ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Σχολή Θετικών Επιστημών

Τμήμα Μαθηματικών

Πτυχιακή Εργασία

Θεωρία και εφαρμογές στα γενικευμένα γραμμικά μοντέλα

Μαρία Τρέντου (311/2014132)

Επιβλέπων Καθηγητής: Δημητράκος Θεοδόσης

Επιτροπή:

Παπαλεξίου Νικόλαος και Παπασαλούρος Ανδρέας

Σάμος, Οκτώβριος 2018

Περίληψη

Σκοπός αυτής της εργασίας είναι η μελέτη των Γενικευμένων Γραμμικών Μοντέλων. Η εργασία αυτή χωρίζεται σε τέσσερα κεφάλαια.

Στο πρώτο κεφάλαιο συναντάμε εισαγωγικές έννοιες από την Στατιστική και γίνεται μια γενική αναφορά στα Γενικευμένα Γραμμικά Μοντέλα.

Στο δεύτερο κεφάλαιο αναλύεται το απλό και πολλαπλό γραμμικό μοντέλο και παρουσιάζεται ένα παράδειγμα πολλαπλής παλινδρόμησης στο στατιστικό πακέτο R.

Στο τρίτο κεφάλαιο αναλύεται η Λογιστική παλινδρόμηση, δηλαδή η παλινδρόμηση για δίτιμες τυχαίες μεταβλητές και παρουσιάζεται και ένα παράδειγμα χρήσης της Λογιστικής παλινδρόμησης στην R.

Τέλος, στο τέταρτο κεφάλαιο θα δείτε την Poisson παλινδρόμηση και ακόμα ένα παράδειγμα που μας επεξηγεί την χρήση την Poisson παλινδρόμησης μέσω της R.

ABSTRACT

The purpose of this work is to study the generalized linear models. This work is divided into four chapters. At the first chapter we come across introductory concepts from statistics and make a general reference to generalized linear models.

At the second chapter we analyze the simple and multiple linear models and we present an example of multiple regression is R.

At the third chapter we analyze the logistic regression, namely the regression for two random variables, and we provide an example for the usage of the logistic regression in R.

At the fourth chapter we will study the Poisson regression. An example is illustrated the usage of Poisson regression through R is presented.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή στα γενικευμένα γραμμικά μοντέλα

1.1. Εισαγωγή

Τα Γενικευμένα Γραμμικά Μοντέλα (Generalized Linear Models ή GLMs) αρχικά αναπτύχθηκαν από τους John Nelder και Robert Wedderburn το 1972. Αργότερα, το 1983, ο Peter McCullagh μαζί με τον John Nelder έγραψαν το βιβλίο Generalized Liner Models. Τα Γενικευμένα Γραμμικά Μοντέλα αποτελούν μια επέκταση των Μοντέλων Παλινδρόμησης και είναι πολύ σημαντικά για την Στατιστική Ανάλυση Δεδομένων.

Για να αναπτύξουμε την θεωρία των γενικευμένων γραμμικών μοντέλων, είναι σημαντική η γνώση κάποιων στοιχείων από την Γραμμική Άλγεβρα και την Στατιστική.

Οι στατιστικές μέθοδοι που θα δούμε αφορούν την ανάλυση των σχέσεων μεταξύ μετρήσεων που γίνονται σε διάφορες ομάδες αντικειμένων, π.χ. τα ύψη ή τα βάρη και οι ηλικίες των αγοριών και των κοριτσιών, η ανάπτυξη ενός φυτού σε διαφορετικές συνθήκες καλλιέργειας.

1.2. Διαχωρισμός μεταβλητών

Για να εφαρμόσουμε τέτοιες μεθόδους διαχωρίζουμε τις μεταβλητές μας ως εξής:

1. Εξαρτημένη μεταβλητή: Είναι η τυχαία μεταβλητή η οποία μεταβάλλεται ελεύθερα και αυτή η οποία εμείς επιδιώκουμε την εκτίμησή της.

2. Ανεξάρτητες μεταβλητές: Θεωρούνται μη τυχαίες μεταβλητές και παίρνουν συγκεκριμένες τιμές ανάλογα με το πρόβλημά μας.

Οι μεταβλητές μπορούν εκ νέου να ταξινομηθούν ως:

- Κατηγορικές ή ποιοτικές π.χ. χρώμα ματιών, τύπος αίματος, φύλο κ.τ.λ.
- Διάταξης π.χ. όταν η ηλικία αναγράφεται ως νέος, μεσήλικας, γέρος κ.τ.λ.

- Συνεχείς μεταβλητές οι παρατηρήσεις μπορούν να πάρουν οποιαδήποτε τιμή σε ένα διάστημα π.χ. ύψος φοιτητών τρίτου έτους.

Μεταξύ των μεταβλητών μας αναζητούμε μια σχέση π.χ. γραμμική, λογαριθμική κτλ. Όταν βρεθεί η σχέση (μοντέλο) αυτή μπορούμε να την χρησιμοποιήσουμε για πρόβλεψη, για να διαπιστώσουμε ποιες από τις ανεξάρτητες μεταβλητές επιδρούν περισσότερο στην εξαρτημένη, για την εκτίμηση παραμέτρων και τον έλεγχο υποθέσεων.

Οι σχέσεις μεταξύ των μεταβλητών χωρίζονται σε δυο κατηγορίες, τις συναρτησιακές σχέσεις και τις στατιστικές σχέσεις.

Μια συναρτησιακή σχέση έχει την εξής μορφή:

y = f(x), όπου η x είναι η ανεξάρτητη και y η εξαρτημένη μεταβλητή.

Π.χ. y = 2x

Μια στατιστική σχέση διαφέρει από τη συναρτησιακή σχέση στο ότι δεν συνιστά μια τέλεια σχέση. Διέπεται από δύο αρχές:

1. Τη διασπορά των σημείων (x_i, y_i) γύρω από την καμπύλη της σχέσης. Το διάγραμμα των σημείων (x_i, y_i) καλείται διάγραμμα διασποράς.
2. Την τάση της εξαρτημένης μεταβλητής να μεταβάλλεται σε σχέση με τις ανεξάρτητες με κάποιο συστηματικό τρόπο.

1.3. Εκθετική οικογένεια κατανομών

Ορισμός: Έστω μία τυχαία μεταβλητή Y της οποίας η συνάρτηση (πυκνότητας) πιθανότητας εξαρτάται από μία παράμετρο θ. Η κατανομή ανήκει στην εκθετική οικογένεια, αν μπορεί να γραφεί στη μορφή:

$$f(y;\theta) = s(y) \cdot t(\theta) \cdot e^{a(y)\cdot b(\theta)}$$

ή ισοδύναμα $f(y;\theta) = exp[a(y) \cdot b(\theta) + c(\theta) + d(y)],\ y \in Y, \theta \in \Theta$

Το στήριγμα της συνάρτησης (πυκνότητας) πιθανότητας S = {y: f(y;θ) > 0} πρέπει να είναι ανεξάρτητο της παραμέτρου θ.

Αν a(y) = y, τότε λέμε πως η κατανομή είναι σε κανονική μορφή και b(θ) είναι η φυσική παράμετρος της κατανομής.

1.3.1. Παραδείγματα κατανομών της εκθετικής οικογένειας

1. Η Κανονική κατανομή με γνωστή διασπορά και σ.π.π:

$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{(y-\mu)^2}{2\cdot\sigma^2}} = e^{\frac{y\cdot\mu-\frac{\mu^2}{2}}{\sigma^2}-\frac{1}{2}\cdot[\frac{y^2}{\sigma^2}+\ln 2\cdot\pi\cdot\sigma^2]}, y\in\mathbb{R}$$

Άρα ανήκει στην εκθετική οικογένεια κατανομών με θ = μ, a(y) = y, b(θ) = $\frac{\mu}{\sigma^2}$, c(θ) = $-(\frac{\mu^2}{2\sigma^2}+\frac{1}{2}\cdot\ln 2\cdot\pi\cdot\sigma^2)$, d(y) = $-\frac{y^2}{2\sigma^2}$ και το στήριγμα είναι το $\mathbb{R}$ που δεν εξαρτάται από το θ.

2. Η Poisson κατανομή με παράμετρο λ και συνάρτηση πιθανότητας:

$$f(y;\lambda) = \frac{\lambda^y\cdot e^{-\lambda}}{y!} = exp[y\cdot\ln\lambda-\lambda-\ln y!]$$

$και\ \theta=\lambda, a(y)=y, b(\theta)=\ln\theta, c(\theta)=-\lambda,\ d(y)=y!$ ανήκει στην εκθετική οικογένεια κατανομών. Το στήριγμα είναι το $\mathbb{R}$, οπότε δεν εξαρτάται από το θ.

3. Η Διωνυμική κατανομή με παραμέτρους n και p και συνάρτηση πιθανότητας:

$$f(y;n,p) = \binom{n}{p}\cdot p^y\cdot(1-p)^{n-y} = exp[y\cdot\ln\frac{p}{1-p}+n\cdot\ln(1-p)+\ln\binom{n}{y}]$$

με a(y) = y, b(θ) = $\ln\frac{p}{1-p}$, c(θ) = $n\cdot\ln(1-p)$, d(y) = $\ln\binom{n}{y}$ ανήκει στην εκθετική οικογένεια κατανομών. Το στήριγμα είναι υποσύνολο των ακεραίων άρα είναι ανεξάρτητο του θ.

1.3.2. Ιδιότητες της εκθετικής οικογένειας κατανομών

1. Ο παραμετρικός χώρος Θ για τις κατανομές τυπικής εκθετικής μορφής είναι κυρτό.
2. Για τις κατανομές τυπικής εκθετικής μορφής η διακύμανση είναι μια αυστηρώς μονότονη συνάρτηση της μέσης τιμής.
3. Ο εκτιμητής μέγιστης πιθανοφάνειας για την φυσική παράμετρο θ υπάρχει πάντα και είναι μεροληπτικός εκτός από την Κανονική κατανομή.
4. Η εκθετική οικογένεια κατανομών έχει την ιδιότητα του μονότονου λόγου πιθανοφάνειας, μια ιδιότητα πολύ σημαντική για τον έλεγχο υποθέσεων.
5. Για την μέση τιμή και την διακύμανση της α(Y), Y συνεχής τ.μ. έχουμε:

$$\int f(y;\theta)\,dy = 1\ ,με\ f(y;\theta)\ \sigma.\pi.\pi.$$

Παραγωγίζουμε ως προς θ και έχουμε:

$$\int\frac{df(y;\theta)}{d\theta}dy = 0$$

Συνεπώς αν παραγωγίσουμε άλλη μια φορά τα δύο μέλη της προηγούμενης σχέσης καταλήγουμε στην:

$$\int \frac{d^2 f(y;\theta)}{d\theta^2} dy = 0$$

Άρα κάνοντας αντικατάσταση στην σχέση με

$$f(y;\theta) = exp[a(y) \cdot b(\theta) + c(\theta) + d(y)]$$

θα έχουμε:

$$\frac{df(y;\theta)}{d\theta} = [a(y) \cdot b^{'}(\theta) + c^{'}(\theta)] \cdot f(y;\theta)$$

Οπότε έχουμε: $\int [a(y) \cdot b^{'}(\theta) + c^{'}(\theta)] \cdot f(y;\theta) = 0$

ή ισοδύναμα $E[a(Y)] \cdot b^{'}(\theta) \ + c^{'}(\theta) = 0 \Longrightarrow E[a(Y)] = -\frac{c^{'}(\theta)}{b^{'}(\theta)}$.

Δουλεύοντας με παρόμοιο τρόπο και χρησιμοποιώντας την 2η παράγωγο έχουμε ότι:

$$Var[a(Y)] = \frac{b^{''}(\theta) \cdot c^{'}(\theta) - c^{''}(\theta) \cdot b^{'}(\theta)}{[b^{'}(\theta)]^3}$$

Αν τώρα πάρουμε την σχέση την ln-πιθανοφάνειας έχουμε:

$$l(\theta; y) = a(y) \cdot b(\theta) + c(\theta) + d(y)$$

και παραγωγίζοντας ως προς θ παίρνουμε την σχέση:

$$U = u(\theta; y) = \frac{dl(\theta; y)}{d\theta} = a(y) \cdot b^{'}(\theta) + c^{'}(\theta)$$

Η συνάρτηση $u(\theta; y)$ καλείται score συνάρτηση και εφόσον εξαρτάται η τιμή της από το y μπορούμε να την θεωρήσουμε τ.μ. με μέση τιμή:

$$E(U) = E[a(Y)] \cdot b^{'}(\theta) + c^{'}(\theta) = -\frac{c^{'}(\theta)}{b^{'}(\theta)} \cdot b^{'}(\theta) + c^{'}(\theta) = 0$$

και διακύμανση:

$$Var(U) = \left[b^{'}(\theta)\right]^2 \cdot Var[a(Y)] = \frac{b^{''}(\theta) \cdot c^{'}(\theta)}{b^{'}(\theta)} - c^{''}(\theta)$$

1.4. Στατιστικό μοντέλο

Για να κάνουμε μια στατιστική μελέτη πρέπει να κάνουμε πρόβλεψη της τιμής μιας μεταβλητής, της μεταβλητής απόκρισης, η οποία γίνεται μέσω γνωστών μεταβλητών, των επεξηγηματικών. Για την δημιουργία στατιστικού μοντέλου αρχικά χρειαζόμαστε μια εξίσωση που να συνδέει τη μεταβλητή απόκρισης με την επεξηγηματική μεταβλητή καθώς και την κατανομή της μεταβλητής απόκρισης. Έπειτα πρέπει να κάνουμε εκτίμηση των παραμέτρων του μοντέλου μας, κάτι που

γίνεται μέσω διαστημάτων εμπιστοσύνης αλλά και ελέγχων υποθέσεων. Τέλος ερμηνεύουμε τις τιμές των αποτελεσμάτων και την καταλληλότητα του μοντέλου μας, πόσο καλά δηλαδή το μοντέλο ερμηνεύει τα δεδομένα μας.

1.5. Γενικευμένα Γραμμικά Μοντέλα

Η σημαντικότερη κατανομή για την δημιουργία στατιστικών μοντέλων είναι η κανονική κατανομή, αφού θα μας βόλευε η μεταβλητή απόκρισης να ακολουθεί την κατανομή αυτήν. Αυτό όμως δεν είναι πάντοτε εφικτό, αφού μια τιμή μπορεί να είναι δίτιμη, δηλαδή π.χ. να λαμβάνει τις τιμές 0, 1. Οπότε οι μεταβλητές απόκρισης παίρνουν τιμές από μια γενικότερη οικογένεια κατανομών. Στα γενικευμένα γραμμικά μοντέλα η μεταβλητή απόκρισης Y δοθείσας της επεξηγηματικής X ακολουθεί κατανομές της εκθετικής οικογένειας κατανομών και ειδικότερα τις Κανονική, Διωνυμική και Poisson.

Ορισμός. Ένα γενικευμένο γραμμικό μοντέλο ορίζεται από ένα σύνολο ανεξάρτητων τ.μ. $Y_1, \dots, Y_N$ καθεμία από τις οποίες ακολουθεί κάποια κατανομή που ανήκει στην εκθετική οικογένεια κατανομών και έχει τις εξής ιδιότητες:

1. Η κατανομή που ακολουθεί κάθε Y_i έχει την κανονική μορφή και εξαρτάται από μια παράμετρο θ_i μόνο,
$$f(y_i; \theta_i) = exp[a(y_i) \cdot b(\theta_i) + c(\theta_i) + d(y_i)]$$
2. Οι Y_i έχουν όλες κατανομές της ίδιας μορφής, π.χ. όλες ακολουθούν κανονική κατανομή. Έτσι οι από κοινού συνάρτηση (πυκνότητας) πιθανότητας των $Y_1, \dots, Y_N$ γίνεται:
$$f(y_1, \dots, y_N, \theta_1, \dots, \theta_N) = \prod_{i=1}^{N} exp[y_i \cdot b(\theta_i) + c(\theta_i) + d(y_i)]$$
$$= exp\left[\sum_{i=1}^{N} y_i \cdot b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)\right]$$

Για να προσδιορίσουμε το μοντέλο χρησιμοποιούμε ένα σύνολο παραμέτρων b μεγέθους p, p<N. Υποθέτουμε ότι $E(Y_i) = \mu_i$, η οποία είναι συνάρτηση του θ_i.

Για ένα γενικευμένο γραμμικό μοντέλο έχουμε έναν μετασχηματισμό του μ_i με
$$g(\mu_i) = x_i^T \cdot b.$$

Η g είναι μονότονη και διαφορίσιμη συνάρτηση και καλείται συνάρτηση σύνδεσης ενώ το διάνυσμα x είναι οι επεξηγηματικές μας μεταβλητές:

$$x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix} \kappa\alpha\iota\ b = \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix}$$

Οι μεταβλητές $x_i, i = 1, \dots, p$ δημιουργούν μια γραμμική πρόβλεψη, την

$$\eta = \sum_{i=1}^{p} x_i \cdot b_i.$$

Συνοπτικά λοιπόν ένα γενικευμένο γραμμικό μοντέλο αποτελείται από:

- $Τις\ Y_1, \dots, Y_N$ με κατανομή από την εκθετική οικογένεια κατανομών
- Τις παραμέτρους β μέσω των οποίων καθορίζονται οι μέσοι των Y_i, μ_i.
- Τις επεξηγηματικές μεταβλητές $X = \begin{bmatrix} X_1^T \\ \vdots \\ X_N^T \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{N1} & \cdots & X_{Np} \end{bmatrix}$
- Την συνάρτηση σύνδεσης (link function): $g(\mu_i) = x_i^T \cdot b$.

Η συνάρτηση σύνδεσης μας δίνει μια σχέση για την γραμμική παράμετρο και την μέση τιμή μ της μεταβλητής απόκρισης Y. Στα γραμμικά μοντέλα η μέση τιμή μ ταυτίζεται με την γραμμική πρόβλεψη, οπότε και η ταυτοτική συνάρτηση σύνδεσης παίρνει οποιαδήποτε πραγματική τιμή. Ωστόσο σε διακριτές κατανομές, όπως την Poisson, πρέπει μ > 0, αλλά η η όχι απαραίτητα θετικό. Συνεπώς δημιουργείται η ανάγκη εισαγωγής της λογαριθμικής συνάρτησης η = log(μ) ως συνάρτηση σύνδεσης ή και της εκθετικής μ = e^{η}, ούτως ώστε να υπάρχει γραμμική σχέση.

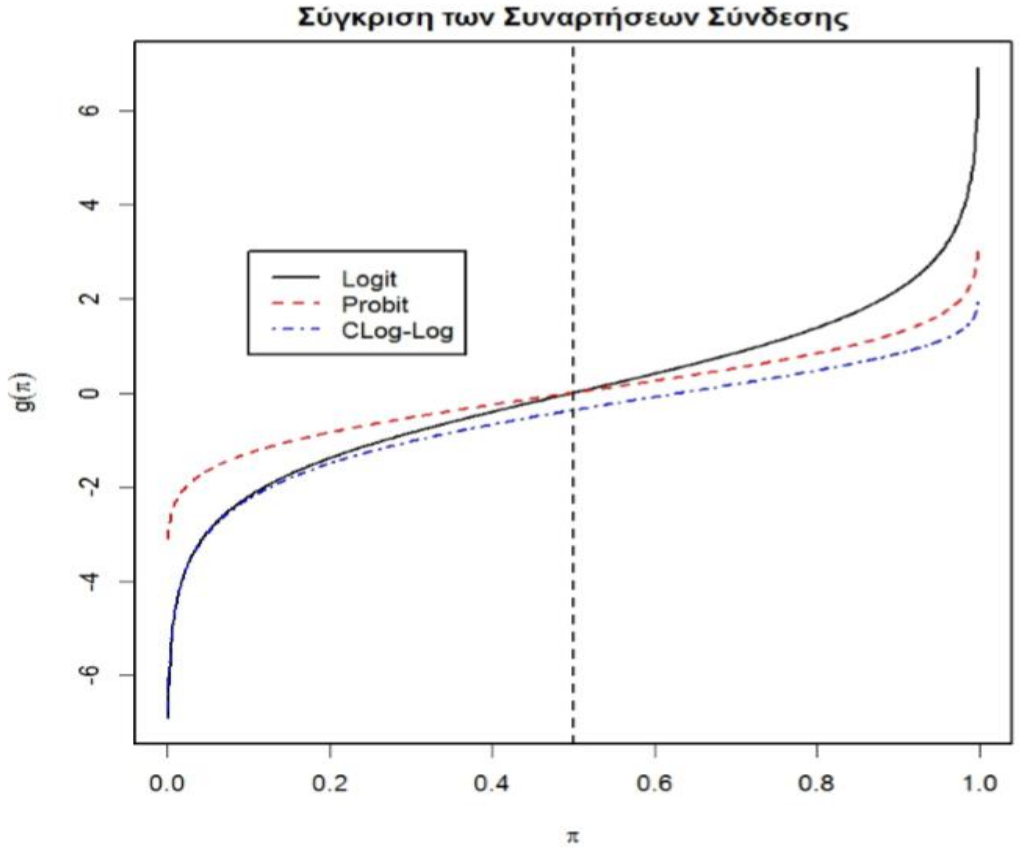
Αν τώρα έχουμε διωνυμική κατανομή, τότε έχουμε τρεις βασικές συναρτήσεις σύνδεσης:

- Logit: $\eta = \log \frac{\mu}{1-\mu}$
- Probit: $\eta = \Phi^{-1}(\mu)$, με Φ η συνάρτηση κατανομής της τυπικής κανονικής κατανομής.
- Complementary log-log: $\eta = \log(-\log(1-\mu))$.

Τέλος αν η = θ, με θ η κανονική παράμετρος, τότε έχουμε:

- Κανονική κατανομή: η = μ
- Poisson κατανομή: η = log(μ)
- Διωνυμική κατανομή: η = $\log \frac{p}{1-\mu}$

Γραφικά οι διαφορά των τριών βασικών συναρτήσεων σύνδεσης παρουσιάζονται στο παρακάτω διάγραμμα:



1.5.1. Εκτιμητική στα Γενικευμένα γραμμικά μοντέλα

Θεωρούμε $Y_1, \ldots, Y_N$ ανεξάρτητες τ.μ.. Επιδιώκουμε να εκτιμήσουμε τις παραμέτρους b τέτοιες ώστε:

$$E(y_i) = \mu_i \; και \;\; g(\mu_i) = x_i^T \cdot b.$$

<u>Εκτίμηση με την μέθοδο Newton – Raphson</u>

Η μέθοδος αυτή πρακτικά μας δίνει απάντηση στην λύση εξίσωσης μιας μεταβλητής, δηλαδή f(x) = 0, κάτι που δίνεται προσεγγιστικά από την σχέση:

$$x^{(m)} = x^{(m-1)} - \frac{f[x^{(m-1)}]}{f'[x^{(m-1)}]}.$$

Εμείς λοιπόν για την σχέση μας

$$f(y;\theta) = exp\left[\sum_{i=1}^{N} y_i \cdot b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)\right]$$

έχουμε παίρνοντας την ln – πιθανοφάνεια για κάθε Y_i :

$$l(\theta; y) = \sum_{i=1}^{N} y_i \cdot b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i).$$

Επειδή οι κατανομές που μας ενδιαφέρουν ανήκουν στην εκθετική οικογένεια κατανομών, ισχύει ότι το ολικό μέγιστο της $l(\theta; y)$ δίνεται μοναδικά από την λύση των εξισώσεων:

$$\frac{\partial l}{\partial \theta} = 0 \Leftrightarrow \frac{\partial l}{\partial b} = 0.$$

Αποδεικνύεται ότι :

$$U_j = \frac{\partial l}{\partial b_j} = \sum_{i=1}^{N} \frac{(y_i - \mu_i) \cdot x_{ij}}{Var(Y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i}, \text{όπου } x_{ij} \text{ το } j - \text{οστό στοιχείο του } x_i^T.$$

Όμως οι εξισώσεις $U_j = 0$ δεν είναι γραμμικές και συνεπώς δεν λύνονται με απλές μεθόδους, αλλά χρειάζονται κάποια επαναληπτική αριθμητική μέθοδο. Μια τέτοια μέθοδος λοιπόν είναι η Newton – Raphson.

Η μέθοδος αυτή μας δίνει την m-οστή προσέγγιση από την σχέση :

$$b^{(m)} = b^{(m-1)} - \left[\frac{\partial^2 l}{\partial b_j \partial \beta_k}\right]^{-1}_{b=b^{(m-1)}} \cdot U^{(m-1)},$$

Όπου $\left[\frac{\partial^2 l}{\partial b_j \partial b_k}\right]^{-1}_{b=b^{(m-1)}}$ είναι ο πίνακας των παραγώγων $2^{ης}$ της l για την τιμή $b = b^{(m-1)}$ και $U^{(m-1)}$ είναι το διάνυσμα των παραγώγων $1^{ης}$ τάξης $U_j = \frac{\partial l}{\partial b_j}$ για $b = b^{(m-1)}$.

1.5.2. Μελέτη καταλληλότητας του μοντέλου

Ψάχνουμε το μοντέλο στο οποίο όλες οι μεταβλητές θα είναι στατιστικά σημαντικές, κάτι το οποίο ελέγχεται μέσω των συντελεστών b.

Για να ελέγξουμε την επάρκεια του μοντέλου θα χρησιμοποιήσουμε μια δειγματική κατανομή της στατιστικής συνάρτησης D (deviance) η οποία σχετίζεται με την X^2_{N-p}.

Η στατιστική συνάρτηση deviance ή αλλιώς στατιστική συνάρτηση αναλογίας λογαριθμικής πιθανοφάνειας είναι η:

$$D = 2\left[l\left(\widehat{b_{max}}; y\right) - l\left(\hat{b}; y\right)\right].$$

Μετασχηματίζοντας αυτήν την συνάρτηση έχουμε:

$$D = 2[l(\widehat{b_{max}}; y) - l(b_{max}; y)] - 2[l(\hat{b}; y) - l(b; y)] + 2[l(b_{max}; y) - l(b; y)]$$

Το πρώτο μέρος της προηγούμενης σχέσης ακολουθεί X_m^2, με m να είναι ο αριθμός των παραμέτρων του πλήρους μοντέλου. Το δεύτερο μέρος ακολουθεί X_p^2, p είναι ο αριθμός των σημαντικών παραμέτρων και τέλος το τρίτο μέρος είναι μια σταθερή ποσότητα που είναι κοντά στο μηδέν όταν το μοντέλο των σημαντικών παραμέτρων περιγράφει τα δεδομένα τόσο καλά όσο το πλήρες μοντέλο. Το πλήρες μοντέλο είναι ένα γενικευμένο γραμμικό μοντέλο που χρησιμοποιεί την ίδια κατανομή με το μοντέλο των σημαντικών παραμέτρων και έχουν την ίδια συνάρτηση σύνδεσης. Όταν έχουμε καλή προσαρμογή των δεδομένων στο μοντέλο, έχουμε ότι:

$$D \sim X_{m-p}^2.$$

Η καταλληλότητα του μοντέλου κρίνεται στο αν η τιμή του D είναι κοντά στην μέση τιμή της κατανομής, δηλαδή αν το μοντέλο με p παραμέτρους περιγράφει καλά το σύνολο των N παρατηρήσεων με $D \sim X_{N-p}^2$ τότε D $\approx N - p$.

Για τον έλεγχο υποθέσεων μέσω της συνάρτησης deviance ορίζουμε ένα μοντέλο για κάθε υπόθεση και συγκρίνουμε τις στατιστικές συναρτήσεις της καλής προσαρμογής για τα μοντέλα αυτά. Τα μοντέλα θα πρέπει να έχουν ίδια κατανομή και ίδια συνάρτηση σύνδεσης. Συνεπώς θα έχουμε μηδενική υπόθεση H_0 και εναλλακτική H_α τέτοιες ώστε:

$$H_0: b_0 = \begin{bmatrix} b_1 \\ \vdots \\ b_q \end{bmatrix} \text{ έναντι } H_\alpha: b_\alpha = \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix}.$$

Από την D έχουμε:

$$\begin{aligned} \Delta D = D_0 - D_\alpha &= 2[l(b_{max}; y) - l(b_0; y)] - 2[l(b_{max}; y) - l(b_\alpha; y)] \\ &= 2[l(b_\alpha; y) - l(b_0; y)]. \end{aligned}$$

Αν τα μοντέλα περιγράφουν καλά τα δεδομένα, τότε έχουμε ότι:

$$D_0 \sim X_{N-q}^2 \text{ και } D_\alpha \sim X_{N-p}^2 \Leftrightarrow \Delta D \sim X_{p-q}^2.$$

Αν τώρα D ~X_{p-q}^2τότε το μοντέλο που προτιμάται είναι το H_0 λόγω απλότητας.

1.5.3. Υπόλοιπα

Τα υπόλοιπα έχουν την μορφή $\widehat{e_i} = y_i - \widehat{y_i}$. Μέσω αυτών ελέγχουμε τα χαρακτηριστικά του μοντέλου.

Τα υπόλοιπα είναι επίσης ένας δείκτης καταλληλότητας ενός μοντέλου. Στα γενικευμένα γραμμικά μοντέλα τα υπόλοιπα έχουν διάφορες μορφές:

- Υπόλοιπα deviance

$$r_D = sign(y-\mu)\sqrt{d_i}, \qquad \sum_{i=1}^{N} r_D^2 = D \; \kappa\alpha\iota \sum_{i=1}^{N} d_i = D.$$

Στην Poisson κατανομή $r_D = sign(y-\mu)\cdot\left[2\left(\log\left(\frac{y}{\mu}\right) - y + \mu\right)\right]^2$.

- Υπόλοιπα Pearson:

$$r_p = \frac{y_i - \mu_i}{\sqrt{var(y_i)}}$$

Χρησιμοποιούνται κυρίως για κανονική κατανομή, ειδάλλως είναι πολύ δύσχρηστα.

- Υπόλοιπα Anscobe:

  Ουσιαστικά αυτά τα υπόλοιπα ορίστηκαν χρησιμοποιώντας μια συνάρτηση A(y), όπου η συνάρτηση αυτή στο Y θα μας οδηγούσε σε μια κατανομή πολύ κοντά στην κανονική. Η συνάρτηση αυτήν είναι η:

$$A(\cdot) = \int \frac{d\mu}{var^{1/3}(\mu)}.$$

  Για την Poisson κατανομή έχουμε:

$$\int \frac{d\mu}{\mu^{1/3}} = \frac{3}{2}\mu^{2/3}.$$

  Τα Anscobe υπόλοιπα λοιπόν είναι:

$$r_A = \frac{\frac{3}{2}\left(y^{\frac{2}{3}} - \mu^{\frac{2}{3}}\right)}{\mu^{\frac{1}{6}}}.$$

1.5.4. Κριτήρια επιλογής μοντέλου

Στα γενικευμένα γραμμικά μοντέλα μπορούμε να χρησιμοποιήσουμε τα κριτήρια AIC και BIC, καθώς και διάφορους συντελεστές προσδιορισμού για την επιλογή βέλτιστου μοντέλου, τα οποία θα αναλυθούν στα επόμενα κεφάλαια.

Κεφάλαιο 2

Το γραμμικό μοντέλο

2.1. Απλή γραμμική παλινδρόμηση

Το μοντέλο της απλής γραμμικής παλινδρόμησης είναι της μορφής:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, με\ i = 1,2,\dots,n \quad όπου,$$

Y_i: είναι οι τιμές της εξαρτημένης μεταβλητής, η οποία είναι τυχαία μεταβλητή

X_i: είναι οι τιμές της ανεξάρτητης μεταβλητής που θεωρούνται ως γνωστές σταθερές

β_0, β_1: οι άγνωστες παράμετροι του μοντέλου όπου καλούνται συντελεστές παλινδρόμησης.

ε_i: τα σφάλματα του μοντέλου τα οποία είναι τυχαίες μεταβλητές

n: είναι το πλήθος των παρατηρήσεων (X_i, Y_i) που διαθέτουμε.

Το μοντέλο αυτό καλείται απλό γιατί περιέχει μόνο μια ανεξάρτητη μεταβλητή και γραμμικό γιατί είναι γραμμικό ως προς τις παραμέτρους.

<u>Υποθέσεις για τα σφάλματα</u>

1. $E(\varepsilon_i) = 0\ και\ \mathrm{Var}(\varepsilon_i) = \sigma^2$=σταθερή (ομοσκεδαστικότητα) για κάθε i=1,…,n
2. Είναι ανά δύο ασυσχέτιστες τ.μ. δηλαδή $\mathrm{cov}(\varepsilon_i, \varepsilon_j)$=0 ∀ i ≠ j
3. Τα σφάλματα ακολουθούν Κανονική κατανομή, δηλαδή $\varepsilon_i \sim N(0, \sigma^2)$

Σχόλια

- Η υπόθεση της κανονικότητας των σφαλμάτων είναι αρκετά ρεαλιστική (λόγω Κ.Ο.Θ.) δεδομένου ότι αποτελούν το αθροιστικό αποτέλεσμα επιμέρους σφαλμάτων.
- Αφού τα σφάλματα είναι ασυσχέτιστες τ.μ. και ισόνομες, τότε είναι και ανεξάρτητες.

Ας επιστρέψουμε όμως στο μοντέλο μας, αφού τα σφάλματα είναι ασυσχέτιστες τ.μ. τότε και τα Y_i είναι ανά δυο ασυσχέτιστες τ.μ. με

$E(Y_i) = E(\beta_0 + \beta_1 X_i + \varepsilon_i) = \beta_0 + \beta_1 X_i + E(\varepsilon_i) = \beta_0 + \beta_1 X_i$

$Var(Y_i) = Var(\beta_0 + \beta_1 X_i + \varepsilon_i) = Var(\varepsilon_i) = \sigma^2$

Η σχέση $E(Y_i) = \beta_0 + \beta_1 X_i$ καλείται συνάρτηση του μοντέλου ή ευθεία παλινδρόμησης.

Παρατηρήσεις

- $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i) = Y_i - E(Y_i)$, δηλαδή το τυχαίο σφάλμα ε_i ισούται με την κατακόρυφη απόκλιση της παρατηρούμενης τιμής Y_i από την άγνωστη ευθεία παλινδρόμησης
- $το\ \beta_1$ είναι η κλίση της ευθείας παλινδρόμησης και δείχνει την μεταβολή της μέσης τιμής της Y για κάθε μοναδιαία αύξηση της τιμής της X
- $το\ \beta_0$ είναι το σημείο στο οποίο η ευθεία παλινδρόμησης τέμνει τον άξονα των Y και δίνει τη μέση τιμή της Y για X=0.

Όμως, τα β_0 και β_1 όπως είπαμε είναι οι άγνωστες παράμετροι του μοντέλου μας, συνεπώς είναι αναγκαίο για την συνέχεια να εκτιμήσουμε στατιστικά την τιμή τους.
Με την μέθοδο ελαχίστων τετραγώνων θα πραγματοποιήσουμε αυτήν την εκτίμηση, δηλαδή οι εκτιμητές μας θα πρέπει να ελαχιστοποιούν την ποσότητα:

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Παίρνοντας μερικές παραγώγους σε αυτήν την σχέση ως προς β_0, β_1 και λύνοντας το σύστημα που προκύπτει έχουμε τους εκτιμητές:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$$

<u>Ερμηνεία των $\widehat{\beta_1}$ και $\widehat{\beta_0}$</u>

$\widehat{\beta_1}$: Για κάθε μοναδιαία αύξηση στην τιμή της X, η μέση τιμή της Y μεταβάλλεται κατά $\widehat{\beta_1}$

$\widehat{\beta_0}$: Η τιμή της $\widehat{\beta_0}$ εκτιμά τη μέση τιμή της Y όταν X=0 με την προϋπόθεση ότι η τιμή X=0 υπήρχε στις παρατηρήσεις βάσει των οποίων εκτιμήθηκε η ευθεία παλινδρόμησης.

Σχόλιο

Οι διαφορές $e_i = Y_i - \widehat{Y_i}$ καλούνται υπόλοιπα (residuals).

2.1.1. Ανάλυση διακύμανσης (ANOVA)

Το μέτρο μεταβλητότητας των τιμών της Y είναι η ποσότητα:

$$\sum_{\mathrm{i}=1}^{n}(Y_i - \bar{Y})^2$$

και καλείται ολική μεταβλητότητα των Y. Συμβολισμός SStot.

Το άθροισμα των τετραγώνων της παλινδρόμησης είναι η ποσότητα :

$$\sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2$$

και συμβολίζεται ως SSreg.

Το άθροισμα των τετραγώνων των υπολοίπων είναι η ποσότητα:

$$\sum_{\mathrm{i}=1}^{n}(Y_i - \widehat{Y}_i)^2$$

και συμβολίζεται ως SSres.

Έχουμε ότι:

SStot = SSreg + SSres

κάτι που προκύπτει από την σχέση:

$$Y_i - \bar{Y} = (Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \bar{Y})$$

υψώνοντας στο τετράγωνο και τα δυο μέλη και αθροίζοντας.

Σε κάθε άθροισμα τετραγώνων αντιστοιχεί ένας αριθμός που καλείται βαθμός/οι ελευθερίας και δείχνει το πλήθος των ανεξάρτητων συναρτήσεων των Y_i που απαιτούνται για τον υπολογισμό του αθροίσματος αυτού.

- SStot έχει n-1 β.ε. επειδή από τις διαφορές $Y_1 - \bar{Y}, \dots, Y_{\mathrm{n}} - \bar{Y}$ που χρειάζονται για τον υπολογισμό του οι n-1 είναι ανεξάρτητες αφού $\sum_{i=1}^{n}(Y_i - \bar{Y})$=0.

- SSreg $= \sum_{i=1}^{n}(\widehat{Y_i} - \bar{Y})^2 = \sum_{i=1}^{n}(\widehat{\beta_0} + \widehat{\beta_1}X_i - \bar{Y})^2 = \sum_{i=1}^{n}(\bar{Y} - \widehat{\beta_1}\bar{X} + \widehat{\beta_1}X_i - \bar{Y})^2 = \widehat{\beta_1}^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$ έχει έναν β.ε. γιατί μπορεί να υπολογιστεί με μια συνάρτηση των Y_i, την $\widehat{\beta_1}$.
- SSres έχει n-2 β.ε. γιατί
  $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(Y_i - \widehat{Y_i}) = 0$ και
  $\sum_{i=1}^{n} X_i e_i = \sum_{i=1}^{n} X_i(Y_i - \widehat{Y_i}) = 0.$ Άρα έχουμε:
  Ο πίνακας ανάλυσης διακύμανσης λοιπόν έχει την εξής μορφή:

Πηγή μεταβλητότητας	**Άθροισμα τετραγώνων**	**Βαθμοί ελευθερίας**	**Μέσα τετράγωνα**	**F-test**
Παλινδρόμηση	SSreg	1	MSreg	$F=\frac{MSreg}{MSres}$
Υπόλοιπα	SSres	n-2	MSres	
Ολική μεταβλητότητα	SStot	n-1		

Όπου

SSreg $= \widehat{\beta_1}^2 \left[\sum X_i^2 - \frac{(\sum X_i)^2}{n}\right]$, MSreg $= \frac{SSreg}{1}$

SSres = SStot – SSreg , MSres $= \frac{SSres}{n-2}$

SStot $= \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$

Σχόλιο

Γενικά για τους βαθμούς ελευθερίας (β.ε.) έχουμε:

- Β.ε. για SStot: (ο αριθμός των παρατηρήσεων) – 1
- Β.ε για SSreg: αριθμός ανεξάρτητων μεταβλητών στο μοντέλο
- Β.ε. για SSres: (αριθμός παρατηρήσεων) – (αριθμός παραμέτρων του μοντέλου)

2.1.2. Συντελεστής προσδιορισμού

$$R^2 = \frac{SSreg}{SStot} = 1 - \frac{SSres}{SStot}$$

Η ποσότητα R^2 καλείται συντελεστής προσδιορισμού και εκφράζει το ποσοστό της μεταβλητότητας των τιμών της Y που εξηγείται από την παλινδρόμηση.

Προφανώς ισχύει ότι $0 \leq R^2 \leq 1$ και είναι «καθαρός» αριθμός, δηλαδή ανεξάρτητος μονάδων.

Σχόλιο

Το R^2 δεν συνδέεται με καλό ή κακό μοντέλο. Στην ιδανική περίπτωση όπου SSres =0 άρα $R^2 = 1$ δεν σημαίνει πως έχουμε σωστό μοντέλο αλλά έχουμε άριστο προγνωστικό παράγοντα.

2.1.3. F-test Έλεγχος σημαντικότητας της παλινδρόμησης

Έχουμε SSreg = $\widehat{\beta_1}^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$, ανεξάρτητο του SSres. Επίσης, $\frac{\text{SSreg}}{\sigma^2} \sim X_1^2$ όταν β_1=0

και $\frac{\text{SSres}}{\sigma^2} \sim X_{\text{n}-2}^2$. Άρα έχουμε:

$$\frac{\frac{\text{SSreg}}{\sigma^2}\Big/_1}{\frac{\text{SSres}}{\sigma^2}\Big/_{\text{n}-2}} = \frac{MSreg}{MSres} \sim F_{1,n-2}, \text{ όταν } \beta_1=0$$

Ο έλεγχος H_0: β_1=0 έναντι H_α: $\beta_1 \neq 0$ επιπέδου σημαντικότητας α μπορεί να γίνει με στατιστική συνάρτηση ελέγχου F = $\frac{MSreg}{MSres} \sim F_{1,n-2}$ υπό την μηδενική υπόθεση και κρίσιμη περιοχή F>$F_{1,n-2,\alpha}$

2.1.4. Συσχέτιση

Για δυο τ.μ. X και Y ο «θεωρητικός» συντελεστής συσχέτισης ρ(X,Y) ορίζεται ως εξής:

ρ(X,Y)= $\frac{\text{cov(X,Y)}}{\sqrt{Var(X)\cdot Var(Y)}}$. Αποτελεί μέτρο γραμμικής σχέσης μεταξύ των τ.μ. X και Y και ρ∈ $[-1,1]$.

- Όσο πλησιάζει το ρ στα άκρα του διαστήματος γίνεται ισχυρότερη η γραμμική συσχέτιση ενώ προς το κέντρο του διαστήματος η γραμμική συσχέτιση είναι πιο ασθενής.
- Για ρ = 1 έχουμε τέλεια θετική συσχέτιση και για ρ = − 1 έχουμε τέλεια αρνητική συσχέτιση.
- Για ρ = 0 οι X, Y δεν σχετίζονται γραμμικά και ονομάζονται ασυσχέτιστες.

Ωστόσο, στην πράξη χρησιμοποιούμε τον δειγματικό συντελεστή συσχέτισης του Pearson r(X,Y), όπου

$$r(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2\cdot\sum_{i=1}^{n}(X_i-\bar{X})^2}} = \frac{\sum X_\iota Y_\iota - \frac{\sum X_\iota \sum Y_\iota}{n}}{\sqrt{[\sum X_i^2 - \frac{(\sum X_i)^2}{n}]\cdot[\sum Y_i^2 - \frac{(\sum Y_i)^2}{n}]}}$$

<u>Συσχέτιση και παλινδρόμηση</u>

Για την σχέση μεταξύ του r και του $\widehat{\beta_1}$ έχουμε:

$$r(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2\cdot\sum_{i=1}^{n}(X_i-\bar{X})^2}} = \frac{\widehat{\beta_1}\cdot\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2\cdot\sum_{i=1}^{n}(X_i-\bar{X})^2}}$$

$$= \widehat{\beta_1}\cdot\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} = \widehat{\beta_1}\cdot\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2/_{n-1}}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2/_{n-1}}} = \widehat{\beta_1}\cdot\frac{s_x}{s_y},$$

όπου

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \text{ και } s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$$

Άσκηση:

Για να διερευνηθεί η σχέση μεταξύ ηλικίας Χ και πίεσης αίματος Υ σε γυναίκες συλλέγονται 12 παρατηρήσεις:

X_i: 36 38 42 42 47 49 55 56 60 63 68 72

Y_i: 118 115 125 140 128 145 150 147 155 149 152 160

Α) Να κατασκευαστεί το διάγραμμα διασποράς.

Β) Υιοθετώντας το μοντέλο απλής γραμμικής παλινδρόμησης να βρεθεί η εκτιμώμενη ευθεία παλινδρόμησης.

Γ) Να δοθεί ο πίνακας anova. Τι ποσοστό μεταβλητότητας των τιμών της Υ εξηγείται από την παλινδρόμηση;

Δ) Να εκτιμηθεί η διασπορά των σφαλμάτων.

Ε) Να δοθεί ένα 95% διάστημα εμπιστοσύνης για τα β_0 και β_1.

Στ) Να γίνει ο έλεγχος σημαντικότητας της παλινδρόμησης χρησιμοποιώντας:

i) T-test.

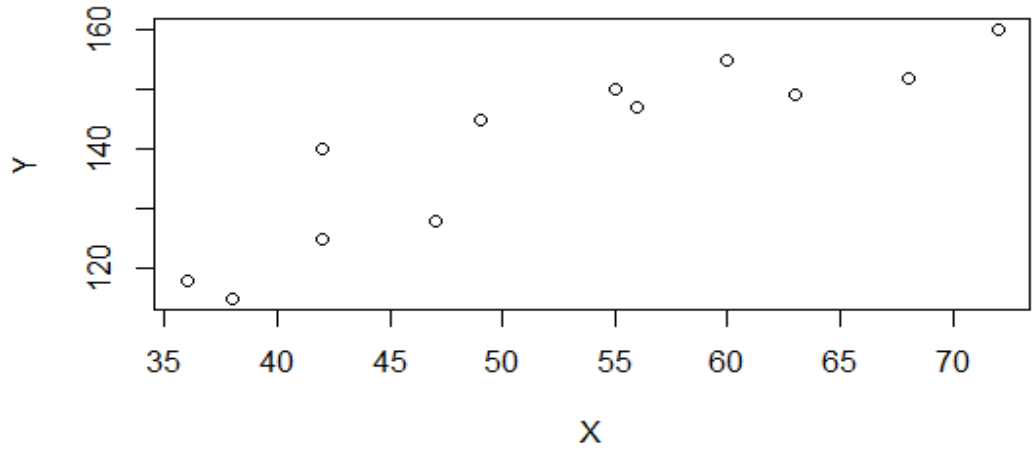
ii) Το δ.ε. για το β_1.

iii) F-test.

Ζ) Να γίνει ο έλεγχος $H_0: \beta_0 = 85$ *έναντι* $H_\alpha: \beta_0 \neq 85$.

Απάντηση:

A)



B)

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
36	118	1296	13924	4248
38	115	1444	13225	4370
42	125	1764	15625	5250
42	140	1764	19600	5880
47	128	2209	16384	6016
49	145	2401	21025	7105
55	150	3025	22500	8250
56	147	3136	21609	8232
60	155	3600	24025	9300
63	149	3969	22201	9387
68	152	4624	23104	10336
72	160	5184	25600	11520
628	**1684**	**34416**	**238822**	**89894**

Οπότε έχουμε:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum X_i \cdot \sum Y_i}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}} = \frac{89894 - \frac{628 \cdot 1684}{12}}{34416 - \frac{624^2}{12}} = 1.138$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{1684}{12} - 1.138 \cdot \frac{628}{12} = 80.778.$$

Άρα $\hat{Y} = 80.778 + 1.138 \cdot X$.

Συνεπώς, για κάθε αύξηση στη ηλικία της γυναίκας κατά ένα έτος κατά μέσο όρο η πίεση του αίματος αυξάνει κατά 1.138.

Γ)

$$SSreg = \hat{\beta}_1{}^2 \sum (X_i - \bar{X})^2 = \hat{\beta}_1{}^2 \left[\sum X_i^2 - \frac{(\sum X_i)^2}{n} \right] = 2008.182$$

$$SStot = \sum (Y_i - \bar{Y})^2 = \left[\sum Y_i^2 - \frac{(\sum Y_i)^2}{n} \right] = 2500.667$$

$$SSres = SStot - SSreg = 2500.667 - 2008.182 = 492.485.$$

Άρα έχουμε:

Πηγή μεταβλητότητας	**Άθροισμα τετραγώνων**	**Βαθμοί ελευθερίας**	**Μέσα τετράγωνα**	**F-test**
Παλινδρόμηση	SSreg=2008.182	1	MSreg=2008.182	F=40.777
Υπόλοιπα	SSres=492.485	n-2=10	MSres=49.2485	
Ολική μεταβλητότητα	SStot=2500.667	n-1=11		

$$R^2 = \frac{SSreg}{SStot} = 0.8031,$$

συνεπώς το 80.31% της μεταβλητότητας των τιμών της Y εξηγείται από την παλινδρόμηση.

Δ) $s^2 = \hat{\sigma}^2 = MSres = 49.2485.$

Ε) Ένα 95% δ.ε. για το β_0:

$$\hat{\beta}_0 \pm t_{n-2,\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right)}$$

$$= 80.778 \pm 2.228 \cdot \sqrt{49.2485 \left(\frac{1}{12} + \frac{\left(\frac{628}{12}\right)^2}{1550.667} \right)}$$

$$= 80.778 \pm 21.262 \rightarrow [59.516, 102.04].$$

Έχουμε 95% βεβαιότητα ότι η τιμή του β_0 είναι μεταξύ των τιμών 59.516 και 102.04.

Ένα 95% δ.ε. για το β_1:

$$\hat{\beta}_1 \pm t_{n-2,\frac{\alpha}{2}}\sqrt{\frac{\hat{\sigma}^2}{\sum(X_i-\bar{X})^2}} = 1.138 \pm 2.228\sqrt{\frac{49.2485}{1550.667}} = 1.138 \pm 0.397$$

$$\rightarrow [0.741, 1.535].$$

Έχουμε 95% βεβαιότητα ότι για κάθε αύξηση κατά 1 έτος στην ηλικία γυναίκας, η πίεση του αίματος θα αυξάνει κατά μέσο όρο από 0.741 εως 1.535.

Στ) $H_0: \beta_1 = 0$ *έναντι* $H_\alpha: \beta_1 \neq 0$.

i) $t_{\beta_1} = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\hat{\sigma}^2}{\sum(X_i-\bar{X})^2}}} = \frac{1.138}{0.178} = 6.393$ *και κ.π.* $|t_{\beta_1}| > t_{n-2,\frac{\alpha}{2}} = t_{10,0.025} =$ 2.228.

Όμως $|t_{\beta_1}| = 6.393 > 2.228$.

Άρα απορρίπτεται η μηδενική υπόθεση, οπότε η παλινδρόμηση είναι στατιστικά σημαντική.

ii) $0 \notin [0.741, 1.535]$. Συνεπώς η H_0 απορρίπτεται.

iii) $F = 40.776 > 4.96 = F_{1,10,0.05} = F_{1,n-2,\alpha}$. Άρα η H_0 απορρίπτεται.

Ζ) Για τον $H_0: \beta_0 = 85$ *έναντι* $H_\alpha: \beta_0 \neq 85$:

$$t_{\beta_0} = \frac{\hat{\beta}_0 - 85}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i-\bar{X})^2}\right)}} = \frac{80.778 - 85}{9.543} = -0.442 \text{ } \textit{και κ.π.} |t_{\beta_0}| > t_{n-2,\frac{\alpha}{2}}$$

$$= t_{10,0.025} = 2.228.$$

Όμως $|t_{\beta_0}| = 0.442 < 2.228 = t_{10,0.025}$. Άρα δεν απορρίπτουμε την μηδενική υπόθεση, δηλαδή $\beta_0 = 85$.

2.2. Πολλαπλή γραμμική παλινδρόμηση

2.2.1. Μέση τιμή και διακύμανση πίνακα

Αν $\mathcal{X} = (X_1 \quad, \ldots, \quad X_p)$ είναι ένα τυχαίο διάνυσμα, τότε:

$$E(\mathcal{X}) = E\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} = \tilde{\mu}, \text{ όπου } E(X_i) = \mu_i, i = 1, \ldots, p$$

$$Var(\mathcal{X}) = cov(\mathcal{X}, \mathcal{X}) = cov[X_i, X_j] = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1\rho} \\ \vdots & \ddots & \vdots \\ \sigma_{\rho 1} & \cdots & \sigma_{\rho\rho} \end{bmatrix} = \Sigma,$$

όπου $\sigma_{ij} = \text{cov}(X_i, X_j) = \text{E}[(X_i - \mu_i)\cdot(X_j - \mu_j)] = \text{E}(X_i \cdot X_j) - \text{E}(X_i) \cdot \text{E}(X_j)$

και

$$\sigma_{ii} = \text{cov}(X_i, X_i) = \text{E}(X_i - \mu_i)^2 = E(X_i^2) - [E(X_i)]^2 = Var(X_i) = \sigma_i^2$$

Ο πίνακας Σ ονομάζεται πίνακας διακυμάνσεων - συνδιακυμάνσεων του τυχαίου διανύσματος $\mathcal{X}$ και είναι συμμετρικός αφού

$\sigma_{ij} = \text{cov}(X_i, X_j) = \text{cov}(X_j, X_i) = \sigma_{ji}$.

2.2.2. Τετραγωνικές μορφές

Μια συνάρτηση της μορφής:

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_{ij} \cdot X_i \cdot X_j = \sum_{i=1}^{n} \alpha_{ii} \cdot X_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_{ii} \cdot X_i \cdot X_j, \quad \text{με } i \neq j$$

όπου α_{ij}, i=1,…,n σταθερές και X_i μεταβλητές, καλείται τετραγωνική μορφή.

Αν A=[(α_{ij})] είναι ένας πίνακας τάξης n τότε η παραπάνω τετραγωνική μορφή γράφεται ως εξής: $Q = \mathcal{X}^T A \mathcal{X}$.

Ορισμός: Ένας συμμετρικός πίνακας A καλείται θετικά ορισμένος αν $\mathcal{X}^T A \mathcal{X} > 0$.

Πρόταση: Αν A συμμετρικός πίνακας τάξης n και $\mathcal{X}$ τυχαίο διάνυσμα n× 1 με μέση τιμή μ και διασπορά Σ, τότε :

$\text{E}[\mathcal{X}^T \cdot A \cdot \mathcal{X}] = \text{tr}(A\cdot\Sigma) + \tilde{\mu}^T \cdot A \cdot \tilde{\mu}$

Πρόταση: Έστω A συμμετρικός πίνακας τάξης n.

A) Αν $\mathcal{X}$ είναι τυχαίο διάνυσμα με $\text{E}(\mathcal{X}) = \tilde{\mu}$ και $\text{Var}(\mathcal{X}) = \Sigma$, τότε

$\text{cov}(\mathcal{X}, \mathcal{X}^T A \mathcal{X}) = 2\cdot\Sigma\cdot A\cdot \tilde{\mu}$

B) Αν $X_1, \ldots, X_n$ τυχαίο δείγμα από $N(\mu, \sigma^2)$, τότε:

$$\text{Var}(\mathcal{X}^T A \mathcal{X}) = 2 \cdot \sigma^4 \cdot \text{tr}(A^2) + 4 \cdot \sigma^2 \cdot \widetilde{\mu^T} \cdot A^2 \cdot \tilde{\mu}, \text{ με } \tilde{\mu} = \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix} \text{ διάστασης n× 1.}$$

2.2.3. Πολυδιάστατη κανονική κατανομή

Συνάρτηση πυκνότητας πιθανότητας μονοδιάστατης κανονικής κατανομής:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(X-\mu)^2}{2\sigma^2}} \cdot I_{\mathbb{R}}(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(X-\mu)\cdot(\sigma^2)^{-1}\cdot(X-\mu)} I_{\mathbb{R}}(x)$$

όπου $\text{E}(X) = \mu$ και $\text{Var}(X) = \sigma^2$.

Σε αντιστοιχία με την παραπάνω περίπτωση έχουμε για την ρ-διάστατη κανονική κατανομή:

$$f(\tilde{x}) = \frac{1}{(\sqrt{2\pi})^{\rho} \cdot |\Sigma|} \cdot \mathrm{e}^{-(\tilde{x}-\tilde{\mu})^T \cdot \Sigma^{-1} \cdot (\tilde{x}-\tilde{\mu})/2} \cdot I_{\mathbb{R}^{\rho}}(\mathrm{x})$$

όπου $E(\tilde{X}) = \tilde{\mu}$ και $Var(\tilde{X}) = \Sigma$, με Σ ρxρ συμμετρικός θετικά ορισμένος πίνακας και |Σ| η ορίζουσά του.

Πρόταση: Έστω $\tilde{X}$ ένα ρx1 τυχαίο διάνυσμα με $\tilde{X} \sim N_{\rho}(\tilde{\mu}, \Sigma)$ και έστω $\tilde{\alpha}$ ένα ρx1 διάνυσμα σταθερών και A ένας κxρ, με κ≤ ρ, πίνακας σταθερών με rank(A) = κ. Τότε:

i. $z = \tilde{\alpha}^T \cdot \tilde{X} \sim N(\tilde{\alpha}^T \cdot \tilde{\mu}, \tilde{\alpha}^T \cdot \Sigma \cdot \tilde{\alpha})$

ii. $\tilde{z} = A \cdot \tilde{X} \sim N_{\kappa}(A \cdot \tilde{\mu}, A \cdot \Sigma \cdot A^T)$

2.2.4. Μη κεντρικές X^2, t, F κατανομές

1. Αν $z_1, \ldots, z_n \ (= z^T)$ είναι τυχαίο δείγμα από N(0,1) τότε:

$$W = \sum_{i=1}^{n} z_i^2 = \tilde{z}^T \cdot \tilde{z} \sim X_n^2$$

❖ Αν $y_1, \ldots, y_n$ ανεξάρτητες τ.μ. με $y_i \sim N(\mu_i, 1)$ τότε

a) $y_i - \mu_i \sim N(0,1)$ και είναι ανεξάρτητες. Άρα

$$\sum_{i=1}^{n} (y_i - \mu_i)^2 = (\tilde{y} - \tilde{\mu})^T \cdot (\tilde{y} - \tilde{\mu}) \sim X_n^2$$

b) Η κατανομή της τ.μ.

$$V = \sum_{i=1}^{n} y_i^2 = \tilde{y}^T \cdot \tilde{y}$$

καλείται μη κεντρική X_n^2 και παράμετρο μη κεντρικότητας λ, όπου

$$\lambda = \frac{1}{2} \sum_{i=1}^{n} \mu_i^2 = \frac{1}{2} \tilde{\mu}^T \cdot \tilde{\mu}$$

και συμβολίζουμε $V \sim X_n^2(\lambda)$.

2. Αν $U \sim X_p^2$ και $V \sim X_q^2$ και U,V ανεξάρτητες τ.μ. τότε:

$$W = \frac{U/p}{V/q} \sim F_{p,q}.$$

❖ Αν $U \sim X_p^2(\lambda)$ και $V \sim X_q^2$ και U,V ανεξάρτητες τ.μ. τότε η κατανομής της

$$W = \frac{U/p}{V/q} \sim F_{p,q}(\lambda)$$

καλείται μη κεντρική F με p,q β.ε. και μη κεντρική παράμετρο λ.

3. Αν $Z \sim N(0,1)$ και $U \sim X_p^2$ και Z, U ανεξάρτητες τ.μ. τότε:

$$\mathrm{T} = \frac{Z}{\sqrt{\frac{\mathrm{U}}{p}}} \sim t_p$$

❖ Αν Y~N(μ, σ^2) και U~ X^2_{p} και Y,U ανεξάρτητες τ.μ. τότε η κατανομή της

$$\mathrm{T} = \frac{Y/\sigma}{\sqrt{\mathrm{U}/p}} \sim t_p\left(\frac{\mu}{\sigma}\right)$$

και καλείται t με p β.ε. και μη κεντρική παράμετρο $\frac{\mu}{\sigma}$.

2.2.5. Κατανομές τετραγωνικών μορφών

Πρόταση: Αν $\tilde{\mathrm{Y}} \sim N_{\mathrm{n}}(\tilde{\mu}, \Sigma)$ τότε $(\tilde{Y} - \tilde{\mu})^T \cdot \Sigma^{-1} \cdot (\tilde{Y} - \tilde{\mu}) \sim X^2_n$.

Θεώρημα 1: Αν $\tilde{\mathrm{Y}} \sim N_p(\tilde{\mu}, \Sigma)$ και A pxp συμμετρικός πίνακας σταθερών με rank(A) = r, r ≤ p, τότε:

$\tilde{Y}^T \cdot A \cdot \tilde{Y} \sim X^2_{\mathrm{r}}(\lambda)$, όπου λ= $\frac{1}{2}\tilde{\mu}^T \cdot \tilde{\mu}$ αν και μόνον αν ΑΣ είναι ταυτοδύναμος.

Θεώρημα 2:

Έστω B κxρ πίνακας σταθερών και A ρxρ συμμετρικός πίνακας σταθερών και $\tilde{\mathrm{Y}} \sim N_{\rho}(\tilde{\mu}, \Sigma)$. Τότε B · Ỹ και $\tilde{Y}^T \cdot A \cdot \tilde{Y}$ είναι ανεξάρτητα αν-ν $\mathrm{B} \cdot \Sigma \cdot \mathrm{A} = \mathbb{O}_{\kappa \mathrm{x} \rho}$.

Θεώρημα 3:

Αν A και B είναι ρxρ συμμετρικοί πίνακες σταθερών και $\tilde{\mathrm{Y}} \sim N_{\rho}(\tilde{\mu}, \Sigma)$, τότε

$\tilde{Y}^T \cdot A \cdot \tilde{Y}$ και $\tilde{Y}^T \cdot B \cdot \tilde{Y}$είναι ανεξάρτητα αν-ν $\mathrm{A} \cdot \Sigma \cdot \mathrm{B} = \mathbb{O}_{\rho \mathrm{x} \rho}$.

2.2.6. Μοντέλο πολλαπλής γραμμικής παλινδρόμησης

$Y_{\mathrm{i}} = \beta_0 + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \ldots + \beta_{p-1} \cdot X_{i,p-1} + \varepsilon_i$ με i=1,…,n

Y_{i} = η τιμή της εξαρτημένης μεταβλητής κατά την i-οστή επανάληψη του πειράματος,

X_{ij} = η i-οστή παρατήρηση για την j ανεξάρτητη μεταβλητή του μοντέλου, δεν είναι τυχαίες μεταβλητές,

β_{i} = οι άγνωστες παράμετροι του μοντέλου,

ε_{i} = τα τυχαία σφάλματα για τα οποία εισάγουμε τις υποθέσεις:

1. $\mathrm{E}(\varepsilon_i) = 0$ και $\mathrm{Var}(\varepsilon_{\mathrm{i}}) = \sigma^2$
2. $\mathrm{Cov}(\varepsilon_{\mathrm{i}}, \varepsilon_j) = 0 \ \forall \ \mathrm{i} \neq \mathrm{j}$.

Σε μορφή πινάκων το μοντέλο πολλαπλής παλινδρόμησης γίνεται:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{\mathrm{n}} \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1,p-1} \\ & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{n,p-1} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{\mathrm{p}-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{\mathrm{n}} \end{bmatrix}$$

ή $\tilde{Y} = \mathrm{X} \cdot \tilde{\beta} + \tilde{\varepsilon}$. Ο πίνακας X καλείται πίνακας σχεδιασμού.

Για την εκτίμηση του $\tilde{\beta}$ ελαχιστοποιούμε την ποσότητα

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \tilde{\varepsilon}^T \cdot \tilde{\varepsilon}$$

Παίρνουμε τις εξισώσεις : $X^T \cdot \mathrm{X} \cdot \tilde{\beta} = X^T \cdot \tilde{Y}$ οι οποίες καλούνται κανονικές εξισώσεις,

$$X^T \cdot \mathrm{X} = \begin{bmatrix} n & \sum_{i=1}^{n} X_{i1} & \cdots & \sum_{i=1}^{n} X_{i,p-1} \\ \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i1}^2 & & \sum_{i=1}^{n} X_{i1} \cdot X_{i,p-1} \\ & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} X_{i,p-1} & \sum_{i=1}^{n} X_{i1} \cdot X_{i,p-1} & \cdots & \sum_{i=1}^{n} X_{i,p-1}^2 \end{bmatrix}$$

$X^T \cdot \tilde{Y} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{i1} \cdot Y_i \\ \vdots \\ \sum_{i=1}^{n} X_{i,p-1} \cdot Y_i \end{bmatrix}$. Άρα με την προϋπόθεση ότι ο πίνακας $(X^T \cdot \mathrm{X})^{-1}$ υπάρχει έχουμε ότι:

$$\hat{\tilde{\beta}} = (X^T \cdot \mathrm{X})^{-1} \cdot X^T \cdot \tilde{Y}.$$

Ερμηνεία των παραμέτρων του μοντέλου:

Η παράμετρος β_{i} i=1,…, p-1, δείχνει την μεταβολή στην μέση τιμή της Y αν η αντίστοιχη ανεξάρτητη μεταβλητή X_i αυξηθεί κατά μια μονάδα και οι τιμές των υπολοίπων παραμείνουν σταθερές.

Εκτιμώμενο μοντέλο και υπόλοιπα

Η εκτίμηση για την τιμή της Y για κάθε i = 1,…,n είναι

$$\widehat{Y_{\mathrm{i}}} = \widehat{\beta_0} + \widehat{\beta_1} \cdot X_{\mathrm{i}1} + \ldots + \widehat{\beta_{\mathrm{p}-1}} \cdot X_{i,p-1} \iff \hat{\tilde{Y}} = X \cdot \hat{\tilde{\beta}}$$

Άρα, αφού το i-οστό υπόλοιπο είναι $\mathrm{e}_i = Y_i - \widehat{Y_{\mathrm{i}}}$, έχουμε:

$\tilde{\mathrm{e}} = \tilde{Y} - \hat{\tilde{Y}} = \tilde{Y} - X \cdot \hat{\tilde{\beta}} = \tilde{Y} - X \cdot (X^T \cdot X)^{-1} \cdot X^T \cdot \tilde{Y} = [I_{\mathrm{n}} - X \cdot (X^T \cdot \mathrm{X})^{-1} \cdot X^T] \cdot \tilde{Y} = (I_n - P) \cdot \tilde{Y},$

Όπου P = $X \cdot (X^T \cdot X)^{-1} \cdot X^T$, ο πίνακας hat-matrix.

Ο hat-matrix είναι:

- Συμμετρικός, αφού

$$P^T = (X \cdot (X^T \cdot X)^{-1} \cdot X^T)^T = X \cdot [(X^T \cdot X)^{-1}]^T \cdot X^T = X \cdot [(X^T \cdot X)^T]^{-1} \cdot X^T$$
$$= X \cdot (X^T \cdot X)^{-1} \cdot X^T = P$$

- Ταυτοδύναμος, αφού
  $P^2 = P \cdot P = X \cdot (X^T \cdot X)^{-1} \cdot X^T \cdot X \cdot (X^T \cdot X)^{-1} \cdot X^T = X \cdot (X^T \cdot X)^{-1} \cdot X^T = P$.

Το P_{ii} στοιχείο του hat-matrix καλείται μόχλευση. Μεγάλη τιμή P_{ii} σημαίνει πως η i-οστή παρατήρηση έχει τιμή μακριά από το κέντρο των παρατηρήσεων και πιθανά να έχει μεγάλη επίδραση στο προσαρμοζόμενο μοντέλο. Μια παρατήρηση με μεγάλη επίδραση καλείται επηρεάζουσα.

Πρόταση:

Για το μοντέλο $\tilde{Y} = X \cdot \tilde{\beta} + \tilde{\varepsilon}$ με rank(X) = p, όπου $E(\tilde{\varepsilon}) = \mathbb{0}$ και $Var(\tilde{\varepsilon}) = \sigma^2 \cdot I_n$ ισχύει ότι: $E(\hat{\tilde{\beta}}) = \tilde{\beta}$ και $Var(\hat{\tilde{\beta}}) = \sigma^2 \cdot (X^T \cdot X)^{-1}$.

Απόδειξη

$E(\hat{\tilde{\beta}}) = E((X^T \cdot X)^{-1} \cdot X^T \cdot \tilde{Y}) = (X^T \cdot X)^{-1} \cdot X^T \cdot E(\tilde{Y}) = (X^T \cdot X)^{-1} \cdot X^T \cdot X \cdot \tilde{\beta} = \tilde{\beta}$

$Var(\hat{\tilde{\beta}}) = Var((X^T \cdot X)^{-1} \cdot X^T \cdot \tilde{Y}) = (X^T \cdot X)^{-1} \cdot X^T \cdot Var(\tilde{Y}) \cdot [(X^T \cdot X)^{-1} \cdot X^T]^T =$
$(X^T \cdot X)^{-1} \cdot X^T \cdot \sigma^2 \cdot I_n \cdot X \cdot (X^T \cdot X)^{-1} = \sigma^2 \cdot (X^T \cdot X)^{-1}$.

BLUE εκτιμητές (Best Linear Unbiased Estimates)

Είναι οι γραμμικοί αμερόληπτοι εκτιμητές με την ελάχιστη διακύμανση μεταξύ των γραμμικών αμερόληπτων εκτιμητών.

Θεώρημα Gauss – Markov

Έστω το μοντέλο $\tilde{Y} = X \cdot \tilde{\beta} + \tilde{\varepsilon}$, όπου X nxp πίνακας με rank(X) = p και $\tilde{\varepsilon}$ τυχαίο διάνυσμα με $E(\tilde{\varepsilon}) = 0$ και $Var(\tilde{\varepsilon}) = \sigma^2 \cdot I_n$. Τότε ο Ε.Ε.Τ. του $\tilde{\beta}, o\ \hat{\tilde{\beta}}$, είναι blue εκτιμητής του $\tilde{\beta}$.

Θεώρημα:

Αν $\tilde{Y} \sim N_n(X \cdot \tilde{\beta}, \sigma^2 \cdot I_n)$, όπου X είναι ένας nxp πίνακας με rank(X) = p τότε:

1. $\hat{\tilde{\beta}} \sim N_p(\tilde{\beta}, \sigma^2 \cdot (X^T \cdot X)^{-1})$
2. $\frac{(\hat{\tilde{\beta}} - \tilde{\beta})^T \cdot X^T \cdot X \cdot (\hat{\tilde{\beta}} - \tilde{\beta})}{\sigma^2} \sim X^2_p$
3. $\hat{\tilde{\beta}}$ ανεξάρτητο του SSres
4. $\frac{SSres}{\sigma^2} \sim X^2_{n-p}$.

2.2.7. Έλεγχος υποθέσεων και διαστήματα εμπιστοσύνης

Αν $\tilde{\varepsilon} \sim N_n(\mathbb{0}, \sigma^2 \cdot I_n)$ έχουμε ότι $\hat{\tilde{\beta}} \sim N_{\mathrm{p}}(\tilde{\beta}, \sigma^2 \cdot (X^T \cdot X)^{-1})$, άρα

$\widetilde{\beta_{\mathrm{i}}} \sim N_p(\beta_i, \sigma^2 \cdot c_{i+1,i+1})$, i=0,…,p-1 και $c_{i+1,i+1}$ είναι το $(i+1, i+1)$ στοιχείο του πίνακα $(X^T \cdot X)^{-1}$.

Οπότε έχουμε $\frac{\widehat{\beta_{\mathrm{i}}} - \beta_i}{\sigma \cdot \sqrt{c_{i+1,i+1}}} \sim N(0,1)$ και $\frac{SSres}{\sigma^2} \sim X^2_{n-\mathrm{p}}$, SSres και $\hat{\tilde{\beta}}$ ανεξάρτητα. Άρα:

$$\frac{\widehat{\beta_{\mathrm{i}}} - \beta_i \Big/ \sigma \cdot \sqrt{c_{i+1,i+1}}}{\sqrt{\frac{SSres}{\sigma^2} \Big/ \mathrm{n} - \mathrm{p}}} = \frac{\widehat{\beta_{\mathrm{i}}} - \beta_i}{\sqrt{MSres \cdot c_{i+1,i+1}}} = \frac{\widehat{\beta_{\mathrm{i}}} - \beta_i}{s \cdot \sqrt{c_{i+1,i+1}}} \sim t_{n-p}$$

όπου $s^2 = \mathrm{MSres}\ (= \widehat{\sigma^2})$

Ο έλεγχος $H_0 : \beta_{\mathrm{i}} = \beta_{\mathrm{i}}^*$, όπου β_{i}^* είναι μια δοσμένη τιμή γίνεται με την στατιστική συνάρτηση ελέγχου $\mathrm{T} = \frac{\widehat{\beta_{\mathrm{i}}} - \beta_{\mathrm{i}}^*}{s \cdot \sqrt{c_{i+1,i+1}}}$, όπου όταν ισχύει η H_0 ακολουθεί t_{n-p} και κρίσιμη περιοχή μεγέθους α: $|\mathrm{t}| > t_{n-p,\frac{\alpha}{2}}$ $\gamma\iota\alpha$ H_α: $\beta_{\mathrm{i}} \neq \beta_{\mathrm{i}}^*$.

Ένα 100(1-α)% διάστημα εμπιστοσύνης για το β_{i}:

$$\widehat{\beta_{\mathrm{i}}} \pm t_{n-p,\frac{\alpha}{2}} \cdot\ s \cdot \sqrt{c_{i+1,i+1}}.$$

2.2.8. Έλεγχος σημαντικότητας της παλινδρόμησης

Έστω $\tilde{\mathrm{b}} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{\mathrm{p}-1} \end{pmatrix}$. Ο έλεγχος σημαντικότητας της παλινδρόμησης είναι:

H_0: $\tilde{\mathrm{b}} = \mathbb{0}$ έναντι H_α: : $\tilde{\mathrm{b}} \neq \mathbb{0}$ γίνεται με σ.σ.ε. $F = \frac{MSreg}{MSres}$, η οποία όταν ισχύει η H_0 ακολουθεί $\mathrm{F}_{p-1,\mathrm{n}-\mathrm{p}}$ και κ.π. επιπέδου σημαντικότητας α: $F > \mathrm{F}_{p-1,\mathrm{n}-\mathrm{p},\alpha}$.

2.2.9. Έλεγχος γραμμικότητας του μοντέλου

Στην απλή γραμμική παλινδρόμηση , στο διάγραμμα διασποράς θα πρέπει να έχουμε γραμμική τάση, διαφορετικά το μοντέλο μας είναι λανθασμένο. Αν τώρα η γραφική παράσταση μας θυμίζει κάποια συνάρτηση τότε μπορούμε να χρησιμοποιήσουμε

κάποιον μετασχηματισμό για να οδηγηθούμε σε γραμμικό μοντέλο. Μερικοί μετασχηματισμοί παρουσιάζονται στον εξής πίνακα:

Μοντέλο	*Συνάρτηση y=f(x)*	*Μετασχηματισμός*	*Γραμμική Μορφή*
M1	$\Upsilon = \gamma_0 \cdot \gamma_1^X$	$Y' = \log Y$	$Y' = \log\gamma_0 + \log\gamma_1 \cdot X$
M2	$\Upsilon = e^{\gamma_0+\gamma_1 \cdot X}$	$Y' = \log Y$	$Y' = \gamma_0 + \gamma_1 \cdot X$
M3	$\Upsilon = \gamma_0 + \gamma_1 \cdot \log X$	$X' = \log X$	$\Upsilon = \gamma_0 + \gamma_1 \cdot X'$
M4	$\Upsilon = \gamma_0 + \gamma_1 \cdot \sqrt{X}$	$X' = \sqrt{X}$	$\Upsilon = \gamma_0 + \gamma_1 \cdot X'$

Στο μοντέλο πολλαπλής γραμμικής παλινδρόμησης όμως, δεν χρησιμοποιούμε τις γραφικές παραστάσεις του Υ με καθένα από τα X_i, αλλά των μερικών υπολοίπων e'_{ij} ως προς τα X_i, όπου $e'_{ij} = e_{ij} + \widehat{\beta_j} \cdot X_{ij}$. Για να πούμε πως έχουμε «καλό» μοντέλο θα πρέπει η γραφική παράσταση των (e'_{ij}, X_{ij}) να προσεγγίζει ευθεία που διέρχεται από την αρχή των αξόνων με κλίση β_j. Η ορθότητα τότε του μοντέλου ελέγχεται με την γραφική παράσταση των υπολοίπων e με τον εκτιμητή του Υ, $\hat{Y}$.

Αν τώρα παρατηρήσουμε πως η γραφική παράσταση των (e'_{ij}, X_{ij}) έχει περίπου την μορφή γνωστής πολυωνυμικής συνάρτησης, εισαγάγουμε έναν επιπλέον όρο στο μοντέλο, βαθμού όσου η συνάρτηση αυτή.

Ορισμός: Ακραίες παρατηρήσεις ονομάζονται αυτές οι οποίες φαίνεται να μην προσαρμόζονται στο μοντέλο μας. Πιθανός λόγος εμφάνισής τους είναι η παραβίαση των υποθέσεων του μοντέλου.

Σχόλια: Οι ακραίες παρατηρήσεις μπορεί:

1. Να μας οδηγούν σε λανθασμένα συμπεράσματα.
2. Να έχουν ισχυρή επίδραση στο προσαρμοζόμενο μοντέλο, ιδιαίτερα όταν παρουσιάζουν υψηλή μόχλευση με αποτέλεσμα αν τα διαγράφουμε να οδηγούμαστε σε διαφορετικά αποτελέσματα.

Αυτό το πρόβλημα το διαχειριζόμαστε ως εξής:

Μελετάμε το πρόβλημά μας με αλλά και χωρίς τις ακραίες παρατηρήσεις και καταγράφουμε τις διαφορές. Αν οι διαφορές είναι μικρές τότε τις διαγράφουμε διαφορετικά τις μετασχηματίζουμε, π.χ. παίρνουμε τον λογάριθμο της παρατήρησης.

2.2.10. Test των Durbin- Watson

Ελέγχει για αυτοσυσχέτιση πρώτου βαθμού, ελέγχει δηλαδή αν τα σφάλματα πληρούν την σχέση: $\varepsilon_t = \mathrm{p} \cdot \varepsilon_{t-1} + u_t$, όπου u_t τ.μ. με $u_t \sim N(0, \sigma^2)$ ανεξάρτητες μεταξύ τους και |p|<1.

1. Ο έλεγχος H_0: $\mathrm{p} = 0$ έναντι H_α: $\mathrm{p} > 0$ γίνεται με την συνάρτηση

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

- Αν $\mathrm{d} < d_L$ απορρίπτουμε την μηδενική υπόθεση σε επίπεδο σημαντικότητας α.
- Αν $\mathrm{d} > d_U$ δεν απορρίπτουμε την μηδενική υπόθεση σε επίπεδο σημαντικότητας α.
- Αν $d_L < \mathrm{d} < d_U$ το τεστ δεν δίνει απόφαση.

d_U *και* d_L σταθερές που δίνονται από πίνακες για διαφορετικές τιμές των κ, α όπου κ είναι ο αριθμός των ανεξάρτητων μεταβλητών του μοντέλου.

2. Αν έχουμε H_0: $\mathrm{p} = 0$ έναντι H_α: $\mathrm{p} < 0$ ακολουθούμε τα ίδια βήματα απόφασης αλλά ως συνάρτηση έχουμε την $\mathrm{d}^* = 4 - d$.
3. Για τον έλεγχο H_0: $\mathrm{p} = 0$ έναντι H_α: $\mathrm{p} \neq 0$ αν $\mathrm{d} < d_L$ ή $4 - \mathrm{d} < d_L$ απορρίπτουμε την μηδενική σε επίπεδο σημαντικότητας 2α και αν $\mathrm{d} > d_U$ ή $4 - \mathrm{d} > d_U$ δεν απορρίπτουμε την μηδενική υπόθεση. Διαφορετικά το τεστ δεν δίνει απόφαση.

Ένας εκτιμητής του p είναι ο εξής: $\hat{\mathrm{p}} = \frac{\sum_{t=2}^{n} e_t \cdot e_{t-1}}{\sum_{t=1}^{n} e_t^2}$ και $\mathrm{d} \approx 2(1-\hat{\mathrm{p}})$. Άρα τιμές του d κοντά στο 2 δίνουν ένδειξη ότι $\mathrm{p} \approx 0$ ενώ τιμές του κοντά στο 0 δείχνουν ότι $\mathrm{p} \approx 1$.Όσο πιο κοντά είναι στο 2 τόσο ισχυρότερη η ένδειξη της μη αυτοσυσχέτισης.

2.2.11. Έλεγχος κανονικότητας των σφαλμάτων

Ο έλεγχος αυτός γίνεται με τα στατιστικά τεστ:
Kolmogorov – Smirnov, Shapiro – Wilk, Anderson – Darling. Το καλύτερο με βάση την ισχύ είναι το Shapiro – Wilk.

2.2.12. Επιλογή του «καλύτερου» μοντέλου

Μέθοδοι επιλογής:

1. (Αν p μικρό) Τρέχουμε όλες τις γραμμικές παλινδρομήσεις με 1, 2,…,p και με τις p ανεξάρτητες μεταβλητές και επιλέγουμε το καλύτερο με βάση τα κριτήρια παρακάτω.
2. Μέθοδος προσθήκης ανεξάρτητων μεταβλητών:

   Βήμα 1: Ξεκινάμε με το μοντέλο $Y = \beta_0 + \varepsilon$.

   Βήμα 2: Κάνουμε παλινδρομήσεις για το μοντέλο $Y = \beta_0 + \beta_1 \cdot X_j + \varepsilon$ για κάθε μια από τις ανεξάρτητες μεταβλητές X_j, j=1,…,p. Επιλέγουμε εκείνη την ανεξάρτητη μεταβλητή στην οποία στο F-test για τον έλεγχο H_0: $\beta_1 = 0$ έδωσε το μικρότερο p-value. Έπειτα προσθέτουμε έναν ακόμα όρο και ακολουθούμε την ίδια διαδικασία έως ότου να εξαντληθούν οι ανεξάρτητες μεταβλητές.
3. Μέθοδος διαγραφής ανεξάρτητων μεταβλητών:

   Βήμα 1: Ξεκινάμε με το μοντέλο $Y = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p + \varepsilon$.

   Βήμα 2: Διαγράφουμε την μεταβλητή X_j για την οποία ο έλεγχος H_0: $\beta_j = 0$ έδωσε το μεγαλύτερο p-value.

   Βήμα 3: Προσαρμόζοντας το μοντέλο με τις p-1 μεταβλητές που απομένουν επαναλαμβάνουμε το προηγούμενο βήμα.

Κριτήρια επιλογής μοντέλου:

1. Θέλουμε το μοντέλο μας να είναι όσο το δυνατόν πιο απλό και κατά συνέπεια πιο οικονομικό. Ουσιαστικά δηλαδή θέλουμε να περιέχει τις πιο σημαντικές μεταβλητές.
2. Συντελεστής προσδιορισμού R^2

   Χρησιμοποιείται για σύγκριση μοντέλων με τον ίδιο αριθμό μεταβλητών και καλύτερο είναι αυτό με την μεγαλύτερη τιμή $του\ R^2$.
3. $R^2_{adj} = 1 - \frac{SSres/_{\beta.\varepsilon.\upsilon\pi o\lambda o\acute{\iota}\pi\omega\nu}}{SStot/_{n-1}}$.

   Καλύτερο αυτό με την μεγαλύτερη τιμή του R^2_{adj}.
4. $s^2 = MSres$.

   Χρησιμοποιείται για την σύγκριση μοντέλων με τον ίδιο αριθμό παρατηρήσεων και καλύτερο είναι αυτό με την μικρότερη τιμή του s^2.

2.2.13. Πολυσυγγραμμικότητα

Το πρόβλημα αυτό προκύπτει όταν κάποιες από τις ανεξάρτητες μεταβλητές μας $X_i, i = 1, \ldots, p-1$ είναι γραμμικά εξαρτημένες και συνεπώς ο πίνακας πληροφορίας

είναι μη αντιστρέψιμος, κάτι που σημαίνει πως δεν μπορούμε να βρούμε τους εκτιμητές των παραμέτρων του μοντέλου. Ένας παράγοντας διάγνωσης της πολυσυγγραμμικότητας είναι ο παράγοντας διόγκωσης διασποράς VIF (Variance Inflation Factor), ο οποίος για την ανεξάρτητη μεταβλητή X_i είναι:

$$VIFi = \frac{1}{1 - R_i^2}, R_i^2 \text{ ο συντελεστής προσδιορισμού του μοντέλου}$$

όταν η X_i είναι η εξαρτημένη μεταβλητή και οι υπόλοιπες p-2 είναι οι ανεξάρτητες. Για τιμές του VIFi μεγαλύτερες του 10 έχουμε πολυσυγγραμμικότητα και συνεπώς η συγκεκριμένη μεταβλητή θα πρέπει να αφαιρεθεί και να κάνουμε την ανάλυσή μας χωρίς αυτήν, ενώ για μικρές, σχεδόν μοναδιαίες τιμές του δείκτη VIFi δεν έχουμε πρόβλημα με την μεταβλητή.

Ένα ομοιόμορφο κριτήριο για τον έλεγχο ύπαρξης ή μη πολυσυγγραμμικότητας είναι ο μέσος όρος του VIF, δηλαδή:

$$\overline{VIF} = \frac{1}{p-1} \cdot \sum_{i=1}^{p-1} VIFi$$

Με αυτό, αν έχουμε τιμές πολύ μεγαλύτερες της μονάδας, τότε κάποιος από τους δείκτες είναι μεγάλος, συνεπώς έχουμε πολυσυγγραμμικότητα.

Τέλος ένας δείκτης για έλεγχο πολυσυγγραμμικότητας είναι ο συντελεστής ανεκτικότητας (Tolerance Intex):

$$TOL = 1 - R_i^2 = \frac{1}{VIFi}$$

Τιμές κοντά στο 0 δείχνουν πιθανή συσχέτιση, ενώ τιμή κοντά στο 1 το ακριβώς αντίθετο.

2.3. Παράδειγμα πολλαπλής γραμμικής παλινδρόμησης στην R

Στην R υπάρχουν δεδομένα τα οποία δίνονται ελεύθερα για χρήση. Τα δεδομένα μας λοιπόν είναι από ένα περιοδικό για αυτοκίνητα του 1974 και σχετίζονται με την κατανάλωση καυσίμων (mpg) και 10 διαφορετικά χαρακτηριστικά αυτοκινήτων και των αποδόσεών τους για 32 διαφορετικά μοντέλα.

Για να «καλέσουμε» τα δεδομένα μας πληκτρολογούμε data=mtcars και για να τα τυπώσουμε στην οθόνη print(data) και έχουμε:

	Mpg	cly	disp	Hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4

Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4

Continental											
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2

AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4

Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

Οι μεταβλητές μας περιγράφουν με την σειρά:

1. Mpg: μίλι/γαλόνι
2. Cly: αριθμό κυλίνδρων
3. Disp: κυβισμό
4. Hp: ιπποδύναμη
5. Drat: αναλογία πίσω άξονα
6. Wt: βάρος
7. Qsec: χρόνος κάτω από το μήκος του ¼ του μιλίου
8. Vs: τύπος κινητήρα (0 για τον κινητήρα με διάταξη V και 1 για τον κινητήρα με ευθεία διάταξη)
9. Am: κιβώτιο ταχυτήτων (0 αυτόματο, 1 χειροκίνητο)
10. Gear: αριθμός ταχυτήτων πλην της όπισθεν
11. Carb: αριθμός καρμπυρατέρ

Ετοιμάζουμε τα δεδομένα μας για χρήση με την εντολή attact(data) και μετονομάζουμε τις μεταβλητές μας ως εξής:

Y=mpg, X1=cly, X2=disp,…, X10=carb.

Για να εφαρμόσουμε το γραμμικό μοντέλο φτιάχνουμε τον πίνακα σχεδιασμού X

X=cbind(rep(1,32), X1, X2, X3, X4, X5, X6, X7, X8, X9, X10) και εφαρμόζουμε γραμμική παλινδρόμηση:

model=lm(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10) και εξάγουμε τον πίνακα:

```
Coefficients:
(Intercept)           X1           X2
   12.30337     -0.11144      0.01334
         X3           X4           X5
   -0.02148      0.78711     -3.71530
         X6           X7           X8
    0.82104      0.31776      2.52023
         X9          X10
    0.65541     -0.19942
```

Ο πίνακας αυτός μας δείχνει ότι το προσαρμοζόμενο μοντέλο είναι το:

$$\hat{Y} = 12.30337 - 0.11144 \cdot X_1 + 0.01334 \cdot X_2 - 0.02148 \cdot X_3 + 0.78711 \cdot X_4 - 3.71530 \cdot X_5 + 0.82104 \cdot X_6 + 0.31776 \cdot X_7 + 2.52023 \cdot X_8 + 0.65541 \cdot X_9 - 0.19942 \cdot X_{10}$$

Έπειτα δημιουργούμε τον πίνακα ανάλυσης διακύμανσης με την εντολή anova(model):

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq  F value
X1         1 817.71  817.71 116.4245
X2         1  37.59   37.59   5.3526
X3         1   9.37    9.37   1.3342
X4         1  16.47   16.47   2.3446
X5         1  77.48   77.48  11.0309
X6         1   3.95    3.95   0.5623
X7         1   0.13    0.13   0.0185
X8         1  14.47   14.47   2.0608
X9         1   0.97    0.97   0.1384
X10        1   0.41    0.41   0.0579
Residuals 21 147.49    7.02
             Pr(>F)
X1        5.034e-10 ***
X2         0.030911 *
X3         0.261031
X4         0.140644
X5         0.003244 **
```

```
X6          0.461656
X7          0.893173
X8          0.165858
X9          0.713653
X10         0.812179
Residuals
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1
```

Ή ισοδύναμα είναι ο πίνακας:

Πηγή μεταβλητότητας	Άθροισμα τετραγώνων	Βαθμοί ελευθερίας	Μέσα τετράγωνα	F-test
Παλινδρόμηση	SSreg=978.55	10	MSreg=97.855	F=13.93
Υπόλοιπα	SSres=147.49	21	MSres7.02	
Ολική μεταβλητότητα	SStot=1126.04	31		

Για να ελέγξουμε τι γίνεται με τις παραμέτρους του μοντέλου εκτελούμε την εντολή summary(model) και έχουμε:

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 +
    X10)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value
(Intercept) 12.30337   18.71788   0.657
X1          -0.11144    1.04502  -0.107
X2           0.01334    0.01786   0.747
X3          -0.02148    0.02177  -0.987
X4           0.78711    1.63537   0.481
```

```
X5          -3.71530    1.89441  -1.961
X6           0.82104    0.73084   1.123
X7           0.31776    2.10451   0.151
X8           2.52023    2.05665   1.225
X9           0.65541    1.49326   0.439
X10         -0.19942    0.82875  -0.241
            Pr(>|t|)
(Intercept)   0.5181
X1            0.9161
X2            0.4635
X3            0.3350
X4            0.6353
X5            0.0633 .
X6            0.2739
X7            0.8814
X8            0.2340
X9            0.6652
X10           0.8122
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869, Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Στο πρώτο μέρος έχουμε μερικά στοιχεία για τα κατάλοιπα. Στην πρώτη στήλη του πίνακα (Estimate) έχουμε ακριβώς τους ίδιους εκτιμητές που είχαμε βρει πριν για τις παραμέτρους, στην δεύτερη στήλη έχουμε τους εκτιμητές ελαχίστων τετραγώνων των παραμέτρων και στην στήλη t-value δίνονται οι τιμές των σ.σ.ε. Παρατηρούμε πως οι τιμές των πιθανοτήτων είναι σε όλες τις παραμέτρους μεγαλύτερες του επιπέδου σημαντικότητας 0.05 και συνεπώς δεν είναι στατιστικά σημαντικοί. Η τιμή του F-test είναι η ίδια που βρήκαμε και στον anova, η τιμή του p-value = 3.793e-07 < 0.001, δηλαδή έχει πολύ μικρή τιμή και συνεπώς απορρίπτουμε την υπόθεση ότι όλες οι παράμετροι του μοντέλου είναι ίσες με το 0 ($H_0: \beta_1 = \beta_2 = \ldots = \beta_{10} = 0$) άρα το μοντέλο είναι στατιστικά σημαντικό. Επίσης από τον εξαγόμενο πίνακα βλέπουμε ότι

$$R^2 = 0.869 \; και \; R^2_{adj} = 0.8066.$$

Οι τιμές του t-test μας δημιούργησαν έναν προβληματισμό για την πιθανή ύπαρξη πολυσυγγραμμικότητας. Συνεπώς πρέπει να εξετάσουμε το ενδεχόμενο αυτό. Όπως είδαμε στην θεωρία αυτό μπορούμε να το κάνουμε με τον παράγοντα διόγκωσης VIF. Προσθέτουμε το πακέτο car όπου υπάρχει η συνάρτηση VIF:

```
> library("car", lib.loc="~/R/win-library/3.5")
```

```
> vif(model)
       X1        X2        X3        X4        X5        X6        X7        X8
15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873  4.648487
       X9       X10
 5.357452  7.908747
>
```

Παρατηρούμε ότι πολλές τιμές είναι μεγαλύτερες του 10, άρα έχουμε πρόβλημα πολυσυγγραμμικότητας.

Για να λύσουμε το πρόβλημα αυτό θα πρέπει να αφαιρέσουμε όσες μεταβλητές δεν μας χρειάζονται. Αυτό όπως έχουμε δει και στην θεωρία γίνεται με δυο μεθόδους, την Backward και την Forward. Εμείς θα ακολουθήσουμε την 1η εκ των μεθόδων, και θα την επαναλάβουμε έως ότου όλες οι μεταβλητές που θα μας έχουν απομείνει να είναι στατιστικά σημαντικές, δηλαδή να έχουν πιθανότητα μικρότερη του επιπέδου σημαντικότητας, οπότε έχουμε:

```
> drop1(lm(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10),test="F")


Single term deletions

Model:
Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10
       Df Sum of Sq    RSS    AIC
<none>              147.49 70.898
X1      1    0.0799 147.57 68.915
X2      1    3.9167 151.41 69.736
X3      1    6.8399 154.33 70.348
X4      1    1.6270 149.12 69.249
X5      1   27.0144 174.51 74.280
X6      1    8.8641 156.36 70.765
X7      1    0.1601 147.66 68.932
X8      1   10.5467 158.04 71.108
X9      1    1.3531 148.85 69.190
X10     1    0.4067 147.90 68.986
       F value  Pr(>F)
```

```
<none>
X1      0.0114 0.91609
X2      0.5576 0.46349
X3      0.9739 0.33496
X4      0.2317 0.63528
X5      3.8463 0.06325 .
X6      1.2621 0.27394
X7      0.0228 0.88142
X8      1.5016 0.23399
X9      0.1926 0.66521
X10     0.0579 0.81218
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1
```

Από τους εξαγόμενους πίνακες εμείς αναζητούμε την μεταβλητή η οποία έχει την μεγαλύτερη πιθανότητα (Pr > F) και την αφαιρούμε από το μοντέλο. Αντί την εντολής drop1 θα μπορούσαμε να κάνουμε απλώς summary, θα είχαμε ακριβώς το ίδιο αποτέλεσμα. Παρατηρούμε λοιπόν πως η μεγαλύτερη τιμή εντοπίζεται στην μεταβλητή X_1, οπότε και την αφαιρούμε. Κάνουμε την ίδια διαδικασία χωρίς την X_1:

```
> drop1(lm(Y~X2+X3+X4+X5+X6+X7+X8+X9+X10),test="F")


Single term deletions

Model:
Y ~ X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10
       Df Sum of Sq    RSS    AIC
<none>              147.57 68.915
X2      1    3.9009 151.47 67.750
X3      1    7.3632 154.94 68.473
X4      1    1.9826 149.56 67.342
X5      1   27.0280 174.60 72.297
X6      1   10.0933 157.67 69.032
X7      1    0.2685 147.84 66.973
X8      1   11.8359 159.41 69.384
X9      1    1.8211 149.40 67.308
X10     1    0.5201 148.09 67.028
       F value  Pr(>F)
<none>
X2      0.5815 0.45381
```

```
X3      1.0977 0.30615
X4      0.2956 0.59214
X5      4.0293 0.05716 .
X6      1.5047 0.23292
X7      0.0400 0.84326
X8      1.7645 0.19768
X9      0.2715 0.60754
X10     0.0775 0.78326
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1
```

Τώρα η μεγαλύτερη τιμή παρατηρείται στην X_7, οπότε την αφαιρούμε και ακολουθούμε την ίδια διαδικασία:

```
> drop1(lm(Y~X2+X3+X4+X5+X6+X8+X9+X10),test="F")

Single term deletions

Model:
Y ~ X2 + X3 + X4 + X5 + X6 + X8 + X9 + X10
       Df Sum of Sq    RSS    AIC
<none>              147.84 66.973
X2      1    3.6467 151.49 65.753
X3      1    7.1060 154.95 66.475
X4      1    2.2139 150.06 65.449
X5      1   27.3799 175.22 70.410
X6      1   15.6830 163.53 68.200
X8      1   11.5694 159.41 67.384
X9      1    2.1437 149.99 65.434
X10     1    0.6855 148.53 65.121
       F value  Pr(>F)
<none>
X2      0.5673 0.45897
X3      1.1055 0.30399
X4      0.3444 0.56301
X5      4.2595 0.05049 .
X6      2.4398 0.13195
X8      1.7999 0.19283
X9      0.3335 0.56922
X10     0.1066 0.74696
---
Signif. codes:
```

```
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1
```

Παρατηρούμε πως η X_{10} έχει την μεγαλύτερη τιμή, οπότε συνεχίζουμε την ίδια διαδικασία χωρίς αυτήν:

```
> drop1(lm(Y~X2+X3+X4+X5+X6+X8+X9),test="F")


Single term deletions

Model:
Y ~ X2 + X3 + X4 + X5 + X6 + X8 + X9
       Df Sum of Sq    RSS    AIC
<none>              148.53 65.121
X2      1    10.110 158.64 65.229
X3      1    14.826 163.35 66.166
X4      1     1.932 150.46 63.535
X5      1    69.127 217.66 75.350
X6      1    26.408 174.94 68.358
X8      1    12.323 160.85 65.672
X9      1     1.565 150.09 63.457
       F value   Pr(>F)
<none>
X2      1.6337 0.213420
X3      2.3956 0.134763
X4      0.3122 0.581508
X5     11.1699 0.002717 **
X6      4.2672 0.049815 *
X8      1.9913 0.171042
X9      0.2529 0.619641
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1
```

Αφαιρούμε την X_9 και έχουμε:

```
> drop1(lm(Y~X2+X3+X4+X5+X6+X8),test="F")

Single term deletions

Model:
Y ~ X2 + X3 + X4 + X5 + X6 + X8
       Df Sum of Sq    RSS    AIC
```

```
<none>              150.09 63.457
X2      1     8.545 158.64 63.229
X3      1    13.285 163.38 64.171
X4      1     3.345 153.44 62.162
X5      1    67.572 217.66 73.351
X6      1    25.574 175.67 66.491
X8      1    20.036 170.13 65.466
       F value   Pr(>F)
<none>
X2      1.4233 0.244054
X3      2.2127 0.149381
X4      0.5571 0.462401
X5     11.2550 0.002536 **
X6      4.2598 0.049551 *
X8      3.3372 0.079692 .
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1
```

Αφαιρούμε την X_4 και συνεχίζουμε:

```
> drop1(lm(Y~X2+X3+X5+X6+X8),test="F")

Single term deletions

Model:
Y ~ X2 + X3 + X5 + X6 + X8
       Df Sum of Sq    RSS    AIC
<none>              153.44 62.162
X2      1     6.629 160.07 61.515
X3      1    12.572 166.01 62.682
X5      1    69.043 222.48 72.051
X6      1    26.470 179.91 65.255
X8      1    32.198 185.63 66.258
       F value   Pr(>F)
<none>
X2      1.1232 0.298972
X3      2.1303 0.156387
X5     11.6993 0.002075 **
X6      4.4853 0.043908 *
X8      5.4559 0.027488 *
---
Signif. codes:
```

```
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1
```

Αφαιρούμε και την X_2 και έχουμε:

```
> drop1(lm(Y~X3+X5+X6+X8),test="F")

Single term deletions

Model:
Y ~ X3 + X5 + X6 + X8
       Df Sum of Sq    RSS    AIC
<none>              160.07 61.515
X3      1     9.219 169.29 61.307
X5      1    78.494 238.56 72.284
X6      1    20.225 180.29 63.323
X8      1    25.993 186.06 64.331
       F value   Pr(>F)
<none>
X3      1.5551 0.223088
X5     13.2403 0.001141 **
X6      3.4115 0.075731 .
X8      4.3845 0.045791 *
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1
```

Τέλος αφαιρούμε την X_3 και κάνουμε τον έλεγχο για να δούμε αν όσες έχουν απομείνει είναι στατιστικά σημαντικές:

```
> drop1(lm(Y~X5+X6+X8),test="F")
Single term deletions

Model:
Y ~ X5 + X6 + X8
       Df Sum of Sq    RSS    AIC
<none>              169.29 61.307
X5      1   183.347 352.63 82.790
X6      1   109.034 278.32 75.217
X8      1    26.178 195.46 63.908
       F value    Pr(>F)
<none>
X5     30.3258 6.953e-06 ***
X6     18.0343 0.0002162 ***
```

```
X8          4.3298 0.0467155 *
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1
```

Παρατηρούμε ότι οι τρεις μεταβλητές που έχουν απομείνει είναι στατιστικά σημαντικές σε επίπεδο σημαντικότητας 5%. Για να δούμε τώρα ποιο θα είναι το προσαρμοζόμενο μοντέλο μετά την αφαίρεση των μεταβλητών έχουμε:

```
> lm(Y~X5+X6+X8)

Call:
lm(formula = Y ~ X5 + X6 + X8)

Coefficients:
(Intercept)           X5           X6
      9.618       -3.917        1.226
         X8
      2.936
```

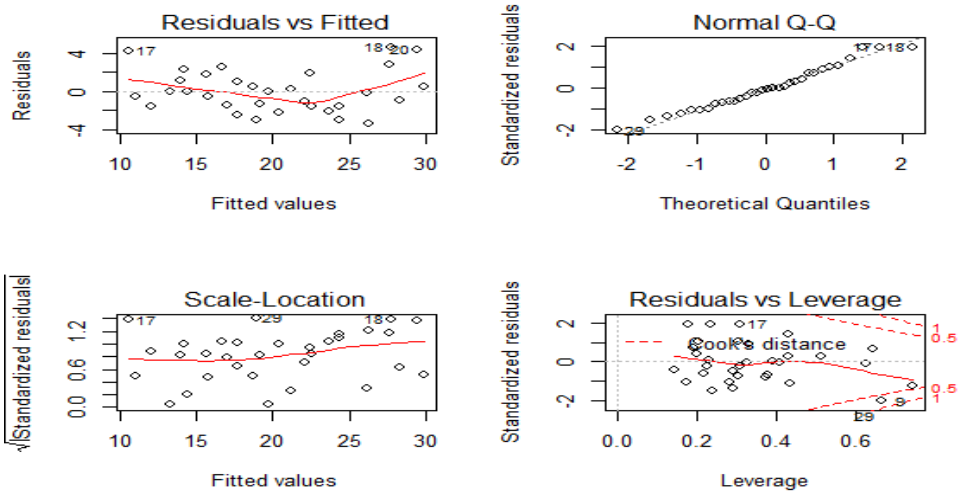
Άρα το προσαρμοζόμενο μοντέλο είναι το:

$$\hat{Y} = 9.618 - 3.917 \cdot X_5 + 1.226 \cdot X_6 + 2.936 \cdot X_8.$$

Για να ελέγξουμε τις υποθέσεις του γραμμικού μοντέλου αρκεί να κάνουμε διαγνωστικά διαγράμματα. Συνεπώς έχουμε εκτελώντας στην R τις εντολές:

```
> par(mfrow=c(2,2))

> plot(model)
```



Στο 1ο διάγραμμα βλέπουμε την γραμμικότητα του μοντέλου, κάτι που εν μέρη το έχουμε αφού η κόκκινη γραμμή προσεγγίζει την διακεκομμένη γραμμή.

Στο 2ο διάγραμμα παρατηρούμε ότι τα σημεία κινούνται πάνω στην διακεκομμένη ευθεία, συνεπώς έχουμε κανονικότητα των υπολοίπων.

Στο 3ο διάγραμμα η κόκκινη γραμμή είναι σχεδόν ευθεία οπότε έχουμε ομοσκεδαστικότητα των υπολοίπων.

Στο 4ο διάγραμμα η κόκκινη συνεχόμενη γραμμή είναι κοντά στην γκρι οριζόντια διακεκομμένη και όλα τα σημεία είναι κοντά σε αυτήν, οπότε δεν έχουμε ακραίες παρατηρήσεις και κατά συνέπεια δεν έχουμε μόχλευση.

Σχόλια:

Αντί του 2ου διαγράμματος για έλεγχο κανονικότητας των υπολοίπων θα μπορούσαμε να κάνουμε έλεγχο Shapiro - Wilk με την εντολή:

```
> shapiro.test(resid(model))

        Shapiro-Wilk normality test

data:  resid(model)
W = 0.95694, p-value = 0.2261
```

Επειδή p-value = 0.2261>0.001 αποδεχόμαστε την μηδενική υπόθεση η οποία είναι η ύπαρξη κανονικότητας των υπολοίπων.

Συμπερασματικά λοιπόν έχουμε καταλήξει στο προσαρμοζόμενο μοντέλο:

$$\hat{Y} = 9.618 - 3.917 \cdot X_5 + 1.226 \cdot X_6 + 2.936 \cdot X_8$$

κάτι που σημαίνει ότι:

- Για κάθε μοναδιαία αύξηση στο βάρος του αυτοκινήτου wt και διατηρώντας σταθερό τον χρόνο κάτω από το μήκος του ¼ του μιλίου αλλά και το ίδιο κιβώτιο ταχυτήτων η μέση τιμή της κατανάλωσης του καυσίμου μειώνεται κατά 3.917.
- Για κάθε μοναδιαία αύξηση του χρόνου qsec και διατηρώντας το βάρος και το κιβώτιο ταχυτήτων η μέση κατανάλωση αυξάνει κατά 1.226.
- Τέλος με αλλαγή στο κιβώτιο ταχυτήτων και διατηρώντας σταθερό το βάρος και τον χρόνο qsec θα έχουμε αύξηση στην μέση κατανάλωση κατά 2.936.

Κεφάλαιο 3

Δίτιμες μεταβλητές και Λογιστική παλινδρόμηση

3.1. Δίτιμες μεταβλητές

Έστω πείραμα όπου η μεταβλητή απόκρισης Y παίρνει μόνο μια από δυο δυνατές τιμές, για παράδειγμα επιτυχία ή αποτυχία, ή αλλιώς 0 ή 1 αντίστοιχα. Τότε όπως γνωρίζουμε η Y ακολουθεί Bernoulli κατανομή με παράμετρο p. Τότε έχουμε:

$$Y = \begin{cases} 1, \alpha\nu\ \acute{\epsilon}\chi ουμε\ \epsilon\pi\iota\tau υχία \\ 0, \alpha\nu\ \acute{\epsilon}\chi ουμε\ \alpha\pi οτυχία \end{cases}$$

Αν p είναι η πιθανότητα επιτυχίας τότε:

$$P(Y = 1) = p\ \kappa\alpha\iota\ P(Y = 0) = 1 - p$$

και η συνάρτηση πιθανότητας της Y είναι:

$$f(y; \pi) = P(Y = y) = \pi^y \cdot (1 - \pi)^{1-y}, y \in 0{,}1.$$

Αν τώρα έχουμε ένα σύνολο από i=1,…, n τέτοιες μεταβλητές με $Y_i \sim Bernoulli(p_i)$ τότε η από κοινού συνάρτηση πιθανότητας είναι:

$$f(y; \pi) = \prod_{i=1}^{n} \pi_i^{y_i} \cdot (1 - \pi_i)^{1-y_i} = exp\left[\sum_{i=1}^{n} y_i \cdot \log\frac{\pi_i}{1 - \pi_i} + \sum_{i=1}^{n} \log(1 - \pi_i)\right],$$

όπου $\pi = [\pi_1 \quad \dots \quad \pi_n]^T\ \kappa\alpha\iota\ y = [y_1 \quad \dots \quad y_n]^T$.

Μπορούμε επίσης να ορίσουμε την συνάρτηση Y η οποία περιγράφει τον αριθμό των επιτυχιών σε n ανεξάρτητες προσπάθειες, αν φυσικά p_i ίσα μεταξύ τους. Τότε η

$$Y = \sum_{i=1}^{n} Y_i$$

ακολουθεί Διωνυμική κατανομή με παραμέτρους n, p και συνάρτηση πιθανότητας:

$$P(Y = y) = \binom{n}{y} \cdot \pi^y \cdot (1 - \pi)^{n-y}, y = 0{,}1, \dots, n$$

Σε αυτήν την περίπτωση η συνάρτηση ln- πιθανοφάνειας για $Y_i \sim Bin(n_i, \pi_i)$ γίνεται:

$$l(\pi_1, \dots, \pi_N; y_1, \dots, y_N) = \sum_{i=1}^{N} \left[y_i \cdot \log\left(\frac{\pi_i}{1 - \pi_i}\right) + n_i \cdot \log(1 - \pi_i) + \log\binom{n_i}{y_i}\right].$$

3.2. Γενικευμένα Γραμμικά Μοντέλα και Δίτιμες Μεταβλητές

Θεωρούμε Ν στο πλήθος ανεξάρτητες μεταβλητές με $Y_i \sim Bin(n_i, \pi_i)$. Θα πρέπει να βρούμε μια σχέση της πιθανότητας επιτυχιών με την επεξηγηματική μεταβλητή. Όπως έχουμε δει μια τέτοια σχέση μπορεί να θεωρηθεί η συνάρτηση στήριξης

$$g(\pi_i) = x_i^T \cdot \beta,$$

όπου x_i είναι το διάνυσμα των επεξηγηματικών μεταβλητών και β το διάνυσμα των παραμέτρων του μοντέλου. Όμως επειδή p_i είναι πιθανότητα θα ανήκει στο διάστημα [0,1] κάτι το οποίο περιορίζει τις επιλογές των x_i και β.

Κάτι τέτοιο αποφεύγεται επιλέγοντας κατάλληλη εκτίμηση της p μέσω μιας κατανομής. Συνεπώς για την Διωνυμική κατανομή επιλέγουμε ως συνάρτηση σύνδεσης μια εκ των:

- Logit: $g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, κάτι που σημαίνει ότι $\pi_i = \frac{e^{n_i}}{1+e^{n_i}}$. Το μοντέλο με συνάρτηση σύνδεσης την logit είναι το πιο συχνά χρησιμοποιούμενο.
- Probit: $g(\pi_i) = \Phi^{-1}(\pi_i)$, όπου Φ είναι η συνάρτηση κατανομής της κανονικής κατανομής και Φ^{-1} η αντίστροφη αυτής. Συνεπώς $p_i = \Phi(n_i)$.
- Complementary log – log: $g(\pi_i) = \log(-\log(1-\pi_i))$ και συνεπώς
$$\pi_i = 1 - exp(-e^{n_i}).$$
Την περίπτωση αυτήν την επιλέγουμε συνήθως όταν έχουμε είτε πολύ μεγάλη είτε πολύ μικρή πιθανότητα να συμβεί ένα γεγονός.

Σαφώς και υπάρχουν και άλλες επιλογές για την συνάρτηση σύνδεσης, αλλά αυτές οι τρεις είναι οι πιο διαδεδομένες. Μια σημαντική παρατήρηση είναι πως και οι τρεις αυτές συναρτήσεις είναι συνεχείς και αύξουσες στο (0,1).

3.3. Λογιστική παλινδρόμηση

Πολλές φορές η μεταβλητή απόκρισης είναι διακριτή και μπορεί να παίρνει παραπάνω από δυο τιμές. Τότε θέλουμε να βρούμε το καλύτερο μοντέλο που να περιγράφει την σχέση της μεταβλητής απόκρισης με τις επεξηγηματικές μεταβλητές. Θέλουμε λοιπόν να κάνουμε μια στατιστική ανάλυση για την εύρεση τέτοιου μοντέλου. Μια μέθοδος για την παραπάνω στατιστική ανάλυση είναι η Λογιστική Παλινδρόμηση. Στην Λογιστική παλινδρόμηση, η μεταβλητή απόκρισης ακολουθεί Bernoulli ή Διωνυμική κατανομή.

Όταν η τ.μ. Υ ακολουθεί Bernoulli κατανομή έχουμε ότι:

$$E(y_i) = \pi_i = P(x_i)\ και\ Var(y_i) = \pi_i \cdot (1 - \pi_i).$$

Αν τώρα θεωρήσουμε το μοντέλο $Y_i = \beta_0 + \beta_i \cdot x_i + \varepsilon_i$ έχουμε:

$$E(Y_i) = \beta_0 + \beta_i \cdot X_i \Rightarrow \beta_0 + \beta_i \cdot X_i = \pi_i.$$

Επίσης, έχουμε ότι $P(Y_i = 1) = \pi_i,\ P(Y_i = 0) = 1 - \pi_i.$

Άρα η μέση τιμή $E(Y_i) = \beta_0 + \beta_i \cdot X_i$ μας δίνει την πιθανότητα $Y_i = 1.$

Έστω ότι y = 1 είναι η επιτυχία με πιθανότητα π και y = 0 η αποτυχία με πιθανότητα $1 - \pi$, τότε έχουμε:

$$E(Y_i) = \frac{exp(\beta_0 + \beta_i \cdot X_i)}{1 + exp(\beta_0 + \beta_i \cdot X_i)}.$$

3.3.1. Γενικό λογιστικό μοντέλο

Το μοντέλο αυτό έχει την εξής μορφή:

$$logit\pi_i = \log\frac{\pi_i}{1 - \pi_i} = x_i^T \cdot \beta,$$

όπου x_i είναι διάνυσμα συνεχών τιμών που αντιστοιχεί σε συμμεταβλητές και το β είναι το διάνυσμα των παραμέτρων β_i.

Το μοντέλο αυτό το επεξεργαζόμαστε με τεχνικές ανάλογες του πολλαπλού γραμμικού μοντέλου, γι' αυτό και είναι αρκετά διαδεδομένο.

Οι εκτιμητές μέγιστης πιθανοφάνειας προκύπτουν από την μεγιστοποίηση της λογαριθμικής συνάρτησης πιθανοφάνειας:

$$l(\pi; y) = \sum_{i=1}^{N} \left\{ y_i \cdot \log \pi_i + (n_i - y_i) \cdot \log(1 - \pi_i) + \log \binom{n_i}{y_i} \right\}.$$

3.3.2. Καλή προσαρμογή του μοντέλου

Για να ελέγξουμε την προσαρμοστικότητα του μοντέλου αντί της μεθόδου εκτίμησης μέγιστης πιθανοφάνειας μπορούμε να εκτιμήσουμε τις παραμέτρους μας ελαχιστοποιώντας το σταθμισμένο άθροισμα τετραγώνων, δηλαδή το:

$$S_w = \sum_{i=1}^{N} \frac{(y_i - n_i \cdot \pi_i)^2}{n_i \cdot \pi_i(1 - \pi_i)}, όπου E(y_i) = n_i \cdot \pi_i, Var(y_i) = n_i \cdot \pi_i(1 - \pi_i).$$

Ισοδύναμα θα μπορούσαμε να ελαχιστοποιήσουμε την στατιστική συνάρτηση χ-τετράγωνο του Pearson,

$$X^2 = \sum_{\iota=1}^{N} \frac{(o_i - e_i)^2}{e_i}, \qquad \mu\varepsilon\ o_i\ \sigma\upsilon\chi\nu\acute{o}\tau\eta\tau\varepsilon\varsigma\ \kappa\alpha\iota\ e_i\ \alpha\nu\alpha\mu\varepsilon\nu\acute{o}\mu\varepsilon\nu\varepsilon\varsigma\ \sigma\upsilon\chi\nu\acute{o}\tau\eta\tau\varepsilon\varsigma.$$

Όταν το X^2 παίρνει τιμές κοντά στις αναμενόμενες συχνότητες, τότε:

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - n_i \cdot \hat{\pi_\iota})^2}{n_i \cdot \hat{\pi_\iota}(1 - \hat{\pi_\iota})}$$

κάτι που με την βοήθεια του αναπτύγματος Taylor για τη σειρά

$s \cdot \log\left(\frac{s}{t}\right)$ για s = t μας δίνει ότι είναι ασυμπτωτικά ισοδύναμη με την deviance:

$$D = 2\sum_{i=1}^{N} \left\{ y_i \log\left(\frac{y_i}{\hat{y_\iota}}\right) + (n_i - y_i) \log\left(\frac{n_i - y_i}{n_i - \hat{y_\iota}}\right) \right\}.$$

και μάλιστα έχουμε:

$$D = 2\sum_{i=1}^{N} \left\{ (y_i - n_i \cdot \hat{\pi_\iota}) + \frac{1}{2} \cdot \frac{(y_i - n_i \cdot \hat{\pi_\iota})^2}{n_i \cdot \hat{p_\iota}} + (n_i - y_i - n_i - n_i \cdot \hat{\pi_\iota}) + \frac{1}{2} \right.$$
$$\left. \cdot \frac{(n_i - y_i - n_i - n_i \cdot \hat{\pi_\iota})^2}{(n_i - n_i \cdot \hat{\pi})} + \cdots \right\}.$$

Δηλαδή αν το μοντέλο μας είναι σωστό έχουμε ότι:

$$D \simeq \sum_{i=1}^{N} \frac{(y_i - n_i \cdot \hat{\pi})^{\ 2}}{n_i \cdot \hat{\pi_\iota}(1 - \hat{\pi_\iota})} = X^2.$$

Πρακτικά αυτό σημαίνει ότι $D \sim X^2_{(N-p)}$ και προσεγγιστικά $X^2 \sim X^2_{(N-p)}$.

Από αυτές τις δύο προσεγγίσεις εμείς επιλέγουμε αυτήν που είναι πιο κοντά στην $X^2_{(N-p)}$, αν και συνίσταται η X^2 προσέγγιση, αφού για μικρές συχνότητες η D επηρεάζεται σε μεγάλο βαθμό.

3.4. Υπόλοιπα

Για την Λογιστική παλινδρόμηση έχουμε δυο τύπους υπολοίπων σχετικά με την καλή προσαρμογή του μοντέλου. Αν έχουμε m διακεκριμένες παρατηρήσεις, τότε μπορούμε να εκτιμήσουμε m υπόλοιπα. Θα συμβολίσουμε στο εξής με:

Y_k τον αριθμό των επιτυχιών, n_k τον αριθμό των επαναλήψεων και $\hat{\pi}_k$ το ποσοστό που εκτιμήσαμε.

3.4.1. Υπόλοιπα Pearson

Τα υπόλοιπα χ-τετράγωνο ή υπόλοιπα Pearson είναι της μορφής:

$$X_k = \frac{(y_k - n_k \cdot \hat{\pi}_k)}{\sqrt{n_k \cdot \hat{\pi}_k}(1 - \hat{\pi}_k)}, k = 1, \dots, m.$$

Πριν είχαμε δει ότι:

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - n_i \cdot \hat{\pi}_i)^2}{n_i \cdot \hat{\pi}_i(1 - \hat{\pi}_i)},$$

οπότε έχουμε ότι:

$$\sum_{i=1}^{m} X_k^2 = X^2, \text{ο οποίος λέγεται Pearson } \chi - \text{τετράγωνο έλεγχος προσαρμογής.}$$

Τα υπόλοιπα Pearson είναι:

$$r_{\pi_k} = \frac{X_k}{\sqrt{1 - h_k}},$$

με h_k να είναι η μόχλευση, την τιμή της οποίας παίρνουμε από τον hat-matrix ($X(X^TX)^{-1}X^T$), αφού το στοιχείο h_{ii} του πίνακα αυτού ονομάζεται μόχλευση της i-οστής παρατήρησης. Παρατηρήσεις με μεγάλη μόχλευση επηρεάζουν την προσαρμογή του μοντέλου.

3.4.2. Υπόλοιπα deviance

Έχουμε:

$$d_k = sign(y_k - n_k \cdot \hat{\pi}_k) \sqrt{\left\{2\left[y_k \cdot \log\left(\frac{y_k}{n_k \cdot \hat{\pi}_k}\right) + (n_k - y_k) \cdot \log\left(\frac{n_k - y_k}{n_k - n_k \cdot \hat{\pi}_k}\right)\right]\right\}}$$

όπου ο όρος πρόσημου $sign(y_k - n_k \cdot \hat{\pi}_k)$ μας εξασφαλίζει ότι d_k ομόσημο του X_k.

Όμως γνωρίζουμε ότι:

$$D = 2\sum_{i=1}^{N}\left[y_i \cdot \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i) \cdot \log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right] \Longrightarrow \sum_{i=1}^{N} d_k^2 = D.$$

Τα τυποποιημένα υπόλοιπα deviance λοιπόν ορίζονται ως εξής:

$$r_{D_k} = \frac{d_k}{\sqrt{1 - h_k}}.$$

Τα υπόλοιπα αυτά χρησιμοποιούνται για τον έλεγχο καταλληλότητας του μοντέλου. Τα σχεδιάζουμε ως προς κάθε επεξηγηματική μεταβλητή και ελέγχουμε αν η υπόθεση της γραμμικότητας ισχύει. Τα τυποποιημένα υπόλοιπα ακολουθούν Κανονική κατανομή με μέση τιμή 0 και τυπική απόκλιση 1.

3.5. Κριτήρια επιλογής μοντέλου στην λογιστική παλινδρόμηση

3.5.1. Κριτήριο AIC

Στην λογιστική παλινδρόμηση έχει την μορφή:

$$AIC = -2\left[\sum_{i=1}^{n} \log\binom{n_i}{y_i} + y_i \log\hat{\pi}_i + (n_i - y_i)\log(1 - \hat{\pi}_i)\right] + 2p,$$

$$\mu\varepsilon\ \hat{\pi}_i = \frac{e^{x_i^T\beta}}{1 + e^{x_i^T\beta}}.$$

3.5.2. Κριτήριο BIC

Έχει την μορφή:

$$BIC = -2\left[\sum_{i=1}^{n} \log\binom{n_i}{y_i} + y_i \log\hat{\pi}_i + (n_i - y_i)\log(1 - \hat{\pi}_i)\right] + plogn,$$

$$\mu\varepsilon\ \hat{\pi}_i = \frac{e^{x_i^T\beta}}{1 + e^{x_i^T\beta}}.$$

3.5.3. Κριτήρια R^2

Όπως και στο γραμμικό μοντέλο, έτσι και εδώ έχουμε τον συντελεστή προσδιορισμού, αλλά έχει παρατηρηθεί ότι δεν είναι ικανοποιητικό μέτρο προσαρμογής των δεδομένων στο μοντέλο, απλώς χρησιμοποιείται για σύγκριση προσαρμογής διαφορετικών μοντέλων στα ίδια δεδομένα.

3.6. Παράδειγμα λογιστικής παλινδρόμησης στην R

Σε μια έρευνα που πραγματοποιήθηκε με εργάτες της αμερικανικής βιομηχανίας βαμβακιού ήθελαν να εξετάσουν αν κάποιος εργάτης πάσχει από κάποια συγκεκριμένη ασθένεια του πνεύμονα. Συγκεντρώθηκαν τιμές για τις ακόλουθες 5 μεταβλητές:

- Φυλή (race): 1=λευκός, 0= άλλο
- Φύλο (sex): 1=άρρεν, 2=θήλυ
- Κάπνισμα (smoking): 1=καπνιστής, 2= μη καπνιστής
- Διάρκεια εργασίας (Empleng): 1=λιγότερο από 10 χρόνια, 2=10-20 χρόνια, 3=περισσότερα από 20 χρόνια
- Σκόνη (ποσοστό σκόνης στον εργασιακό χώρο) (dust): 1=ψηλό, 2=μέτριο, 3=χαμηλό.

Το πρόβλημα που καλούνταν να απαντήσουν ήταν αν οι επεξηγηματικές μεταβλητές είναι σημαντικές στην εμφάνιση αυτής της ασθένειας. Οι πρώτες δυο στήλες των δεδομένων καταγράφουν την συχνότητα των εργατών με ή χωρίς την ασθένεια στις αντίστοιχες τιμές των επεξηγηματικών μεταβλητών.

Αρχικά εισάγουμε τα δεδομένα μας και τα διαβάζουμε:

```
logistic1 <- read.csv("C:/Users/Maria Trentou/Desktop/log
istic1.txt", sep="")
View(logistic1)
attach(logistic1)
```

I	Yes	No	Dust	race	sex	smoking	Empleng

1	3	37	1	1	1	1	1
2	0	74	2	1	1	1	1
3	2	258	3	1	1	1	1
4	25	139	1	2	1	1	1
5	0	88	2	2	1	1	1
6	3	242	3	2	1	1	1
7	0	5	1	1	2	1	1
8	1	93	2	1	2	1	1
9	3	180	3	1	2	1	1
10	2	22	1	2	2	1	1
11	2	145	2	2	2	1	1
12	3	260	3	2	2	1	1
13	0	16	1	1	1	2	1
14	0	35	2	1	1	2	1
15	0	134	3	1	1	2	1
16	6	75	1	2	1	2	1

17	1	47	2	2	1	2	1
18	1	122	3	2	1	2	1
19	0	4	1	1	2	2	1
20	1	54	2	1	2	2	1
21	2	169	3	1	2	2	1
22	1	24	1	2	2	2	1
23	3	142	2	2	2	2	1
24	4	301	3	2	2	2	1
25	8	21	1	1	1	1	2
26	1	50	2	1	1	1	2
27	1	187	3	1	1	1	2
28	8	30	1	2	1	1	2
29	0	5	2	2	1	1	2
30	0	33	3	2	1	1	2
31	0	0	1	1	2	1	2
32	1	33	2	1	2	1	2

33	2	94	3	1	2	1	2
34	0	0	1	2	2	1	2
35	0	4	2	2	2	1	2
36	0	3	3	2	2	1	2
37	2	8	1	1	1	2	2
38	1	16	2	1	1	2	2
39	0	58	3	1	1	2	2
40	1	9	1	2	1	2	2
41	0	0	2	2	1	2	2
42	0	7	3	2	1	2	2
43	0	0	1	1	2	2	2
44	0	30	2	1	2	2	2
45	1	90	3	1	2	2	2
46	0	0	1	2	2	2	2
47	0	4	2	2	2	2	2
48	0	4	3	2	2	2	2

49	31	77	1	1	1	1	3
50	1	141	2	1	1	1	3
51	12	495	3	1	1	1	3
52	10	31	1	2	1	1	3
53	0	1	2	2	1	1	3
54	0	45	3	2	1	1	3
55	0	1	1	1	2	1	3
56	3	91	2	1	2	1	3
57	3	176	3	1	2	1	3
58	0	1	1	2	2	1	3
59	0	0	2	2	2	1	3
60	0	2	3	2	2	1	3
61	5	47	1	1	1	2	3
62	0	39	2	1	1	2	3
63	3	182	3	1	1	2	3
64	3	15	1	2	1	2	3

65	0	1	2	2	1	2	3
66	0	23	3	2	1	2	3
67	0	2	1	1	2	2	3
68	3	187	2	1	2	2	3
69	2	340	3	1	2	2	3
70	0	0	1	2	2	2	3
71	0	2	2	2	2	2	3
72	0	3	3	2	2	2	3

και προσαρμόζουμε το μοντέλο μας:

```
out1<-glm( cbind(Yes, No)~dust+race+sex+smoking+Empleng,
family=binomial)
out1

Call:  glm(formula = cbind(Yes, No) ~ dust + race + sex +
 smoking +
    Empleng, family = binomial)

Coefficients:
(Intercept)         dust         race          sex      s
moking      Empleng
    -0.4852      -1.3751       0.2463      -0.2590      -
0.6292       0.3856

Degrees of Freedom: 64 Total (i.e. Null);  59 Residual
Null Deviance:          322.5
Residual Deviance: 69.51  AIC: 188.2
```

Οπότε έχουμε τις εκτιμήσεις των παραμέτρων, την απόκλιση deviance και τους βαθμούς ελευθερίας του μοντέλου καθώς και την τιμή του κριτηρίου AIC.

Με την παρακάτω εντολή παίρνουμε περισσότερες πληροφορίες για τις παραμέτρους του μοντέλου:

```
> summary(out1)

Call:
glm(formula = cbind(Yes, No) ~ dust + race + sex + smoking +
    Empleng, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q
-3.4126  -0.7573  -0.2421   0.3688
    Max
 1.9804

Coefficients:
            Estimate Std. Error z value
(Intercept)  -0.4852     0.6060  -0.801
dust         -1.3751     0.1155 -11.901
race          0.2463     0.2061   1.195
sex          -0.2590     0.2116  -1.224
smoking      -0.6292     0.1931  -3.259
Empleng       0.3856     0.1069   3.607
            Pr(>|z|)
(Intercept)  0.42331
dust         < 2e-16 ***
race         0.23203
sex          0.22095
smoking      0.00112 **
Empleng      0.00031 ***
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 322.527  on 64  degrees of freedom
Residual deviance:  69.509  on 59  degrees of freedom
AIC: 188.19

Number of Fisher Scoring iterations: 5
```

Βλέπουμε την ελάχιστη και την μέση τιμή των υπολοίπων deviance στην αρχή, έπειτα βλέπουμε τους εκτιμητές των παραμέτρων μας, καθώς και το τυπικό σφάλμα της καθεμιάς και το p-value. Από το τελευταίο εξάγουμε το συμπέρασμα ότι οι μεταβλητές dust, smoking και Empleng είναι στατιστικά σημαντικές για την πρόβλεψη της ασθένειας του πνεύμονα, κάτι που επιβεβαιώνεται και από τα deviance υπόλοιπα στον anova.

Έπειτα θα κάνουμε ανάλυση απόκλισης:

```
> anova(out1)
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Yes, No)

Terms added sequentially (first to last)
```

```
        Df Deviance Resid. Df
NULL                       64
dust     1  221.963        63
race     1    1.054        62
sex      1    5.967        61
smoking  1   10.726        60
Empleng  1   13.308        59
        Resid. Dev
NULL        322.53
dust        100.56
race         99.51
sex          93.54
smoking      82.82
Empleng      69.51
```

Στην αρχή μας εξηγεί ότι έχουμε διωνυμικό μοντέλο και συνάρτηση σύνδεσης την logit και βλέπουμε και τα υπόλοιπα deviance αναλυτικότερα.

Συνεχίζουμε λοιπόν την λογιστική παλινδρόμηση κρατώντας μόνο τις στατιστικά σημαντικές παραμέτρους. Προσαρμόζουμε μοντέλο:

```
> out2<-glm( cbind(Yes, No)~dust+smoking+Empleng, family=
binomial)
```

Κάνουμε ανάλυση απόκλισης και αναζητούμε και το p-value:

```
> anova(out2,out1)
Analysis of Deviance Table

Model 1: cbind(Yes, No) ~ dust + smoking + Empleng
Model 2: cbind(Yes, No) ~ dust + race + sex + smoking + Emplen
g
  Resid. Df Resid. Dev Df Deviance
1        61     72.562
2        59     69.509  2   3.0527
>
> 1-pchisq(3.053,2)
[1] 0.2172949
```

Το p-value παίρνει αυτήν την μορφή γιατί θέλουμε να εξετάσουμε ποιο από τα δύο μοντέλα ταιριάζει καλύτερα στα δεδομένα, κάτι που όπως έχουμε πει γίνεται με την X^2 κατανομή.

Η μηδενική μας υπόθεση είναι ότι το νέο μοντέλο ταιριάζει καλύτερα στα δεδομένα μας. Όμως p-value = 0.2172949 > 0.05, άρα απορρίπτεται η υπόθεσή μας.

Στην συνέχεια θα βρούμε εκτιμητές των παραμέτρων του δεύτερου μοντέλου, που έχει τις λιγότερες μεταβλητές για να δούμε ποιο είναι τελικά το προσαρμοζόμενο μοντέλο.

```
> summary(out2)

Call:
```

```
glm(formula = cbind(Yes, No) ~ dust + smoking + Empleng, famil
y = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q
-3.3421  -0.7700  -0.2518   0.4001
    Max
 2.0523

Coefficients:
            Estimate Std. Error z value
(Intercept) -0.14177    0.34120  -0.415
dust        -1.46572    0.10578 -13.856
smoking     -0.67781    0.18871  -3.592
Empleng      0.33313    0.08861   3.760
            Pr(>|z|)
(Intercept) 0.677783
dust         < 2e-16 ***
smoking     0.000328 ***
Empleng     0.000170 ***
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 322.527  on 64  degrees of freedom
Residual deviance:  72.562  on 61  degrees of freedom
AIC: 187.24

Number of Fisher Scoring iterations: 5
```

Άρα και οι τρεις μεταβλητές που μας έχουν απομείνει, ακόμα και στο καινούργιο μοντέλο είναι στατιστικά σημαντικές, συνεπώς το προσαρμοζόμενο μοντέλο μας είναι:

$$\log\frac{\hat{\pi}}{1-\hat{\pi}} = -0.14177 - 1.46572 \cdot dust - 0.67781 \cdot smoking + 0.33313 \cdot Empleng.$$

Έχοντας το προσαρμοζόμενο μοντέλο μπορούμε να υπολογίσουμε την εκτιμώμενη πιθανότητα κάποιος εργάτης να πάσχει από ασθένεια του πνεύμονα για κάθε συνδυασμό των επεξηγηματικών μεταβλητών. Για παράδειγμα, αν ο εργάτης δουλεύει σε χώρο με υψηλό ποσοστό σκόνης (dust=1), καπνίζει (smoking=1) και δουλεύει πάνω από 20 χρόνια (Empleng=3) έχουμε:

$$\log\frac{\hat{\pi}}{1-\hat{\pi}} = -0.14177 - 1.46572 \cdot 1 - 0.67781 \cdot 1 + 0.33313 \cdot 3 = -1.28591$$

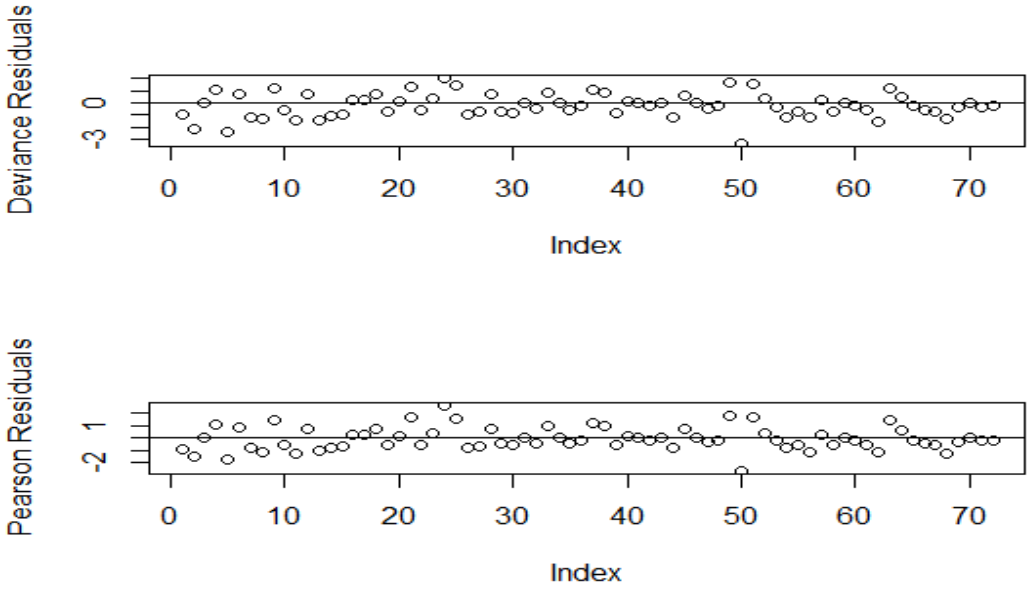
Άρα $\hat{\pi} = 0.2165$.

Τέλος κάνουμε μια ανάλυση των υπολοίπων Deviance και Pearson:

```
> residuals(out2, type="d")
          1           2           3           4           5
  6
-1.01112785 -2.18492306  0.03337745  1.05930402 -2.38265754  0.78600
008
          7           8           9          10          11
 12
-1.15218087 -1.35054620  1.20244845 -0.64235463 -1.41172867  0.67924
307
         13          14          15          16          17
 18
-1.49239385 -1.07494649 -1.01392707  0.24210300  0.23378161  0.67100
850
         19          20          21          22          23
 24
-0.74619693  0.10394697  1.33677067 -0.58548725  0.39372115  2.05225
780
         25          26          27          28          29
 30
 1.49151953 -0.94215644 -0.76417675  0.72544140 -0.66877671 -0.83273
904
         31          32          33          34          35
 36
 0.00000000 -0.43343323  0.88110339  0.00000000 -0.59817207 -0.25108
027
         37          38          39          40          41
 42
 1.04689406  0.83623911 -0.78766664  0.09325771  0.00000000 -0.27363
867
         43          44          45          46          47
 48
 0.00000000 -1.17367854  0.64751598  0.00000000 -0.42856681 -0.20685
139
         49          50          51          52          53
 54
 1.71825610 -3.34213262  1.57828362  0.41914791 -0.35176827 -1.14748
773
         55          56          57          58          59
 60
-0.69863438 -1.25301136  0.24410484 -0.69863438  0.00000000 -0.24191
165
         61          62          63          64          65
 66
-0.61128684 -1.57719361  1.20318680  0.53824354 -0.25255310 -0.58560
110
         67          68          69          70          71
 72
-0.72477260 -1.36106257 -0.35333036  0.00000000 -0.35716402 -0.21149
426
> residuals(out2, type="pear")
          1           2           3           4           5
  6
-0.94535299 -1.55751672  0.03350853  1.09147978 -1.69847122  0.85776
035
          7           8           9          10          11
 12
-0.84251659 -1.16707971  1.39123912 -0.60853460 -1.25426475  0.73131
877
         13          14          15          16          17
 18
-1.07391310 -0.76324957 -0.71764281  0.24581181  0.24358053  0.77245
880
         19          20          21          22          23
 24
-0.53695655  0.10578012  1.66583217 -0.54375649  0.40972260  2.62599
183
         25          26          27          28          29
 30
```

```
 1.60218002 -0.84268234 -0.69200893  0.74984259 -0.47823387 -0.59038
552
         31           32           33           34           35
 36
 0.00000000 -0.40853019  1.00018002  0.00000000 -0.42774538 -0.17800
793
         37           38           39           40           41
 42
 1.19196123  1.00011236 -0.55770997  0.09454217  0.00000000 -0.19375
077
         43           44           45           46           47
 48
 0.00000000 -0.83470237  0.74057705  0.00000000 -0.30479021 -0.14646
181
         49           50           51           52           53
 54
 1.77849560 -2.65712352  1.72114034  0.42524849 -0.25263516 -0.81437
316
         55           56           57           58           59
 60
-0.52573925 -1.14642147  0.25005104 -0.52573925  0.00000000 -0.17168
494
         61           62           63           64           65
 66
-0.59074669 -1.12419537  1.39228570  0.56308506 -0.18001533 -0.41485
545
         67           68           69           70           71
 72
-0.52978675 -1.23313690 -0.34015897  0.00000000 -0.25458012 -0.14982
818
> par(mfrow=c(2,1))
> plot(residuals(out2, type="d"), xlab="Index", ylab="Deviance
 Residuals")
> abline(h=0)
> plot(residuals(out2, type="pear"), xlab="Index", ylab="Pears
on Residuals")
> abline(h=0)
```

Εξάγονται λοιπόν τα εξής διαγράμματα:



Και στα δυο διαγράμματα παρατηρούμε ότι το εύρος τιμών τόσο των deviance υπολοίπων, όσο και των Pearson υπολοίπων κυμαίνεται γύρω από το μηδέν, συνεπών έχουμε κάνει καλή επιλογή συνάρτησης σύνδεσης και κατ' επέκταση ορθή επιλογή μοντέλου.

Κεφάλαιο 4

Poisson Παλινδρόμηση

4.1. Κατανομή Poisson

Η συνάρτηση πιθανότητας της κατανομής είναι:

$$f(y) = \frac{e^{-\mu} \cdot \mu^{y}}{y!}.$$

Η παράμετρος μ είναι θετική και μάλιστα είναι η μέση τιμή της κατανομής καθώς και η διασπορά, δηλαδή:

$$E(y) = \mu \ \ και \ \ Var(y) = \mu.$$

4.2. Poisson παλινδρόμηση

Στην Poisson παλινδρόμηση οι μεταβλητές απόκρισης ακολουθούν Poisson κατανομή, συνήθως τέτοια παραδείγματα έχουμε όταν έχουμε δεδομένα συχνοτήτων.

Θεωρούμε ότι μια μεταβλητή y_i μπορεί να πάρει τις τιμές 0,1,…, n με $y_i \sim Poisson(\mu_i)$. Τότε η συνάρτηση σύνδεσης της μέσης τιμής E(y_i) με τις επεξηγηματικές μεταβλητές είναι:

$$g(\mu_i) = n_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k = x_i^T \beta.$$

Έστω ότι E(y_i) = $\mu_i = n_i \theta_i$, όπου για παράδειγμα y_i είναι ο αριθμός των αιτήσεων για αποζημίωση αγροτών για ζημία στη καλλιέργεια φασολιών. Αυτό θα εξαρτάται από τον αριθμό των αγροτών που όντως έχουν υποστεί ζημιά, n_i καθώς και άλλες μεταβλητές που επηρεάζουν το θ_i, όπως οι καιρικές συνθήκες που επικρατούσαν στην περιοχή ή ο τύπος χώματος της καλλιέργειας ή το είδος φυτοφαρμάκων που έχουν χρησιμοποιηθεί σε αυτήν. Για την ανάλυση τέτοιου τύπου δεδομένων, μπορούμε να χρησιμοποιήσουμε το μοντέλο:

$$\theta_i = e^{x_i^T \beta},$$

και συνεπώς το γενικευμένο γραμμικό μοντέλο γίνεται:

$$E(y_i) = \mu_i = n_i \cdot e^{x_i^T \beta}, y_i \sim Poisson(\mu_i).$$

Η εκτίμηση της παραμέτρου β γίνεται με την χρήση λογαριθμικής πιθανοφάνειας για γενικευμένα γραμμικά μοντέλα.

Έστω $Y_1, \ldots, Y_N$ οι ανεξάρτητες μεταξύ τους μεταβλητές απόκρισης για τις οποίες έχουμε $Y_i \sim Poisson(\lambda_i)$. Τότε η ln- πιθανοφάνεια είναι:

$$l(\beta; y) = \sum_{i=1}^{N} y_i \log(\lambda_i) - \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \log(y_i!).$$

Αν τώρα όλα τα λ_i είναι διαφορετικά, τότε $\beta = [\lambda_1, \ldots, \lambda_N]^T$ και $\hat{\lambda}_i = y_i$, άρα η μέγιστη τιμή της ln-πιθανοφάνειας είναι:

$$l(\beta_{max}; y) = \sum_{i=1}^{N} y_i \log(y_i) - \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \log(y_i!).$$

4.3. Εκτίμηση παραμέτρων

Για να εκτιμήσουμε τις παραμέτρους χρησιμοποιούμε την μέθοδο μέγιστης πιθανοφάνειας και έχουμε:

$$L(y; \beta) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \cdot {\mu_i}^{y_i}}{y_i!} = \frac{\prod_{i=1}^{n} {\mu_i}^{y_i} \cdot exp(-\sum_{i=1}^{n} \mu_i)}{\prod_{i=1}^{n} y_i!}.$$

Παίρνουμε τη λογαριθμική συνάρτηση πιθανοφάνειας:

$$log\{L(y; \beta)\} = \log \prod_{i=1}^{n} \left(\frac{e^{-\mu_i} \cdot {\mu_i}^{y_i}}{y_i!} \right) = \sum_{i=1}^{n} \left[y_i \cdot x_i \beta - e^{x_i \beta} - \log(y_i!) \right].$$

Μεγιστοποιούμε την σχέση με την βοήθεια της παραγώγου, δηλαδή:

$$\frac{\partial}{\partial \beta} [log\{L(y; \beta)\}] = 0,$$

$$\sum_{i=1}^{n} \left[y_i \cdot x_i - e^{x_i \beta} \cdot x_i \right] = 0 \Longrightarrow \sum_{i=1}^{n} (y_i - \mu_i) x_i = 0.$$

Ισοδύναμα με χρήση πινάκων έχουμε ότι $X(y - \mu) = 0$.

4.4. Καταλληλότητα του μοντέλου

4.4.1. Deviance σε Poisson μοντέλο

Έστω ότι έχουμε ένα μοντέλο με $p < N$ παραμέτρους, τότε έχουμε ότι ο εκτιμητής μέγιστης πιθανοφάνειας β μπορεί να υπολογίσει εκτιμήτριες $\hat{\lambda}_i$ και αφού $\hat{\lambda}_i = \hat{y}_i$, αφού $E(Y_i) = \lambda_i$, άρα η μέγιστη τιμή της λογαριθμικής συνάρτησης πιθανοφάνειας θα είναι:

$$l(\beta; y) = \sum_{i=1}^{N} y_i \log(\hat{y}_i) - \sum_{i=1}^{N} \hat{y}_i - \sum_{i=1}^{N} \log(y_i!),$$

Συνεπώς η deviance είναι:

$$D = 2[l(\beta_{max}; y) - l(\beta; y)] = 2\left[\sum y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - \sum (y_i - \hat{y}_i)\right].$$

Επίσης, στα περισσότερα μοντέλα μπορούμε να δείξουμε ότι

$$\sum y_i = \sum \hat{y}_i.$$

Οπότε η D γίνεται:

$$D = 2 \sum o_i \cdot \log \frac{o_i}{e_i},$$

όπου το o_i δηλώνει την τιμή του y_i και το e_i την αναμενόμενη τιμή $\hat{y}_i$.

Η τιμή της D μπορεί να υπολογιστεί και η τιμή της να συγκριθεί με την X^2_{N-p}.

4.4.2. Υπόλοιπα

Επειδή $Var(Y_i) = E(Y_i)$ στην κατανομή Poisson, το τυπικό σφάλμα του Y_i εκτιμάται ως $\sqrt{e_i}$, οπότε τα υπόλοιπα Pearson είναι:

$$r_i = \frac{o_i - e_i}{\sqrt{e_i}}.$$

Στην Poisson κατανομή η σχέση των υπολοίπων με τον έλεγχο καλής προσαρμογής χ-τετράγωνο είναι:

$$X^2 = \sum_i r_i^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}.$$

Η απόκλιση ενός μοντέλου Poisson ξέρουμε πως είναι:

$$D = 2\left[\sum_i o_i \log \frac{o_i}{e_i} - \sum_i (o_i - e_i)\right].$$

οπότε αφού στα περισσότερα μοντέλα έχουμε ότι

$$\sum o_i = \sum e_i$$

έχουμε ότι:

$$D = 2\left[\sum_i o_i \log \frac{o_i}{e_i}\right] \kappa\alpha\iota \;\; d_i = sign(o_i - e_i)\sqrt{2\left[o_i \log \frac{o_i}{e_i} - (o_i - e_i)\right]},$$

$$\delta\eta\lambda\alpha\delta\acute{\eta}\; D = \sum d_i^2.$$

Τα X^2 και D χρησιμοποιούνται για να ελέγξουμε την καλή προσαρμογή αφού οι τιμές τους μπορούν να υπολογιστούν άμεσα από τα δεδομένα του προσαρμοζόμενου μοντέλου. Μπορούν επίσης να συγκριθούν με την X^2 με N-p βαθμούς ελευθερίας, όπου η τιμή του p είναι ο αριθμός των παραμέτρων που έχουν εκτιμηθεί.

4.5. Κριτήρια επιλογής μοντέλου

4.5.1. Κριτήριο AIC

Έχουμε:

$$AIC = -2\sum_{i=1}^{n}\{-\hat{\mu}_i + y_i \log \hat{\mu}_i - \log(y_i!)\} + 2p.$$

4.5.2. Κριτήριο BIC

Έχουμε:

$$BIC = -2\sum_{i=1}^{n}\{-\hat{\mu}_i + y_i \log \hat{\mu}_i - \log(y_i!)\} + p \log n.$$

4.6. Λογαριθμικά – γραμμικά μοντέλα

Για μοντέλα που βασίζονται στην Poisson κατανομή έχουμε:

$$f(y;\mu) = \prod_{i=1}^{N} \mu_i^{y_i} e^{-\mu_i} y_i!\,, με\ \mu\ διάνυσμα\ των\ \mu_i.$$

Οι αναμενόμενες συχνότητες είναι: $E(y_i) = \mu_i, i = 1, \dots, N$.

Οπότε η συνάρτηση σύνδεσης που παίρνουμε είναι η:

$$\log\{E(y_i)\} = x_i^T \beta$$

το οποίο μπορεί να χαρακτηριστεί ως λογαριθμικό γραμμικό μοντέλο.

4.6.1. Υπερδιασπορά

Υπερδιασπορά έχουμε όταν η διασπορά είναι μεγαλύτερη από την μέση τιμή, παρόλο που στην Poisson κατανομή είναι ίσες. Η αρνητική διωνυμική μας δίνει αυτό το φαινόμενο, όπου έχουμε ένα εναλλακτικό μοντέλο, το:

$$var(y_i) = \varphi E(y_i), \qquad \varphi > 1$$

με φ μια παράμετρο που μπορούμε να εκτιμήσουμε. Η υπερδιασπορά συνήθως οφείλεται στην έλλειψη ανεξαρτησίας μεταξύ των παρατηρήσεων.

4.7. Παράδειγμα Poisson παλινδρόμησης

Στην εφαρμογή αυτή χρησιμοποιούμε στοιχεία από τις καταστροφές, που προκλήθηκαν σε συγκεκριμένους τύπους αεροσκαφών του Ναυτικού των Ηνωμένων Πολιτειών. Έχουμε συλλέξει δείγμα από 30 δοκιμαστικές αποστολές, οι οποίες πραγματοποιήθηκαν κατά τη διάρκεια του πολέμου του Βιετνάμ και χρησιμοποιήθηκαν δύο τύποι αεροσκαφών. Τα δεδομένα μας αποτελούνται από 4 μεταβλητές, τις:

damage: είναι ο αριθμός των τμημάτων του αεροσκάφους, που σημειώθηκαν καταστροφές κατά τη διάρκεια των δοκιμαστικών αποστολών.

type: είναι μια δίτιμη μεταβλητή, που αναφέρεται στον τύπο του αεροσκάφους, που χρησιμοποιήθηκε σε κάθε αποστολή (0 για το A4, 1 για το A6).

bombload: είναι το φορτίο της βόμβας (σε τόνους).

airexp: η εμπειρία του πληρώματος (συνολικός αριθμός μηνών).

Σε αυτή την εφαρμογή μπορούμε να χρησιμοποιήσουμε την κατανομή Poisson για τον αριθμό καταστροφών (damage) που προκλήθηκαν στα αεροσκάφη μετά από κάθε αποστολή. Συνεπώς, θεωρώντας ότι η μεταβλητή απόκρισης είναι η μεταβλητή damage μπορούμε να προσαρμόσουμε ένα μοντέλο της παλινδρόμησης Poisson στα δεδομένα μας με επεξηγηματικές μεταβλητές τις type, bombload, airexp.

Εισάγουμε τα δεδομένα μας στην R:

```
> aircraft <- read_excel("C:/Users/Maria Trentou/Desktop/
aircraft.xlsx")
> View(aircraft)
> library(readxl)
> attach(aircraft)
```

Τα δεδομένα μας δίνονται από τον παρακάτω πίνακα:

I	damage	Type	bombload	airexp
1	0	0	4	91.5
2	1	0	4	84.0

3	0	0	4	76.5
4	0	0	5	69.0
5	0	0	5	61.5
6	0	0	5	80.0
7	1	0	6	72.5
8	0	0	6	65.0
9	0	0	6	57.5
10	2	0	7	50.0
11	1	0	7	103.0
12	1	0	7	95.5
13	1	0	8	88.0
14	1	0	8	80.5
15	2	0	8	73.0
16	3	1	7	116.1
17	1	1	7	100.6
18	1	1	7	85.0

19	1	1	10	69.4
20	2	1	10	53.9
21	0	1	10	112.3
22	1	1	12	96.7
23	1	1	12	81.1
24	2	1	12	65.6
25	5	1	8	50.0
26	1	1	8	120.0
27	1	1	8	104.4
28	5	1	14	88.9
29	5	1	14	73.7
30	7	1	14	57.8

Η μεταβλητή type είναι κατηγορική, συνεπώς πρέπει να την δούμε ως διάνυσμα, συνεπώς έχουμε:

```
> type<-as.factor(type)
```

Τώρα είμαστε έτοιμοι δούμε τα χαρακτηριστικά της μεταβλητής απόκρισης, δηλαδή την damage:

```
> summary(damage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   0.250   1.000   1.533   2.000   7.000 
```

Η μέγιστη τιμή λοιπόν για την μεταβλητή αυτή είναι 7 και η ελάχιστη 0, έχει διάμεσο 1 και δειγματικό μέσο 1.533. Στο 1ο και στο 3ο τεταρτημόριο λαμβάνει τις τιμές 0,250 και 2 αντίστοιχα.

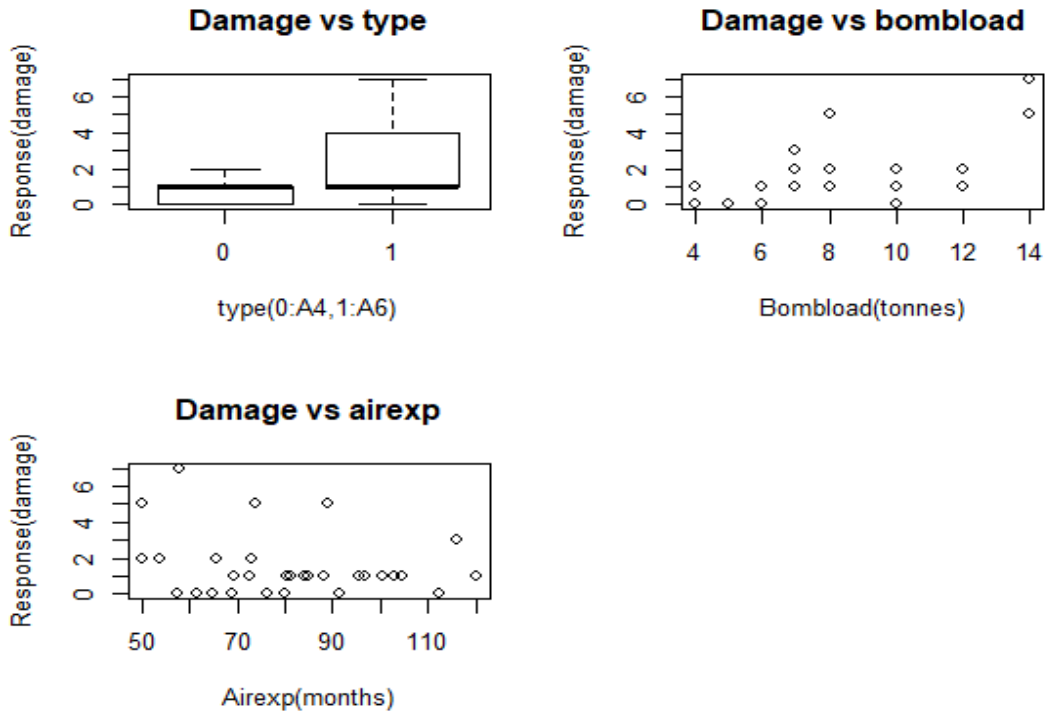
Η δειγματική διασπορά της είναι:

```
> var(damage)
[1] 3.154023
```

Πριν προσαρμόσουμε μοντέλο, ας ελέγξουμε γραφικά την σχέση της κάθε επεξηγηματικής μεταβλητής με την μεταβλητή απόκρισης:

```
> par(mfrow=c(2,2))
> plot(damage~type,ylab="Response(damage)",xlab="type(0:A
4,1:A6)",main="Damage vs type")
> plot(damage~bombload,ylab="Response(damage)",xlab="Bomb
load(tonnes)",main ="Damage vs bombload")
> plot(damage~airexp,ylab="Response(damage)",xlab="Airexp
(months)",main="Damage vs airexp")
```

Και έχουμε:



Παρατηρούμε λοιπόν ότι η μεταβλητή damage σχετίζεται με όλες τις μεταβλητές μας. Επειδή η type είναι κατηγορική, δημιουργούμε θηκόγραμμα, ενώ για τις υπόλοιπες έχουμε διάγραμμα διασποράς.

Ωστόσο πρέπει να εξετάσουμε αν υπάρχει συσχέτιση μεταξύ των μεταβλητών, κάτι που γίνεται με τον συντελεστή συσχέτισης Pearson.

Οι συντελεστές συσχέτισης δίνονται με την εντολή:

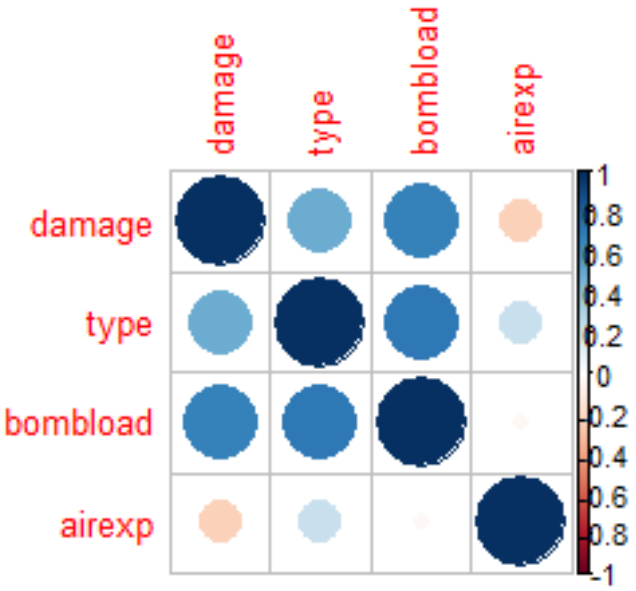
```
> cor.aircraft<-cor(aircraft)
```

```
> cor.aircraft
             damage      type    bombload      airexp
damage    1.0000000 0.4963423  0.67608073 -0.23337694
type      0.4963423 1.0000000  0.71237621  0.22322455
bombload  0.6760807 0.7123762  1.00000000 -0.03241923
airexp   -0.2333769 0.2232245 -0.03241923  1.00000000
```

Ο παραπάνω πίνακας είναι ο πίνακας συσχετίσεων. Πριν τον αναλύσουμε όμως, ας δούμε άλλον έναν τρόπο για να ελέγξουμε την συσχέτιση.

```
> library(graphics)
> library(corrplot)
> corrplot(cor.aircraft)
```

Έχουμε:



Παρατηρούμε λοιπόν, τόσο από τον πίνακα, όσο και από το διάγραμμα ότι υπάρχει συσχέτιση μεταξύ των μεταβλητών και συγκεκριμένα την μεγαλύτερη συσχέτιση έχουν οι damage – bombload και type – bombload.

Η μεγάλη συσχέτιση μεταξύ των type – bombload μας προβληματίζει, διότι πιθανότατα να έχουμε πολυσυγγραμμικότητα στο μοντέλο.

Πάμε τώρα να προσαρμόσουμε το μοντέλο για Poisson παλινδρόμηση με συνάρτηση σύνδεσης log:

```
> air.model<-glm(damage~type+bombload+airexp,family="poisson",data=aircraft)
```

Κάνουμε summary και έχουμε:

```
> summary(air.model)

Call:
glm(formula = damage ~ type + bombload + airexp, family =
 "poisson",
```

```
    data = aircraft)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6418  -1.0064  -0.0180   0.5581   1.9094

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.406023   0.877489  -0.463   0.6436
type         0.568772   0.504372   1.128   0.2595
bombload     0.165425   0.067541   2.449   0.0143 *
airexp      -0.013522   0.008281  -1.633   0.1025
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 53.883  on 29  degrees of freedom
Residual deviance: 25.953  on 26  degrees of freedom
AIC: 87.649

Number of Fisher Scoring iterations: 5
```

Έτσι έχουμε τους εκτιμητές των παραμέτρων του μοντέλου, τα τυπικά σφάλματά τους, τις τιμές των στατιστικών ελέγχων Wald για τους συντελεστές του μοντέλου και τα p-values. Έχουμε και την τιμή της στατιστικής συνάρτησης Deviance και τέλος την τιμή του AIC. Αυτό που εμείς παρατηρούμε είναι ότι στατιστικά σημαντική είναι μόνο η μεταβλητή bombload, κάτι που το αναμέναμε λόγω της πολυσυγγραμμικότητας και πρακτικά μας λέει ότι ο αριθμός των καταστροφών πράγματι σχετίζεται με το φορτίο την βόμβας, όταν οι άλλες μεταβλητές παραμένουν σταθερές.

Επίσης, από την τιμή της ελεγχοσυνάρτησης deviance του μοντέλου με μόνο τον σταθερό όρο και του υποψήφιου μοντέλου, βλέπουμε ότι υπάρχει ένδειξη καλής προσαρμογής στο μοντέλο, αφού και οι τρεις μεταβλητές μαζί μειώνουν σημαντικά την τιμή της. Αυτό βέβαια μπορούμε να το διαπιστώσουμε κάνοντας έναν έλεγχο υποθέσεων, η μηδενική υπόθεση θα είναι να ισχύει το μοντέλο με μόνο τον σταθερό όρο και εναλλακτική να ισχύει το υποψήφιο μοντέλο. Αυτό στην R γίνεται ως εξής:

```
> ddev<-air.model$null.deviance-air.model$deviance
> df<-air.model$df.null-air.model$df.residual
> pvalue<-1-pchisq(ddev,df)
> data.frame(ddev,df,pvalue)
     ddev df       pvalue
1 27.9299  3 3.757192e-06
```

Παρατηρούμε ότι η τιμή του p-value για τον έλεγχο αυτόν είναι πολύ μικρή, συνεπώς απορρίπτουμε την μηδενική υπόθεση, άρα το μοντέλο μας είναι προτιμότερο από αυτό που περιέχει μόνο τον σταθερό όρο.

Ένα ακόμα πρόβλημα που καλούμαστε να εξετάσουμε είναι αυτό της υπερμεταβλητότητας, αφού από την θεωρία μας πρέπει η μέση τιμή και η διασπορά να είναι ίσες. Αυτό μπορούμε να το εξετάσουμε μέσω της συνάρτησης φ = deviance/df, με deviance να είναι η τιμή της ελεγχοσυνάρτησης και df να είναι οι βαθμοί ελευθερίας της χ-τετράγωνο κατανομής που ακολουθεί η deviance.

Συνεπώς έχουμε:

```
> deviance(air.model)/air.model$df.residual
[1] 0.9981985
```

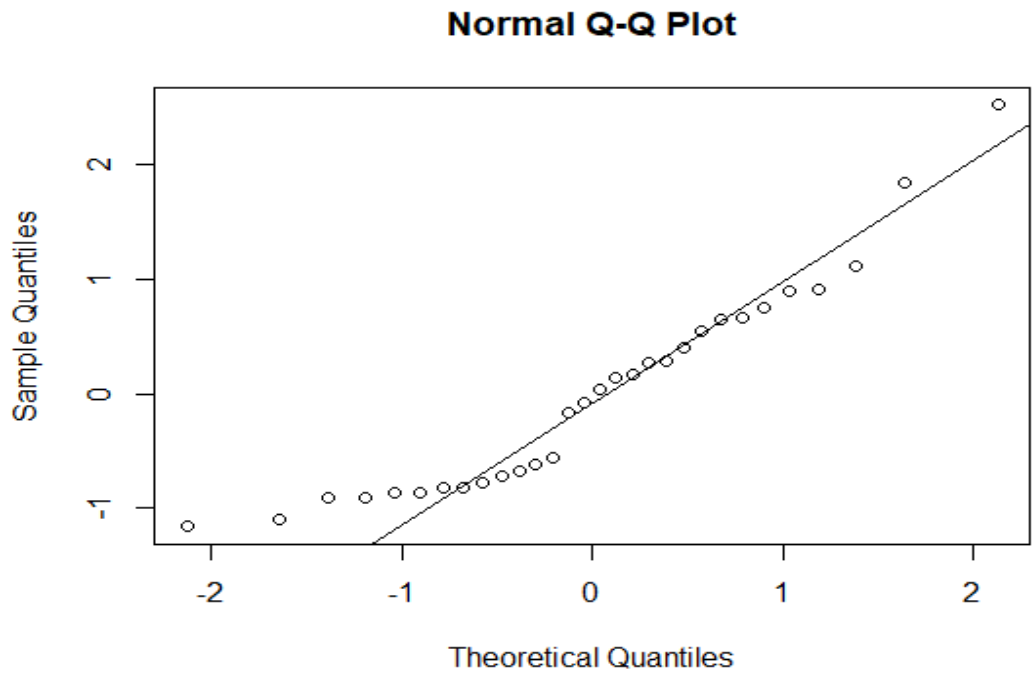
Αφού είναι μικρότερη του 1, συμπεραίνουμε ότι δεν έχουμε υπερμεταβλητότητα στα δεδομένα μας.

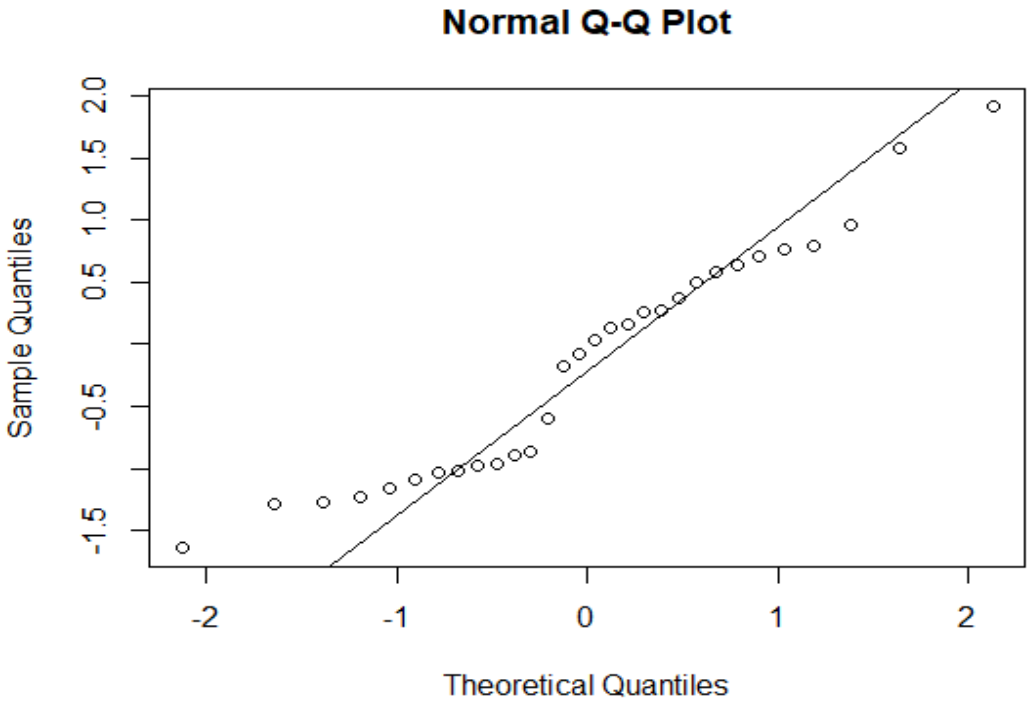
Τέλος, πριν την επιλογή μοντέλου θα πρέπει να εξετάσουμε τις υποθέσεις του μοντέλου. Αυτό συνήθως γίνεται μέσω των υπολοίπων deviance και Pearson, τα οποία υπολογίζονται:

```
> res_pearson<-residuals(air.model,type="pearson")
> res_deviance<-residuals(air.model)
```

Αν και ξέρουμε πως δεν ακολουθούν κανονική κατανομή, ας κάνουμε έναν έλεγχο κανονικότητας:

```
> qqnorm(res_pearson)
> qqline(res_pearson)
> qqnorm(res_deviance)
> qqline(res_deviance)
```

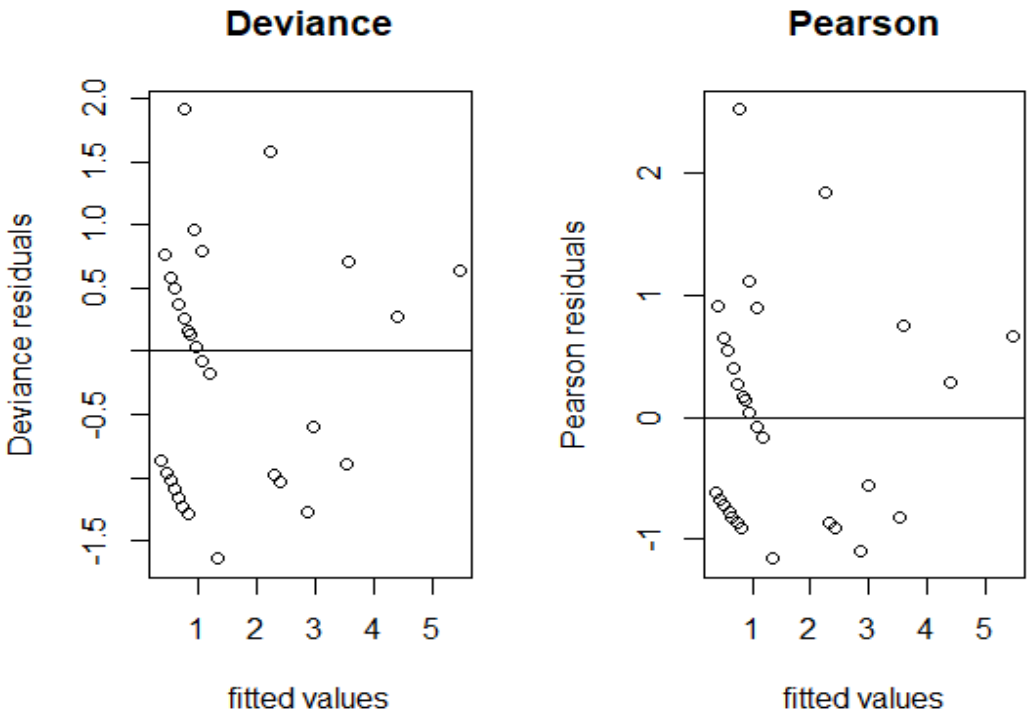




Από τα γραφήματα, σαφώς και βλέπουμε την μη ύπαρξη κανονικότητας, αλλά αυτό που ενδεχομένως μπορούμε να σχολιάσουμε είναι η καλή προσαρμογή του μοντέλου μας, αφού για τα υπόλοιπά μας βλέπουμε ότι κινούνται πάνω σε μια σχετικά καλά ορισμένη ευθεία.

Εξετάζουμε τώρα τα γραφήματα των υπολοίπων deviance και Pearson σε σχέση με τις προσαρμοζόμενες τιμές $\hat{y}_i = \hat{\mu}_i = exp(x_i^T \beta)$:

```
> par(mfrow=c(1,2))
> plot(fitted.values(air.model),res_deviance,xlab="fitted
 values",ylab="Deviance residuals",main="Deviance")
> abline(h=0)
> plot(fitted.values(air.model),res_pearson,xlab="fitted
values",ylab="Pearson residuals",main="Pearson")
> abline(h=0)
```



Αυτό που μπορούμε να πούμε πλέον είναι ότι τα υπόλοιπα συμπεριφέρονται πολύ τυχαία.

Ήρθε η ώρα λοιπόν να επιλέξουμε το βέλτιστο μοντέλο. Αυτό θα γίνει με την βοήθεια των κριτηρίων AIC, BIC, κάνοντας συγκρίσεις πιθανών μοντέλων:

```
> x1.model<-glm(damage~type,family="poisson",data=aircraf
t)
> x2.model<-glm(damage~bombload,family="poisson",data=air
craft)
> x3.model<-glm(damage~airexp,family="poisson",data=aircr
aft)
> x1x2.model<-glm(damage~type+bombload,family="poisson",d
ata=aircraft)
> x1x3.model<-glm(damage~type+airexp,family="poisson",dat
a=aircraft)
> x2x3.model<-glm(damage~bombload+airexp,family="poisson"
,data=aircraft)
> x1x2x3.model<-glm(damage~type+bombload+airexp,family="p
oisson",data=aircraft)
```

Έπειτα κάνουμε συγκρίσεις μέσω anova:

```
> anova(x1.model,x1x2x3.model,test="Chisq")
Analysis of Deviance Table

Model 1: damage ~ type
Model 2: damage ~ type + bombload + airexp
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        28     38.283
2        26     25.953  2    12.33 0.002101 **
---
```

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

```
> anova(x2.model,x1x2x3.model,test="Chisq")
Analysis of Deviance Table

Model 1: damage ~ bombload
Model 2: damage ~ type + bombload + airexp
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        28     29.206
2        26     25.953  2   3.2527   0.1966
```

```
> anova(x3.model,x1x2x3.model,test="Chisq")
Analysis of Deviance Table

Model 1: damage ~ airexp
Model 2: damage ~ type + bombload + airexp
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        28     50.537
2        26     25.953  2   24.584 4.588e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1
```

```
> anova(x1x2.model,x1x2x3.model,test="Chisq")
Analysis of Deviance Table

Model 1: damage ~ type + bombload
Model 2: damage ~ type + bombload + airexp
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        27     28.634
2        26     25.953  1   2.6812   0.1015
```

```
> anova(x1x3.model,x1x2x3.model,test="Chisq")
Analysis of Deviance Table

Model 1: damage ~ type + airexp
Model 2: damage ~ type + bombload + airexp
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        27     32.192
2        26     25.953  1   6.2386   0.0125 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1
```

```
> anova(x2x3.model,x1x2x3.model,test="Chisq")
Analysis of Deviance Table

Model 1: damage ~ bombload + airexp
Model 2: damage ~ type + bombload + airexp
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        27     27.220
```

```
2        26     25.953  1   1.2667   0.2604
```

Κάνουμε και τον έλεγχο μέσω των AIC BIC:

```
> AIC(x1x2x3.model,x1x2.model,x1x3.model,x2x3.model,x1.mo
del,x2.model,x3.model)
              df       AIC
x1x2x3.model  4  87.64922
x1x2.model    3  88.33037
x1x3.model    3  91.88781
x2x3.model    3  86.91589
x1.model      2  95.97952
x2.model      2  86.90196
x3.model      2 108.23330
```

```
> BIC(x1x2x3.model,x1x2.model,x1x3.model,x2x3.model,x1.mo
del,x2.model,x3.model)
              df       BIC
x1x2x3.model  4  93.25401
x1x2.model    3  92.53396
x1x3.model    3  96.09140
x2x3.model    3  91.11948
x1.model      2  98.78192
x2.model      2  89.70435
x3.model      2 111.03570
```

Παρατηρούμε μέσω των τιμών των p-value στις συγκρίσεις ότι το πλήρες μοντέλο είναι καλύτερο από κάθε άλλο μοντέλο, παρόλο που περιέχει μη στατιστικά σημαντική μεταβλητή. Η διαφορά αυτή ίσως να οφείλεται στο φαινόμενο της πολυσυγγραμικότητας των δεδομένων μας.

Συμπερασματικά, το μοντέλο Poisson για τα δεδομένα μας θα έχει την μορφή:

$$damage_i \sim Poisson(\mu_i)$$

$$\log(\mu_i) = \beta_0 + \beta_1 type_i + \beta_2 bombload_i + \beta_3 airexp_i, i = 1, \dots, 30.$$

Το προσαρμοζόμενο μοντέλο θα είναι:

$$\log \hat{\mu}_i = b_0 + b_1 type_i + b_2 bombload_i + b_3 airexp_i$$
$$= -0.406 + 0.569 type_i + 0.165 bombload_i - 0.014 airexp_i.$$

Ισοδύναμα έχουμε:

$$\hat{\mu}_i = exp(= -0.406 + 0.569 type_i + 0.165 bombload_i - 0.014 airexp_i).$$

Τα συμπεράσματά μας λοιπόν είναι:

Ο αναμενόμενος αριθμός καταστροφών, που προκλήθηκαν στον τύπο αεροσκάφους A6 (type=1) είναι exp(0.569) 1.77 φορές μεγαλύτερος από τον αντίστοιχο αριθμό καταστροφών του αεροσκάφους A4 (type=0) όταν και τα δύο αεροσκάφη κατείχαν το ίδιο φορτίο βόμβας και ήταν επανδρωμένα με πλήρωμα με ίδια εμπειρία.

Κάθε επιπλέον τόνος, που χρησιμοποιήθηκε για το φορτίο της βόμβας φαίνεται να αυξάνει τον αναμενόμενο αριθμό καταστροφών, που προκλήθηκαν κατά $(\exp(0.165) - 1) \times 100\% \approx (1.18 - 1) \times 100\% = 18\%$ υπό την προϋπόθεση ότι τα αεροσκάφη είναι του ίδιου τύπου (A4 ή A6) και είναι επανδρωμένα με πλήρωμα με την ίδια εμπειρία.

Κάθε επιπλέον μήνας εμπειρίας του πληρώματος μειώνει την αναμενόμενη τιμή των καταστροφών των αεροσκαφών κατά ποσοστό $(1 - \exp(0.014)) \times 100\% \approx (1 - 0.99) \times 100\% = 1\%$ υπό την προϋπόθεση ότι τα αεροσκάφη είναι του ίδιου τύπου (A4 ή A6) και κατέχουν το ίδιο φορτίο βόμβας.

Βιβλιογραφία

1) AN INTRODUCTION TO GENERALIZED LINEAR MODELS SECOND EDITION, Annette J. Dobson.(2002)
2) Generalized linear models, P. McCullagh and J.A. Nelder, second edition. (1989).
3) Linear and Generalized Linear Mixed Models and Their Applications, Jiming Jiang, Springer.
4) An introduction to Generalized Linear Models, George H. Dunteman Moon-Ho R. Ho. SAGE Publications.(2006)
5) Applied Logistic Regression, Hosmer D. W. and Lemeshov S. second edition, John Wiley & Sons(2001).
6) Extending The Linear Model with R, Faraway J.J. Chapman & Hall/CRC (2006)
7) An Introduction to R, Venables W.N., Smith D.M. an the R Development Core Team(1999)
8) Generalized Linear Models: An Applied Approach, Olsson U., Studentlitteratur. Lund (2002).
9) Introduction to Linear Regression Analysis, Montgomery, D., Peck, E. & Vining, G.G. (2006).. Fourth edition. Wiley. Hoboken. New Jersey.
10) Generalized Linear Models: A Unified Approach, Gill, J. (2001).. Sage University Papers Series on Quantitative Applications in the Social Sciences. Sage. New York.
11) Everitt B.S.and HothornT. (2010). A Handbook of Statistical Analyses Using R. Second Edition .Chapman & Hall /CRC.
12) Σημειώσεις μαθήματος 'Μοντέλα Παλινδρόμησης' ακ. Έτους 2016-2017 εαρινό εξάμηνο, Διδάσκων: Δημητρακοπούλου Θεοδώρα, Τμήμα Στατιστικής και Αναλογιστικών Χρηματοοικονομικών Μαθηματικών, Πανεπιστήμιο Αιγαίου.
13) Στατιστική Θεωρία και Εφαρμογές, Κοκολάκης Γ. & Φουσκάκης Δ. εκδόσεις Συμεών (2009).
14) Διπλωματική Εργασία, Γενικευμένα Γραμμικά Μοντέλα με χρήση του στατιστικού πακέτου R, Νταϊλιάνας Χρίστος.

15) Διπλωματική Εργασία, Γενικευμένα Γραμμικά μοντέλα με χρήση της Γλώσσας προγραμματισμού R, Αθανάσιος Στεφ. Τάτσιος.
16) Διπλωματική Εργασία, Γραμμική και μη γραμμική παλινδρόμηση με εφαρμογές στην R, Ελένη Ι. Κουτσουδάκη.
17) Εισαγωγή στην R, Πρόχειρες σημειώσεις, Κωνσταντίνος Φωκιανός & Χαράλαμπος Χαραλάμπους Τμήμα Μαθηματικών & Στατιστικής Πανεπιστήμιο Κύπρου.
18) Διπλωματική εργασία, ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΟ ΣΤΑΤΙΣΤΙΚΟ ΠΑΚΕΤΟ R, ΠΑΝΑΓΙΩΤΟΠΟΥΛΟΥ ΜΑΡΙΑ – ΕΛΕΥΘΕΡΙΑ.
19) Παρουσίαση και δεδομένα από την κ. Παρπούλα Χριστίνα από το Θερινό σχολείο 'R and Big Data Analytics 2018, Σάμος'.