5.65	5.25	4.27	4.58	4.60	2.797	2.07	4.00	2.46	1.57	3.05	2.50	14.34	8.67	13.49	13.00	10.26	4.13	3.37	3.62	25.02	4.56	6.72	9.54	12.13
Min	67	42.4	32	12.5	12	67.6	20	4	16	9	4	8	1	0	17	21	71	-12.6	-21	-2.4	-12	-6	-2	-9	3	-3	0	-4	-4
Max	95	61.9	43	47.0	32	83.3	34	18	32	19	11	20	8	5	25	29	116	14.9	20	39.1	22	7	8	3	77	13	20	29	31

Πίνακας Π.6 *Συντελεστές γραμμικής συσχέτισης του Pearson για όλες τις μεταβλητές*

	Pts	**X2Fg**	**X2Fg_at**	**X3Fg**	**X3Fg_at**	**Ft**	**Ft_at**	**Reb_o**	**Reb_d**	**As**	**St**	**To**	**Bl_fv**	**Bl_ag**	**Fl_cm**	**Fl_rv**	**Rkg**
Pts	1.000	0.484	0.329	**0.516**	0.038	0.033	0.277	0.065	0.223	**0.540**	0.245	-0.216	0.001	-0.134	0.106	0.172	**0.886**
X2Fg	0.484	1.000	0.020	0.022	-0.109	-0.175	-0.157	-0.303	0.041	0.400	0.196	-0.051	-0.022	-0.354	-0.072	-0.170	**0.518**
X2Fg_at	0.329	0.020	1.000	-0.053	-0.477	-0.010	-0.131	0.375	0.170	0.131	0.169	-0.360	0.034	0.234	-0.022	-0.136	0.294
X3Fg	**0.516**	0.022	-0.053	1.000	0.019	-0.003	-0.044	-0.215	0.069	0.397	0.016	0.042	-0.083	-0.063	0.008	-0.065	0.419
X3Fg_at	0.038	-0.109	-0.477	0.019	1.000	-0.007	-0.108	0.064	-0.033	0.069	-0.086	-0.039	0.098	-0.025	-0.008	-0.073	-0.049
Ft.	0.033	-0.175	-0.010	-0.003	-0.007	1.000	-0.023	-0.007	-0.142	-0.091	0.016	-0.083	-0.010	-0.012	0.036	-0.066	0.034
Ft_at	0.277	-0.157	-0.131	-0.044	-0.108	-0.023	1.000	0.210	0.125	-0.063	0.056	0.113	-0.010	-0.022	0.276	**0.827**	0.237
Reb_o	0.065	-0.303	0.375	-0.215	0.064	-0.007	0.210	1.000	0.021	0.015	-0.066	0.079	0.120	0.221	0.072	0.167	0.060
Reb_d	0.223	0.041	0.170	0.069	-0.033	-0.142	0.125	0.021	1.000	0.141	-0.042	0.153	0.235	0.023	-0.114	0.079	0.392
As	**0.540**	0.400	0.131	0.397	0.069	-0.091	-0.063	0.015	0.141	1.000	0.251	-0.075	0.072	-0.071	-0.104	-0.135	**0.665**
St	0.245	0.196	0.169	0.016	-0.086	0.016	0.056	-0.066	-0.042	0.251	1.000	-0.001	0.016	-0.141	-0.046	-0.047	0.351
To	-0.216	-0.051	-0.360	0.042	-0.039	-0.083	0.113	0.079	0.153	-0.075	-0.001	1.000	0.089	-0.100	0.185	0.158	-0.229
Bl_fv	0.001	-0.022	0.034	-0.083	0.098	-0.010	-0.010	0.120	0.235	0.072	0.016	0.089	1.000	0.072	-0.051	0.043	0.158
Bl_ag	-0.134	-0.354	0.234	-0.063	-0.025	-0.012	-0.022	0.221	0.023	-0.071	-0.141	-0.100	0.072	1.000	0.047	-0.001	-0.227
Fl_cm	0.106	-0.072	-0.022	0.008	-0.008	0.036	0.276	0.072	-0.114	-0.104	-0.046	0.185	-0.051	0.047	1.000	0.328	-0.169
Fl_rv	0.172	-0.170	-0.136	-0.065	-0.073	-0.066	**0.827**	0.167	0.079	-0.135	-0.047	0.158	0.043	-0.001	0.328	1.000	0.155
Rkg	**0.886**	**0.518**	0.294	0.419	-0.049	0.034	0.237	0.060	0.392	**0.665**	0.351	-0.229	0.158	-0.227	-0.169	0.155	1.000
dif_2Fg	0.372	**0.720**	0.089	-0.088	-0.143	-0.155	-0.001	-0.139	0.411	0.343	0.205	-0.039	0.233	-0.287	-0.054	-0.077	**0.533**
dif_3Fg	0.333	-0.014	0.044	**0.684**	-0.069	0.012	-0.115	-0.133	0.297	0.330	0.141	0.123	0.050	0.016	-0.018	-0.133	0.356
dif_Ft	-0.019	-0.164	-0.037	0.040	-0.005	**0.675**	-0.067	0.000	0.019	0.034	0.014	0.002	0.031	0.009	-0.076	-0.019	0.072
dif_Reb	0.330	0.126	0.242	0.071	-0.099	-0.059	0.193	**0.547**	**0.600**	0.215	-0.034	0.307	0.088	0.014	-0.039	0.121	0.428
dif_As	0.435	0.348	0.090	0.232	0.071	-0.095	-0.027	0.028	0.256	**0.750**	0.270	-0.029	0.173	-0.126	-0.034	-0.046	**0.589**
dif_St	0.277	0.181	0.332	-0.018	-0.105	-0.021	0.023	-0.055	-0.118	0.233	**0.771**	-0.458	-0.088	-0.060	-0.052	-0.100	0.347
dif_To	-0.296	-0.112	-0.338	0.070	-0.010	-0.079	-0.033	0.108	0.288	-0.101	**-0.507**	**0.688**	0.075	0.086	0.040	0.010	-0.282
dif_Rkg	**0.623**	0.416	0.240	0.200	-0.135	-0.014	0.238	0.084	**0.521**	**0.509**	0.431	-0.109	0.270	-0.265	-0.217	0.180	**0.860**
dif_Q1	0.432	0.348	0.185	0.142	-0.100	0.016	0.079	-0.003	0.240	0.243	0.295	-0.083	0.102	-0.126	-0.047	-0.045	0.480
dif_Q2	0.483	0.341	0.206	0.152	-0.106	0.075	0.104	0.004	0.316	0.377	0.351	-0.076	0.134	-0.116	-0.010	0.023	**0.579**
dif_Q3	**0.544**	0.385	0.263	0.198	-0.151	0.052	0.075	0.004	0.394	0.418	0.388	-0.135	0.207	-0.129	-0.044	0.005	**0.681**
dif_Pts	**0.640**	0.413	0.328	0.226	-0.152	-0.015	0.139	0.104	0.496	0.471	0.420	-0.157	0.159	-0.183	-0.110	0.034	**0.800**

* Τονίζονται όσες συσχετίσεις είναι κατά απόλυτη τιμή ≥ 0.5, με εξαίρεση την τιμή 1 που αφορά ίδιες μεταβλητές.

Σε συνέχεια του προηγούμενου πίνακα …

	dif_2Fg	dif_3Fg	dif_Ft	dif_Reb	dif_As	dif_St	dif_To	dif_Rkg	dif_Q1	dif_Q2	dif_Q3	dif_Pts
Pts	0.372	0.333	-0.019	0.330	0.435	0.277	-0.296	**0.623**	0.432	0.483	**0.544**	**0.640**
X2Fg	**0.720**	-0.014	-0.164	0.126	0.348	0.181	-0.112	0.416	0.348	0.341	0.385	0.413
X2Fg_at	0.089	0.044	-0.037	0.242	0.090	0.332	-0.338	0.240	0.185	0.206	0.263	0.328
X3Fg	-0.088	**0.684**	0.040	0.071	0.232	-0.018	0.070	0.200	0.142	0.152	0.198	0.226
X3Fg_at	-0.143	-0.069	-0.005	-0.099	0.071	-0.105	-0.010	-0.135	-0.100	-0.106	-0.151	-0.152
Ft.	-0.155	0.012	**0.675**	-0.059	-0.095	-0.021	-0.079	-0.014	0.016	0.075	0.052	-0.015
Ft_at	-0.001	-0.115	-0.067	0.193	-0.027	0.023	-0.033	0.238	0.079	0.104	0.075	0.139
Reb_o	-0.139	-0.133	0.000	**0.547**	0.028	-0.055	0.108	0.084	-0.003	0.004	0.004	0.104
Reb_d	0.411	0.297	0.019	**0.600**	0.256	-0.118	0.288	**0.521**	0.240	0.316	0.394	0.496
As	0.343	0.330	0.034	0.215	**0.750**	0.233	-0.101	**0.509**	0.243	0.377	0.418	0.471
St	0.205	0.141	0.014	-0.034	0.270	**0.771**	**-0.507**	0.431	0.295	0.351	0.388	0.420
To	-0.039	0.123	0.002	0.307	-0.029	-0.458	**0.688**	-0.109	-0.083	-0.076	-0.135	-0.157
Bl_fv	0.233	0.050	0.031	0.088	0.173	-0.088	0.075	0.270	0.102	0.134	0.207	0.159
Bl_ag	-0.287	0.016	0.009	0.014	-0.126	-0.060	0.086	-0.265	-0.126	-0.116	-0.129	-0.183
Fl_cm	-0.054	-0.018	-0.076	-0.039	-0.034	-0.052	0.040	-0.217	-0.047	-0.010	-0.044	-0.110
Fl_rv	-0.077	-0.133	-0.019	0.121	-0.046	-0.100	0.010	0.180	-0.045	0.023	0.005	0.034
Rkg	**0.533**	0.356	0.072	0.428	**0.589**	0.347	-0.282	**0.860**	0.480	**0.579**	**0.681**	**0.800**
dif_2Fg	1.000	-0.071	-0.160	0.258	0.466	0.208	-0.091	**0.638**	0.442	0.460	**0.544**	**0.628**
dif_3Fg	-0.071	1.000	0.052	0.161	0.321	0.049	0.104	0.348	0.195	0.289	0.387	0.428
dif_Ft	-0.160	0.052	1.000	0.030	0.006	-0.047	0.004	0.101	0.026	0.126	0.117	0.057
dif_Reb	0.258	0.161	0.030	1.000	0.242	-0.145	0.362	0.497	0.270	0.319	0.354	0.489
dif_As	0.466	0.321	0.006	0.242	1.000	0.260	-0.145	**0.665**	0.349	0.485	**0.546**	**0.599**
dif_St	0.208	0.049	-0.047	-0.145	0.260	1.000	**-0.755**	0.419	0.311	0.351	0.380	0.442
dif_To	-0.091	0.104	0.004	0.362	-0.145	**-0.755**	1.000	-0.296	-0.207	-0.220	-0.289	-0.333
dif_Rkg	**0.638**	0.348	0.101	0.497	**0.665**	0.419	-0.296	1.000	**0.550**	**0.671**	**0.793**	**0.931**
dif_Q1	0.442	0.195	0.026	0.270	0.349	0.311	-0.207	**0.550**	1.000	**0.741**	**0.637**	**0.582**
dif_Q2	0.460	0.289	0.126	0.319	0.485	0.351	-0.220	**0.671**	**0.741**	1.000	**0.821**	**0.691**
dif_Q3	**0.544**	0.387	0.117	0.354	**0.546**	0.380	-0.289	**0.793**	**0.637**	**0.821**	1.000	**0.837**
dif_Pts	**0.628**	0.428	0.057	0.489	**0.599**	0.442	-0.333	**0.931**	**0.582**	**0.691**	**0.837**	1.000

Πίνακας Π.7 *Συντελεστές συσχέτισης point-biserial για το αποτέλεσμα και polyserial για τα γκρουπ δυναμικότητας, με τις συνεχείς μεταβλητές*

	Pts	X2Fg	X2Fg_at	X3Fg	X3Fg_at	Ft	Ft_at	Reb_o	Reb_d	As	St	To	Bl_fv	Bl_ag	Fl_cm	Fl_rv
Win	**0.554**	0.289	0.227	0.238	-0.116	-0.042	0.260	0.029	0.359	0.363	0.317	-0.207	0.060	-0.177	-0.030	0.173
Group	-0.206	-0.240	-0.023	-0.029	0.047	-0.069	-0.025	-0.006	-0.063	-0.196	-0.048	0.153	-0.100	0.091	0.202	-0.061
dif_Group	-0.311	-0.292	-0.105	-0.025	0.007	-0.056	-0.047	-0.094	-0.135	-0.233	-0.118	0.070	-0.182	0.161	0.233	-0.056

Σε συνέχεια του παραπάνω πίνακα …

	Rkg	dif_2Fg	dif_3Fg	dif_Ft	dif_Reb	dif_As	dif_St	dif_To	dif_Rkg	dif_Q1	dif_Q2	dif_Q3	dif_Pts
Win	**0.653**	0.421	0.356	0.021	0.322	0.450	0.362	-0.341	**0.725**	0.484	**0.542**	**0.617**	**0.754**
Group	-0.329	-0.307	0.066	-0.063	-0.131	-0.284	-0.053	0.151	-0.349	-0.285	-0.283	-0.340	-0.287
dif_Group	-0.424	-0.347	-0.010	-0.050	-0.225	-0.350	-0.093	0.148	-0.455	-0.324	-0.310	-0.331	-0.384

Πίνακας Π.8 *Συντελεστής συσχέτισης Somers' D για το αποτέλεσμα με τα γκρουπ δυναμικότητας*

	Group	dif_Group
Win	-0.226	-0.333

Σχέση υπολογισμού point-biserial συντελεστή

$$r_{pb} = \frac{(Y_1 - Y_0)}{\sigma_Y}\sqrt{pq}$$

Όπου Y_0 & Y_1: μέση τιμή για τις τιμές της συνεχούς μεταβλητής Y που αντιστοιχούν στο επίπεδο 0 και 1 αντίστοιχα της δίτιμης μεταβλητής X | q & p: ποσοστά επί του συνόλου τιμών 0 και 1 της X αντίστοιχα | σ_Y: πληθυσμιακή τυπική απόκλιση της Y

Συντελεστής ασυμμετρίας (Skewness)

Για τον υπολογισμό χρησιμοποιήθηκε ο προσαρμοσμένος Fisher-Pearson τυποποιημένος συντελεστής:

$$G_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Όπου n: μέγεθος δείγματος | x_i: i τιμή της x | $\bar{x}$: μέση τιμή του δείγματος | s: τυπική απόκλιση τους δείγματος

Συντελεστής κύρτωσης (Kurtosis)

Ο εκτιμητής της υπερβάλλουσας κύρτωσης (excess kurtosis) που χρησιμοποιήθηκε είναι:

$$G_2 = \frac{(n+1)n(n-1)}{(n-2)(n-3)} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^2} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

Έλεγχοι καλής προσαρμογής (goodness of fit tests)

Τα παρακάτω κριτήρια συγκρίνουν την εμπειρική αθροιστική συνάρτηση κατανομής με την αντίστοιχη της προσαρμοσμένης. Όπου

$z_{(i)} = \hat{F}(x_{(i)})$ η προσαρμοσμένη συνάρτηση κατανομής για κάθε παρατήρηση i.

Kolmogorov-Smirnov D

Η μέγιστη απόσταση της εμπειρικής συνάρτησης κατανομής, πάνω από την αντίστοιχη προσαρμοσμένη συμβολίζεται με D^+ και η μέγιστη απόσταση της εμπειρικής συνάρτησης κατανομής, κάτω από την αντίστοιχη προσαρμοσμένη συμβολίζεται με D^-. Το D είναι η μεγαλύτερη από τις δύο αυτές αποστάσεις.

$$D^+ = \max_i \left\{ \frac{1}{n} - z_{(i)} \right\} \ \& \ D^- = \max_i \left\{ z_{(i)} - \frac{i-1}{n} \right\}$$

$$D = \max(D^+, D^-)$$

Cramer-Von Mises W^2

$$W^2 = \sum_{i=1}^{n} \left(z_{(i)} - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}$$

Anderson-Darling A^2

$$A^2 = -n - \frac{\sum_{i=1}^{n}\left((2i-1)\ln(z_{(i)}) + (2n+1-2i)\ln(1-z_{(i)})\right)}{n}$$

<u>Συναρτήσεις πυκνότητας πιθανότητας των παρακάτω συνεχών Κατανομών</u>

Αντίστροφη Γκαουσιανή (Inverse Gaussian)

$\lambda > 0$ (shape)

$\mu > 0$ (mean)

$0 < x < +\infty$

$$f(x) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right)$$

Logistic

μ (location)

$\sigma > 0$ (scale)

$-\infty < x < +\infty$

$$f(x) = \frac{\exp\left(-\frac{x-\mu}{\sigma}\right)}{\sigma\left(1+\exp\left(-\frac{x-\mu}{\sigma}\right)\right)^2}$$

Log-logistic με 3 παραμέτρους

$\alpha > 0$ (shape)

$\beta > 0$ (scale)

γ (location) όταν γ=0, τότε προκύπτει η Log-Logistic με 2 παραμέτρους

$\gamma \le x < +\infty$

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{-2}$$

Generalized Logistic

k (shape)

σ > 0 (scale)

μ (location)

$1+k\frac{(x-\mu)}{\sigma}>0$ για κ≠0

$-\infty < x < +\infty$ για κ=0

$$f(x)=\begin{cases}\dfrac{(1+kz)^{-1-1/k}}{\sigma(1+(1+kz)^{-1/k})^2} & \text{για } k\neq 0\\ \dfrac{\exp(-z)}{\sigma(1+\exp(-z))^2} & \text{για } k=0\end{cases} \quad \text{, όπου } z=\frac{x-\mu}{\sigma}$$

Log-normal με 3 παραμέτρους

σ > 0 (shape)

μ ∈ **R** (log-scale)

γ (location) όταν γ=0, τότε προκύπτει η log-normal με 2 παραμέτρους

$\gamma < x < +\infty$

$$f(x)=\frac{\exp\left(-\frac{1}{2}\left(\frac{\ln(x-\gamma)-\mu}{\sigma}\right)^2\right)}{(x-\gamma)\sigma\sqrt{2\pi}}$$

Γάμμα με 3 παραμέτρους

α > 0 (shape)

β > 0 (scale)

γ (location) όταν γ=0, τότε προκύπτει η Γάμμα με 2 παραμέτρους

$\gamma \le x < +\infty$

$$f(x)=\frac{(x-\gamma)^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)}\exp(-(x-\gamma)/\beta)$$

Weibull

α > 0 (shape)

β > 0 (scale)

$0 \le x < +\infty$

$$f(x)=\frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right)$$

Πίνακας Π.9 *Αποτελέσματα ανάλυσης λογιστικού μοντέλου με βάση την Rkg*

```
> glm<-
glm(Win~Rkg+dif_Q3+Ft_at+dif_Reb+dif_To+dif_3Fg,family=binomial)
> anova(glm,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Win

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                       184    239.863
Rkg       1  112.677       183    127.186 < 2.2e-16 ***
dif_Q3    1   27.979       182     99.207 1.226e-07 ***
Ft_at     1    7.560       181     91.647 0.0059680 **
dif_Reb   1    5.826       180     85.821 0.0157919 *
dif_To    1   11.748       179     74.073 0.0006091 ***
dif_3Fg   1   10.653       178     63.420 0.0010989 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(glm)

Call:
glm(formula = Win ~ Rkg + dif_Q3 + Ft_at + dif_Reb + dif_To +
    dif_3Fg, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.74387  -0.10000   0.01927   0.14918   2.50277

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.97077    3.10785  -3.852 0.000117 ***
Rkg           0.12337    0.03748   3.292 0.000995 ***
dif_Q3        0.16779    0.05432   3.089 0.002010 **
Ft_at         0.13407    0.05395   2.485 0.012949 *
dif_Reb       0.21480    0.05912   3.633 0.000280 ***
dif_To       -0.39160    0.10952  -3.575 0.000350 ***
dif_3Fg       0.08309    0.03031   2.741 0.006121 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 239.86  on 184  degrees of freedom
Residual deviance:  63.42  on 178  degrees of freedom
AIC: 77.42

Number of Fisher Scoring iterations: 8
```

Πίνακας Π.10 *Αποτελέσματα ανάλυσης βέλτιστου logit μοντέλου με 5 μεταβλητές*

```
> glm<-glm(Win~dif_2Fg+dif_3Fg+dif_To+dif_Reb+Ft_at,family=binomial)
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> anova(glm,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Win

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                       184    239.863
dif_2Fg   1   38.679       183    201.184 4.996e-10 ***
dif_3Fg   1   38.712       182    162.472 4.912e-10 ***
dif_To    1   37.601       181    124.871 8.680e-10 ***
dif_Reb   1   59.734       180     65.137 1.086e-14 ***
Ft_at     1   25.626       179     39.511 4.143e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(glm)

Call:
glm(formula = Win ~ dif_2Fg + dif_3Fg + dif_To + dif_Reb +
    Ft_at, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.64637  -0.01267   0.00133   0.05708   2.24855

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.44355    1.66416  -3.271 0.001071 **
dif_2Fg      0.41294    0.10736   3.846 0.000120 ***
dif_3Fg      0.31523    0.07439   4.238 2.26e-05 ***
dif_To      -1.04200    0.25009  -4.166 3.09e-05 ***
dif_Reb      0.46272    0.11017   4.200 2.67e-05 ***
Ft_at        0.30871    0.08826   3.498 0.000469 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 239.863  on 184  degrees of freedom
Residual deviance:  39.511  on 179  degrees of freedom
AIC: 51.511

Number of Fisher Scoring iterations: 9

> vif(glm)
 dif_2Fg  dif_3Fg   dif_To  dif_Reb    Ft_at
3.912702 6.884043 9.880836 4.022474 1.833132
```

Πίνακας Π.11 *Αποτελέσματα ανάλυσης αναθεωρημένου μοντέλου*

```
> glm<-glm(Win ~ dif_2Fg + dif_3Fg + dif_To + Ft_at, family=binomial)
> anova(glm,test="Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: Win

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                      178    228.331
dif_2Fg  1   40.995       177    187.336 1.526e-10 ***
dif_3Fg  1   43.077       176    144.259 5.263e-11 ***
dif_To   1   36.582       175    107.677 1.464e-09 ***
Ft_at    1   36.939       174     70.738 1.219e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(glm)

Call:
glm(formula = Win ~ dif_2Fg + dif_3Fg + dif_To + Ft_at, family =
binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.13817  -0.12682   0.03926   0.19690   1.96989

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.18425    1.07923  -3.877 0.000106 ***
dif_2Fg      0.25445    0.05017   5.072 3.94e-07 ***
dif_3Fg      0.18716    0.03610   5.184 2.17e-07 ***
dif_To      -0.41902    0.09170  -4.570 4.89e-06 ***
Ft_at        0.25539    0.05591   4.568 4.93e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 228.331  on 178  degrees of freedom
Residual deviance:  70.738  on 174  degrees of freedom
AIC: 80.738

Number of Fisher Scoring iterations: 7
```

Βιβλιογραφία

Ελληνική

Τζαβελάς Γ. (2007). *Γενικευμένα Γραμμικά Μοντέλα, Μέρος Α – Λογιστική Παλινδρόμηση*, Πανεπιστήμιο Πειραιώς.

Ξένη

Agresti, A. (2006). *Building and Applying Logistic Regression Models, An Introduction to Categorical Data Analysis*, 2nd Edition, Wiley, New Jersey.

Belsley, D. A., E. Kuh, and R. E. Welsh. (1980). *Regression Diagnostics*, Wiley, New York.

Bollen, Kenneth A.; and Jackman, Robert W. (1990). *Regression diagnostics: An expository treatment of outliers and influential cases*, in Fox, John; and Long, J. Scott (eds.), Modern Methods of Data Analysis (pp. 257-91), Newbury Park, CA: Sage.

Boulier, B. and Stekler, H. O., (1999). Are Sports Seedings Good Predictors?: An Evaluation, *International Journal of Forecasting*, 15, 83-91.

Bradford S. Jones (2009). *Logit and Probit,* Department of Political Science, University of California.

Burnham, Kenneth P., Anderson, David R. (2002). *Model Selection and Multimodel Inference*, 2nd ed., Springer, New York.

Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M. Ingersoll (2002). *An Introduction to Logistic Regression Analysis and Reporting*, Indiana University-Bloomington, Ebsco Publishing.

Chatterjee, S, Hadi, A.S., Price, B. (2000). *Regression Analysis by Example, 3rd edition*, Wiley, New York, p104.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman & Hall, New York.

Cooper, H., DeNeve, K. M., and Mosteller, F. (1992). *Predicting Professional Game Outcomes From Intermediate Game Scores*, CHANCE, 5, 3-4, 18-22.

Dayton C. M. (1992). *Logistic Regression Analysis*, Department of Measurement, Statistics & Evaluation, University of Maryland.

DS (2009). *Guide to Credit Scoring in R,* Handbook.

Everitt B. S. and Hothorn Torsten (2011). *A Handbook of Statistical Analyses Using R*.

Friendly M. (2012). *Visualizing Categorical Data with SAS and R*, York University.

Gabel A. and Redner S. (2011). *Random Walk Picture of Basketball Scoring*, Center for Polymer Studies and Department of Physics, Boston University, Massachusetts.

Gujarati D. (1988). *Basic Econometrics*, 2nd Edition, McGraw-Hill Book Company, New York.

Habing B. (2004). *More on Outlier Diagnostics*, University of South Carolina.

Hallett C. D. (1999). *Goodness of Fit Tests in Logistic Regression,* University of Toronto.

Harrell E. Frank (2001). *Regression Modeling Strategies*: With Applications to Linear Models, Logistic Regression, and Survival Analysis, Springer, New York.

Harville A. David and Smith H. Michael (1994). The Home-Court Advantage: How Large Is It, and Does It Vary from Team to Team, *The American Statistician*, Vol. 48, No. 1, pp. 22-28.

Huber B. (1999). *Fitting Distributions to Data*. Practical Issues in the Use of Probabilistic Risk Assessment Sarasota, Florida.

Ibáñez J. Sergio, Javier García, Sebastian Feu, Alberto Lorenzo and Jaime Sampaio (2009). Effects of consecutive basketball games on the game-related statistics that discriminate winner and losing teams, *Journal of Sports Science and Medicine* (2009) 8, 458-462.

Kubatko, Justin; Oliver, Dean; Pelton, Kevin; and Rosenbaum, Dan T. (2007). A Starting Point for Analyzing Basketball Statistics, *Journal of Quantitative Analysis in Sports*: Vol. 3: Iss. 3, Article 1.

Kvam P. and Sokol S. Joel (2006). A Logistic Regression/Markov Chain Model For NCAA Basketball, *Naval Research Logistics* 53.

McDonald B. (2002). *A Teaching Note on Cook's Distance – A Guideline*, Institute of Information and Mathematical Science, Massey University at Albany, Auckland.

Newson R. (2006). Confidence intervals for rank statistics: Somers' D and extensions, *The Stata Journal*, 6, Number 3, pp. 309–334.

O'Halloran S. (2005). *Logit/Probit*, Columbia University, New York.

Parmesan S. and Mooney W. (2011). *Forecasting Basketball Games Using Bayes Networks*.

Peduzzi, P., Concato, J., Kemper E., Holford, T.R. and Feinstein, A.R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49: 1372-1379.

Ricci V. (2005). *Fitting Distributions in R*.

Rossiter D. G. (2011). *Technical Note: Analyzing land cover change with logistic regression in R.*

Schwab A. J. (2003). *Logistic Regression – Complete Problems*, University of Texas.

Shirley K. (2007). *A Markov Model for Basketball,* Applied Statistics Center, Columbia University, New York.

Smith T. & Schwertman C. Neil (1999). Can the NCAA Basketball Tournament Seeding be Used to Predict Margin of Victory?, *The American Statistician*, Vol. 53, No. 2., pp. 94-98.

Statgraphics manual (2006). *Distribution Fitting (Uncensored Data),* StatPoint, Inc.

Statgraphics manual (2006). *Probability Distributions,* StatPoint, Inc.

Stekler H.O. and Klein A. (2011). *Predicting the Outcomes of NCAA Basketball Championship Game*, Research Program on Forecasting, Department of Economics, George Washington University.

Stern S. H. (1994). A Brownian Motion Model for the Progress of Sports Scores, Source: *Journal of the American Statistical Association*, Vol. 89, No. 427, pp. 1128-1134.

Storvik Geir (2011). *Numerical optimization of likelihoods*, University of Oslo.

Wickham H. (2009). *ggplot2: Elegant Graphics for Data Analysis*, Springer, Houston.

Williams R. (2011). *Logistic Regression, Part III: Hypothesis Testing, Comparisons to OLS*, University of Notre Dame.

Witkos R. (2010). *Determining the Success of NCAA Basketball Teams through Team Characteristics.*

Σύνδεσμοι

http://aje.oxfordjournals.org/content/165/6/710.full

http://www.andrews.edu/~calkins/math/edrm611/edrm13.htm#POINTB (*Keith_G.Calkins, 2005)*

http://www.ats.ucla.edu/stat/mult_pkg/faq/general/psuedo_rsquareds.htm

http://www.cookbook-r.com/Graphs/

http://cran.r-project.org/web/views/Distributions.html

http://en.wikipedia.org/wiki/Variance_inflation_factor

http://www.euroleague.net/main/results/

http://www.in-the-game.org/?page_id=10227

http://www.inside-r.org/packages/cran/corrplot/docs/corrplot

http://www.inside-r.org/packages/cran/fitdistrplus/docs/fitdist

http://www.mathwave.com/help/easyfit/html/analyses/distributions/_continuous.html

http://www.r-bloggers.com/search/ggplot2/

http://rocr.bioinf.mpi-sb.mpg.de/ROCR.pdf

http://rstudio-pubs-static.s3.amazonaws.com/2107_4eb1adc1e4d44b93b6fde7eb801519fe.html

http://sas-and-r.blogspot.gr/2010/09/example-87-hosmer-and-lemeshow-goodness.html

http://www.statmethods.net/index.html