Ανάλυση παλινδρόμησης

Γιώργος Ευσταθίου, gefstath@psych.uoa.gr

Εισαγωγικές έννοιες

- Έννοια της ευθύγραμμης σχέσης – έννοια της γραμμής παλινδρόμησης

Προσαρμόζουμε ένα μοντέλο πρόβλεψης στα δεδομένα μας, με σκοπό να χρησιμοποιήσουμε το μοντέλο αυτό για να προβλέψουμε την τιμή της εξαρτημένης μεταβλητής ή της μεταβλητής κριτήριο (outcome) από τις τιμές μίας ή περισσότερων ανεξάρτητων μεταβλητών ή μεταβλητών πρόβλεψης (predictors).

Απλή ανάλυση παλινδρόμησης: Μία μεταβλητή πρόβλεψης
Πολλαπλή ανάλυση παλινδρόμησης: Πολλές μεταβλητές πρόβλεψης

Το μοντέλο πρόβλεψης είναι ευθύγραμμο, δηλαδή βασίζεται σε μία ευθεία γραμμή.

Μία ευθεία γραμμή περιγράφεται από:

την κλίση (slope ή gradient) ή οποία συνήθως συμβολίζεται με β_1 ή (b_1 στο SPSS) και

το σημείο στο οποίο η γραμμή τέμνει τον κάθετο άξονα του γραφήματος (intercept ή constant) που συμβολίζεται συνήθως με β_0 (b_0 στο SPSS).

Η εξίσωση που περιγράφει το μοντέλο (τη γραμμή ή το επίπεδο παλινδρόμησης) είναι η ακόλουθη:

$Y_i = (\beta_0 + \beta_1 X_1)\ [+ e_i]$ ή $Y_i = (b_0 + b_1 X_1)\ [+ e_i]$

$Y_i = (\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_n X_n)\ [+ e_i]$ ή $Y_i = (b_0 + b_1 X_1 + b_2 X_2 \ldots + b_n X_n)\ [+ e_i]$

Όπου: Y_i είναι η μεταβλητή που προσπαθούμε να προβλέψουμε
b_0 είναι το σημείο που η γραμμή τέμνει τον κάθετο άξονα
b_1 είναι η κλίση της γραμμής παλινδρόμησης
X_i είναι το σκορ του συμμετέχοντα στη μεταβλητή πρόβλεψης

Τα b_0 και b_1 ονομάζονται συντελεστές παλινδρόμησης (regression coefficients)

e_i κατάλοιπο ή σφάλμα (residual term) – η διαφορά μεταξύ της τιμής που προβλέπεται βάσει της γραμμής για τον συμμετέχοντα i και της παρατηρούμενης τιμής για το συμμετέχοντα i.

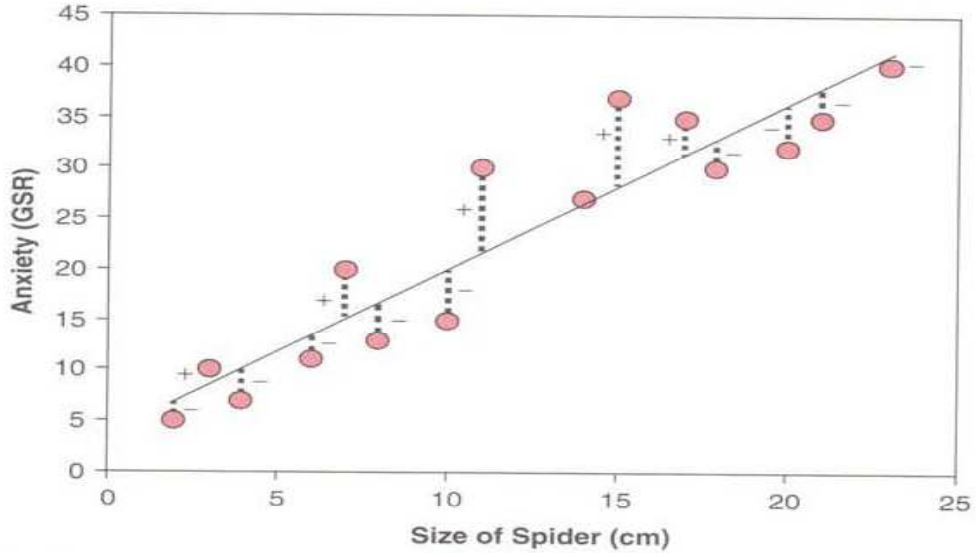
Επιλέγουμε τη γραμμή εκείνη που περιγράφει καλύτερα τα δεδομένα, δηλαδή τη γραμμή στα σημεία της οποίας τείνουν να συγκεντρωθούν η τιμές της μεταβλητής που επιθυμούμε να προβλέψουμε (παλινδρόμηση). Προκειμένου να εντοπίσουμε τη γραμμή που περιγράφει καλύτερα, χρησιμοποιούμε διάφορες μεθόδους, συνήθως αυτή των ελαχίστων τετραγώνων.

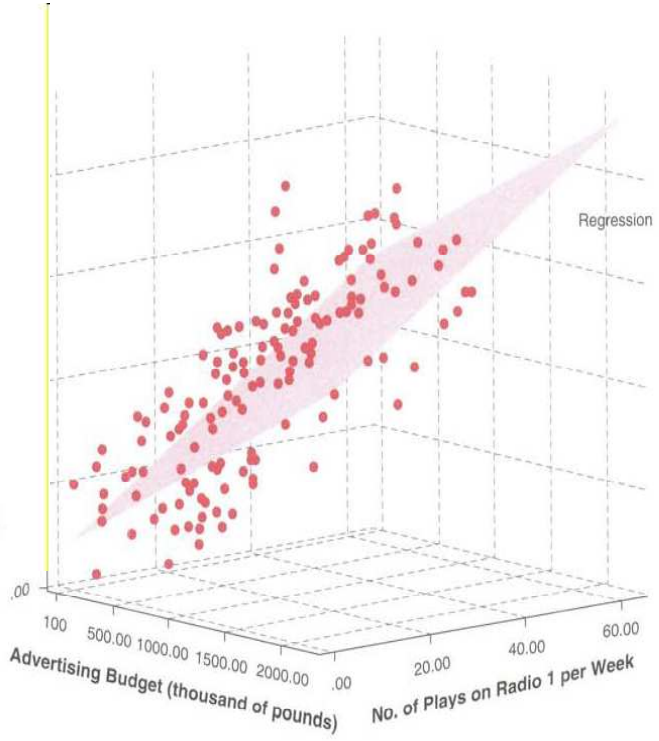
Η καλύτερη γραμμή, είναι η γραμμή εκείνη η οποία από όλες τις γραμμές που θα μπορούσαν να σχεδιαστούν προσφέρει τη μικρότερη απόσταση μεταξύ των παρατηρούμενων δεδομένων και της γραμμής.

Αν δηλώσουμε τα παρατηρούμενα δεδομένα ως σημεία σε ένα διάγραμμα, κάποια θα πέσουν ακριβώς πάνω στη γραμμή, κάποια θα βρίσκονται πάνω από τη γραμμή (θετικές διαφορές) και κάποια κάτω από τη γραμμή (αρνητικές διαφορές).

Οι διαφορές αυτές ονομάζονται κατάλοιπα (residuals). Αν όμως αθροίζαμε θετικές και αρνητικές διαφορές θα αλληλοεξουδετερώνονταν, οπότε υψώνουμε τις διαφορές στο τετράγωνο πριν τις αθροίσουμε. Επιλέγουμε τη γραμμή με τη μικρότερο άθροισμα υψωμένων στο τετράγωνο διαφορών.

Προσοχή: όταν μιλάμε για πολλαπλή ανάλυση παλινδρόμησης μιλάμε για επίπεδο και όχι για γραμμή.


Anxiety (GSR)
Size of Spider (cm)
0
5
10
15
20
25
30
35
40
45


Regression
Advertising Budget (thousand of pounds)
No. of Plays on Radio 1 per Week
100
500.00
1000.00
1500.00
2000.00
20.00
40.00
60.00

- Συνάφεια

Μέτρηση της ευθύγραμμης σχέσης μεταξύ δύο μεταβλητών.

<u>Συντελεστές ή Δείκτες συνάφειας</u> (μέγεθος, κατεύθυνση, στατιστική σημαντικότητα)

Pearson r

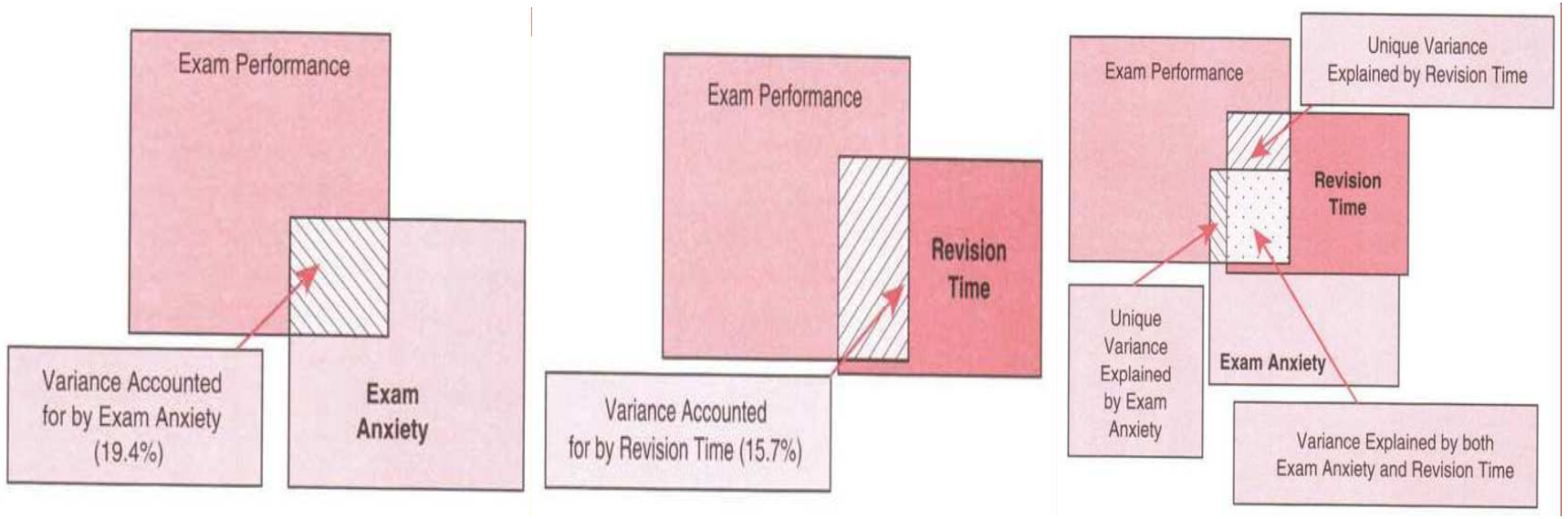
Αιτιότητα: Κατεύθυνση, τρίτες μεταβλητές

<u>Συντελεστής προσδιοριστίας (R^2):</u> το ποσοστό της διακύμανσης μίας μεταβλητής που εξηγείται από μία άλλη μεταβλητή

Π.χ. Αν r=-0,38 τότε $R^2=(-0,38)^2$ = 0,14 (δηλαδή 14% της μεταβλητής εξηγείται από την άλλη μεταβλητή και το 86% ΔΕΝ εξηγείται)

- Μερική συνάφεια

Μερική συνάφεια (partial correlation): Η συνάφεια μεταξύ δύο μεταβλητών, όταν η επίδραση άλλων μεταβλητών διατηρείται σταθερή (ταυτόχρονος έλεγχος της επίδρασης δύο μεταβλητών, η συμμετοχή της άλλης/ων μεταβλητής/ων αφαιρείται και από την ανεξάρτητη και από την εξαρτημένη).



semi partial (part) correlation: Έλεγχος της επίδρασης της τρίτης μεταβλητής μόνο στη μία μεταβλητή και όχι και στις δύο (όπως γίνεται στη μερική συνάφεια, δηλαδή μόνο από την ανεξάρτητη και όχι από την εξαρτημένη)

Αποτελέσματα της ανάλυσης

- Έλεγχος καλής προσαρμογής ή ταυτοσημίας (Goodness of fit)

Αφού εντοπίσουμε την καλύτερη γραμμή, αξιολογούμε πόσο καλά ταιριάζει στα δεδομένα. Μπορεί να είναι η καλύτερη δυνατή γραμμή και παρόλα αυτά να μην είναι ικανοποιητική.

Χρησιμοποιούμε το μέσο όρο του Y ως μοντέλο και συγκρίνουμε με τις παρατηρούμενες τιμές (Συνολικό άθροισμα τετραγώνων)
Χρησιμοποιούμε τις τιμές που προκύπτουν από το μοντέλο με τις παρατηρούμενες (Άθροισμα τετραγώνων των residuals)
Αφαιρώντας το δεύτερο με το πρώτο διαπιστώνουμε πόσο βελτιώνεται η πρόβλεψή μας αν χρησιμοποιήσουμε το μοντέλο αντί του μέσου όρου (Άθροισμα τετραγώνων μοντέλου).

Από το πηλίκο του αθροίσματος τετραγώνων του μοντέλου δια του συνολικού αθροίσματος τετραγώνων προκύπτει το R^2, το οποίο ερμηνεύεται ακριβώς με τον ίδιο τρόπο.

Το R^2 x 100 μας δίνει το ποσοστό της διακύμανσης της μεταβλητής που προσπαθούμε να προβλέψουμε που εξηγείται από το μοντέλο.

Model [c]

Mode	R	R	Adjuste R	Std. Error the	Change R Chang	F	df1	df2	Sig. F	Durbin Watso
1	,286 [a]	,082	,081	2,8534	,082	101,21	1	1136	,000	
2	,317 [b]	,101	,099	2,8251	,019	23,83	1	1135	,000	1,869

a. Predictors: (Constant), GCAS subscale 2 -

b. Predictors: (Constant), GCAS subscale 2 - Liking,

c. Dependent Variable:

- Στατιστική σημαντικότητα μοντέλου

Υπολογίζεται ένα F-test που μας δείχνει ότι η κλίση της γραμμής παλινδρόμησης διαφέρει από το μηδέν (δηλαδή από την οριζόντια θέση).

Βασίζεται στη βελτίωση που επιφέρει το μοντέλο στην πρόβλεψη δια του βαθμού ανακρίβειας του μοντέλου και αναφέρεται στο μοντέλο συνολικά και όχι σε μεμονωμένες μεταβλητές πρόβλεψης.

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	824,137	1	824,137	101,218	,000[a]
	Residual	9249,527	1136	8,142		
	Total	10073,663	1137			
2	Regression	1014,409	2	507,204	63,546	,000[b]
	Residual	9059,255	1135	7,982		
	Total	10073,663	1137			

a. Predictors: (Constant), GCAS subscale 2 - Liking

b. Predictors: (Constant), GCAS subscale 2 - Liking, f_webb

c. Dependent Variable: Σενάριο Δ3

- Μεμονωμένες μεταβλητές πρόβλεψης (predictors)

Για κάθε μεταβλητή πρόβλεψης (predictor) εξάγεται ένας συντελεστής παλινδρόμησης (b).

Το SPSS συμβολίζει τους συντελεστές με B.

b_0: Η τιμή της εξαρτημένης όταν η μεταβλητή πρόβλεψης (ή όλες οι μεταβλητές πρόβλεψης) παίρνουν την τιμή 0. Μπορεί να έχει ή να μην έχει νόημα π.χ. αυτοεκτίμηση κοπέλας με βάρος μηδέν.

b_1: Η αλλαγή στην εξαρτημένη μεταβλητή που σχετίζεται με την αλλαγή μίας μονάδας στη μεταβλητή πρόβλεψης (ανεξάρτητη).

Με το t-test γίνεται έλεγχος για τη σημαντικότητα της συνεισφοράς στην πρόβλεψη.

Μας λέει πόσο επηρεάζει κάθε μεταβλητή πρόβλεψης την εξαρτημένη, όταν οι άλλες μεταβλητές πρόβλεψης κρατιούνται σταθερές.

Το πρόσημο καθορίζει την κατεύθυνση της σχέσης.

β: οι τυπικές αποκλίσεις που θα αλλάξει η εξαρτημένη ως συνέπεια της αλλαγής μίας τυπικής απόκλισης στην μεταβλητή πρόβλεψης. Ευκολότερη εκτίμηση της σημαντικότητας μίας μεταβλητής πρόβλεψης στο μοντέλο.

Διαστήματα εμπιστοσύνης: Όσο καλύτερο το μοντέλο τόσο πιο μικρά. Αν περιλαμβάνουν το μηδέν, σημαίνει ότι σε μερικά δείγματα η κατεύθυνση της σχέσης θα είναι αρνητική και σε άλλα θετική.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	,815	,428		1,905	,057	-,024	1,654					
	GCAS subscale 2 - Liking	,112	,011	,286	10,061	,000	,090	,134	,286	,286	,286	1,000	1,000
2	(Constant)	,318	,436		,730	,465	-,537	1,173					
	GCAS subscale 2 - Liking	,080	,013	,205	6,253	,000	,055	,105	,286	,182	,176	,740	1,351
	f_webb	,422	,086	,160	4,882	,000	,253	,592	,264	,143	,137	,740	1,351

a. Dependent Variable: Σενάριο Δ3

- Πρόβλεψη μεμονωμένων περιπτώσεων

Καταλληλότητα παροχής υπηρεσιών μέσω $= b_0 + b_1X_1 + b_2X_2$

X_1=26 και X_2=2 $= 0,318 + 0,080x26 + 0,422x2$

$= 0,318 + 2,08 + 0,844$

$= 3,242$

X_1=42 και X_2=4 κοκ

Είδη πολλαπλής ανάλυσης παλινδρόμησης

- Απλή (enter)

Όλες οι μεταβλητές πρόβλεψης εισέρχονται υποχρεωτικά στο μοντέλο

- Ιεραρχική (hierarchical)

Οι μεταβλητές πρόβλεψης εισέρχονται με προδιαγεγραμμένη σειρά στο μοντέλο, βάσει ευρημάτων προηγούμενων ερευνών. Μπορεί να εισέλθουν μεμονωμένα ή σε ομάδες.

- Κατά βήματα (stepwise)

Με προοδευτική προσθήκη: Ο υπολογιστής ξεκινά με ένα μοντέλο στο οποίο συμπεριλαμβάνεται μόνο το b_0 και στη συνέχεια αναζητά τη μεταβλητή πρόβλεψης που προβλέπει καλύτερα την εξαρτημένη (έχει το υψηλότερο r) και στη συνέχεια προστίθεται η μεταβλητή που έχει την υψηλότερη semi-partial με την εξαρτημένη και η διαδικασία συνεχίζεται όσο η μεταβλητή που προστίθεται βελτιώνει το μοντέλο πρόβλεψης.

Με προοδευτική αφαίρεση: Ο υπολογιστής ξεκινά με ένα μοντέλο στο οποίο συμπεριλαμβάνεται όλες οι μεταβλητές πρόβλεψης και σταδιακά αφαιρούνται μεταβλητές ανάλογα με τη συμβολή τους στο μοντέλο.

Κατά βήματα: Όπως η μέθοδος με προοδευτική προσθήκη, αλλά κάθε φορά που εισάγεται μία μεταβλητή πρόβλεψης, γίνεται έλεγχος για τη δυνατότητα διαγραφής της μεταβλητής πρόβλεψης με τη μικρότερη συμβολή στο μοντέλο.

Προϋποθέσεις εφαρμογής - παραδοχές

- Πολυσυγραμμικότητα (Multicollinearity)

Υπάρχει όταν υπάρχει υψηλή συνάφεια μεταξύ δύο ή περισσότερων μεταβλητών (>0,80). Αυξάνεται η πιθανότητα για σφάλμα τύπου ΙΙ και μειώνεται το μέγεθος του R.

VIF – Variance Inflation Factor: Δείκτης του κατά πόσο η μεταβλητή έχει ισχυρή ευθύγραμμη σχέση με κάποια άλλη μεταβλητή πρόβλεψης. Ουδός >10, ανησυχούμε αν το μέσο VIF είναι σημαντικά μεγαλύτερο του 1 (αθροίζουμε τους VIF και διαιρούμε δια τον αριθμό των μεταβλητών πρόβλεψης).
Tolerance (1/VIF): Πρόβλημα <0,1, ανησυχούμε από 0,2.

- Ομοσκεδασμός (Homoscedasticity)

Η διασπορά των καταλοίπων θα πρέπει να είναι σταθερή σε κάθε επίπεδο της μεταβλητής ή των μεταβλητών πρόβλεψης. Όταν τα επίπεδα σε κάθε επίπεδο των μεταβλητών πρόβλεψης έχουν πολύ διαφορετική διασπορά τότε μιλάμε για ετεροσκεδασμό (Heteroscedasticity).

- Ανεξαρτησία σφαλμάτων

Για οποιεσδήποτε δύο παρατηρήσεις θα πρέπει τα κατάλοιπα να μην έχουν συνάφεια ή να είναι ανεξάρτητα. Ελέγχεται με το τεστ Durbin-Watson. Παίρνει τιμές από 0 έως 4. Ως πρακτικό κανόνα θεωρούμε προβληματικές τιμές μικρότερες του 1 και μεγαλύτερες του 3.

- Κανονικότητα σφαλμάτων

Θεωρείται ότι τα κατάλοιπα είναι τυχαίες μεταβλητές που σχηματίζουν κανονική κατανομή. ΔΕΝ απαιτείται να σχηματίζουν κανονική κατανομή οι μεταβλητές πρόβλεψης. Κάνουμε K-S στα τυποποιημένα κατάλοιπα (standardized residuals) ή επισκοπούμε τα σχετικά γραφήματα.

- Ανεξαρτησία μετρήσεων

Κάθε τιμή της εξαρτημένης μεταβλητής έχει διαφορετική πηγή. Δεν μπορούμε να έχουμε μετρήσεις από τα ίδια άτομα σε διαφορετικές στιγμές

- Γραμμικότητα (linearity)

Η σχέση μεταξύ των μεταβλητών πρόβλεψης και της εξαρτημένης είναι ευθύγραμμη.

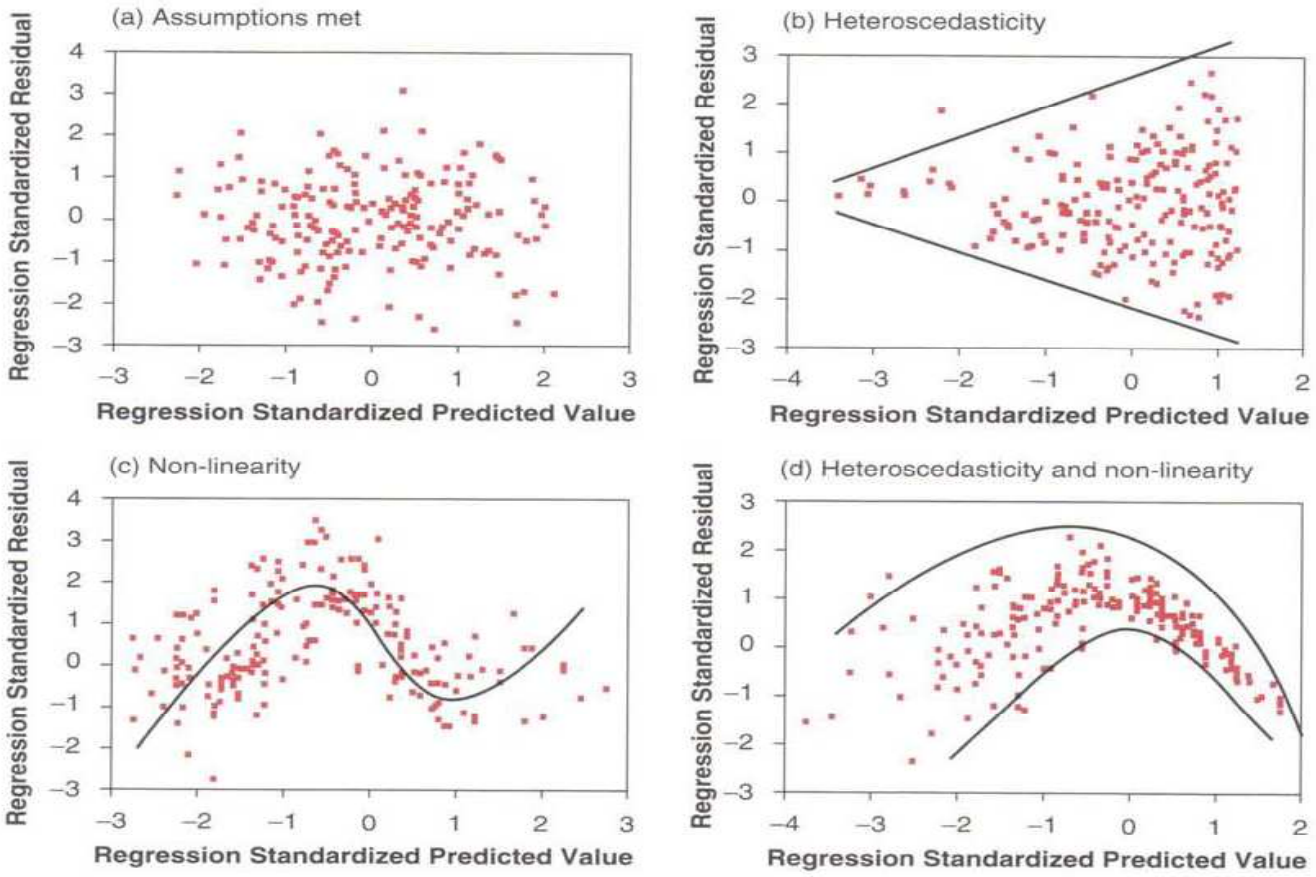


Figure 5.18 Plots of *ZRESID against *ZPRED

Έλεγχος

- Μέγεθος δείγματος

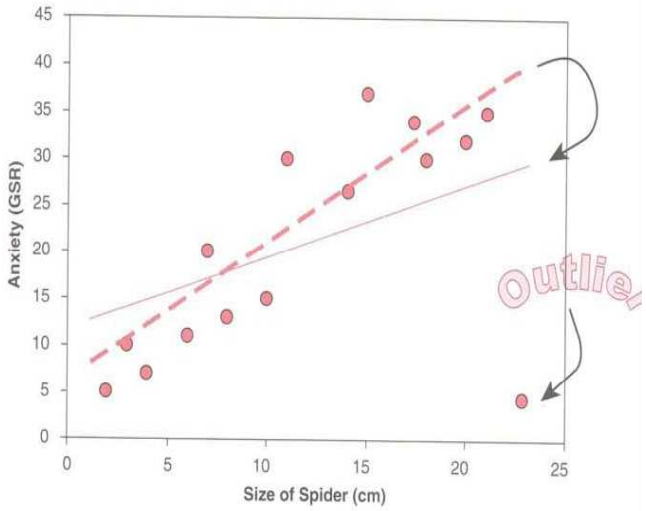
Εμπειρικός κανόνας: για τον έλεγχο της πολλαπλής συνάφειας 50 + 8 x μεταβλητές πρόβλεψης
για τον έλεγχο μεμονωμένων μεταβλητών 104 + αριθμός μεταβλητών προβλ.

Σε περίπτωση που επιλέξουμε stepwise, 40 για κάθε μεταβλητή πρόβλεψης.

Αν το δείγμα είναι πολύ μεγάλο, η ανάλυση διακύμανσης για τον έλεγχο του κατά πόσο η κλίση είναι διάφορη του μηδενός βγαίνει στατιστικά σημαντική γιατί αυξάνονται οι βαθμοί ελευθερίας (Τύπου Ι).

- Ακραίες τιμές (outliers)

Ακραία τιμή είναι εκείνη που διαφέρει σημαντικά από την κεντρική τάση των δεδομένων και επηρεάζει σημαντικά τους συντελεστές παλινδρόμησης.



- Κατάλοιπα (Residuals)

Οι διαφορές μεταξύ των τιμών της εξαρτημένης μεταβλητής που προβλέπονται από το μοντέλο και των τιμών που παρατηρούνται στο δείγμα. Ουσιαστικά αντιπροσωπεύουν το σφάλμα που εμπεριέχεται στο μοντέλο.

Τα απλά ή μη τυποποιημένα κατάλοιπα (unstandardized residuals) είναι στη μονάδα μέτρησης της εξαρτημένης μεταβλητής και η ερμηνεία τους παρουσιάζει δυσκολίες μεταξύ των μοντέλων.

Τα τυποποιημένα κατάλοιπα (standardized residuals) είναι εκφρασμένα σε z-τιμές (Μέσος όρος 0, τυπική απόκλιση 1, δηλαδή 95% μεταξύ, -1,96 και 1,96, 99% μεταξύ −2,58 και 2,58 και 99,9% μεταξύ −3,29 και 3,29. Μας ενδιαφέρουν πάνω από 3,29 και τα ποσοστά στο δείγμα 1% και 5%. (βλ. και studentised residuals).

- Επηρεάζουσες περιπτώσεις (influential cases)

Προσαρμοσμένη προβλεπόμενη τιμή (adjusted predicted values): η προβλεπόμενη τιμή για μία περίπτωση όταν εξαιρείται από τον υπολογισμό του μοντέλου.

DFFit: Η διαφορά μεταξύ της αρχικής τιμής και της προσαρμοσμένης προβλεπόμενης τιμής.

Mahalanobis distance: Απόσταση της περίπτωσης από τους μέσους όρους των μεταβλητών πρόβλεψης. Παίρνει τη μορφή της χ^2 κατανομής, με βαθμούς ελευθερίας ίσους με τον αριθμό των μεταβλητών πρόβλεψης. Για να καθοριστούν οι πολυμεταβλητές ακραίες τιμές, αναζητούμε την κρίσιμη τιμή του χ^2 για το επιθυμητό επίπεδο στατισικής σημαντικότητας (α). Για παράδειγμα η κρίσιμη τιμή του χ^2 για α=0,001 με df=3 είναι 16,266. Κάθε περίπτωση με τιμή μεγαλύτερη από αυτή είναι ακραία τιμή.

- Δυνατότητα γενίκευσης (cross validation)

- Προσαρμοσμένος συντελεστής προσδιοριστίας (adj. R^2)

Διόρθωση του R^2 ώστε να μην αφορά μόνο το συγκεκριμένο δείγμα.

- Διαίρεση του δείγματος

Χωρίζουμε με τυχαίο τρόπο το δείγμα μας σε δύο μικρότερα δείγματα, εκτελούμε την ανάλυση και συγκρίνουμε τα αποτελέσματα. [π.χ. Data > Select Cases > Random sample]

ή

Χρησιμοποιούμε το 80% του δείγματος για να κάνουμε την ανάλυση παλινδρόμησης. Στη συνέχεια υπολογίζουμε τιμές για το άλλο 20% του δείγματος βάσει των συντελεστών παλινδρόμησης που προέκυψαν και στη υπολογίζουμε τη συνάφεια μεταξύ των προβλεπόμενων και των παρατηρούμενων τιμών (είναι το R^2 για το μικρό δείγμα – αν υπάρχει μεγάλη διαφορά στα R^2 τότε η δυνατότητα γενίκευσης είναι περιορισμένη).

Table 5.2 How to report multiple regression

	B	*SE B*	β
Step 1			
Constant	134.14	7.54	
Advertising Budget	0.10	0.01	.58*
Step 2			
Constant	−26.61	17.35	
Advertising Budget	0.09	0.01	.51*
Plays on BBC Radio 1	3.37	0.28	.51*
Attractiveness	11.09	2.44	.19*

Note $R^2 = .34$ for Step 1; $\Delta R^2 = .33$ for Step 2 ($ps < .001$). * $p < .001$.