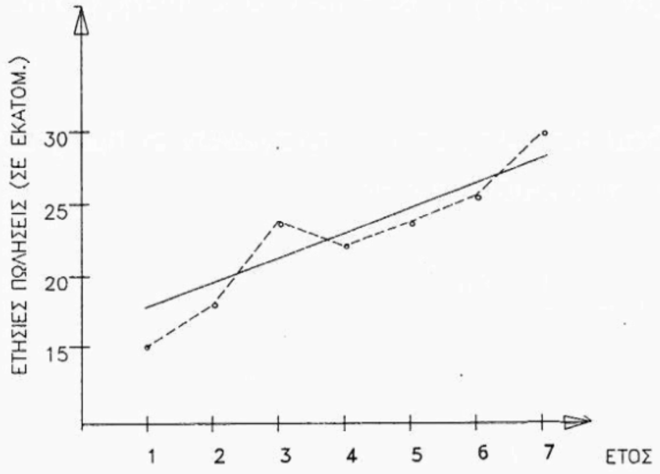
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Μονάδα Προβλέψεων & Στρατηγικής
Forecasting & Strategy Unit

Επιχειρηματικές Προβλέψεις: Μέθοδοι & Τεχνικές

Γραμμική Παλινδρόμηση Διάλεξη 8

Απλή Γραμμική Παλινδρόμηση

Επιχειρηματικές Προβλέψεις: Μέθοδοι & Τεχνικές



$$\hat{Y}_i = a + bX_i$$

$$\sum e_i^2 = \sum \left(Y_i - \hat{Y}_i\right)^2$$

$$\bar{Y} = \frac{\sum Y_i}{n}$$

$$\bar{X} = \frac{\sum X_i}{n}$$

$$b = \frac{\frac{\sum X_i Y_i}{n} - \bar{X}\bar{Y}}{\frac{\sum X_i{}^2}{n} - \bar{X}^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$a = \bar{Y} - b\bar{X}$$

FSU

Βασικές Υποθέσεις

Επιχειρηματικές Προβλέψεις: Μέθοδοι & Τεχνικές

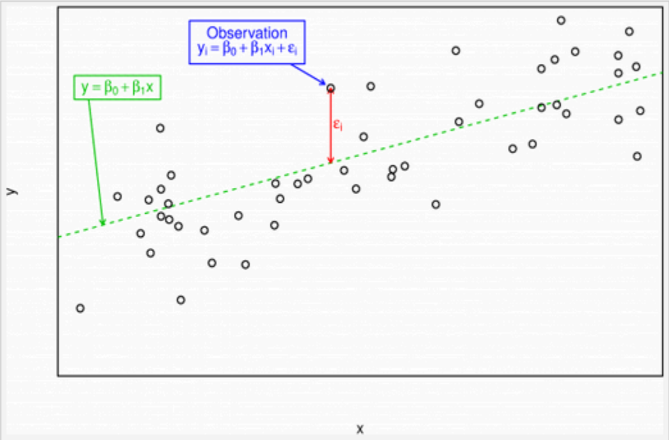


Figure 4.1: An example of data from a linear regression model.

https://www.otexts.org/fpp/4/1

- Έχουν μηδενική μέση τιμή
- Δεν παρουσιάζουν αυτοσυσχέτιση
- Είναι ασυσχέτιστα με την υπό – πρόβλεψη τιμή
- Είναι χρήσιμο να παρουσιάζουν κανονική κατανομή

Συντελεστής Συσχέτισης

- Βασική προϋπόθεση για την εφαρμογή της απλής γραμμικής παλινδρόμησης είναι ότι η τιμή μιας μεταβλητής εξαρτάται από την τιμή ή τη μεταβολή της τιμής κάποιας άλλης. Συχνά όμως δύο μεταβλητές μπορεί να σχετίζονται χωρίς να μπορεί να θεωρηθεί πως η τιμή της μίας επηρεάζει ή εξαρτάται από την τιμή της άλλης.
- Ο συντελεστής συσχέτισης ***r*** αποτελεί ένα μέτρο του βαθμού συσχέτισης που μπορεί να υπάρχει μεταξύ δύο μεταβλητών. Μπορεί να ερμηνευθεί με δύο τρόπους:
    - Ως ένδειξη της κατεύθυνσης της σχέσης ανάμεσα σε δύο μεταβλητές (πχ. αν οι τιμές τους αυξάνονται ή μειώνονται συγχρόνως ή αν η αύξηση της μιας συνεπάγεται μείωση της άλλης ή αν είναι ανεξάρτητες/ασυσχέτιστες μεταξύ τους)
    - Ως ένδειξη του βαθμού συσχέτισης, καθώς όσο η τιμή του συντελεστή απομακρύνεται από το μηδέν, τόσο πιο ισχυρή θεωρείται η συσχέτιση ανάμεσα στις δύο μεταβλητές.

Συντελεστής Συσχέτισης

Επιχειρηματικές Προβλέψεις: Μέθοδοι & Τεχνικές

Συνδιακύμανση των X και Y

$$Cov_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

Διακύμανση του X

$$Cov_{XX} = \frac{\sum(X_i - \bar{X})^2}{n} = Var_X = S_X^2$$

Διακύμανση του Y

$$Cov_{YY} = \frac{\sum(Y_i - \bar{Y})^2}{n} = Var_Y = S_Y^2$$

$$r_{XY} = \frac{Cov_{XY}}{\sqrt{Cov_{YY} \cdot Cov_{XX}}} = \frac{Cov_{XY}}{S_Y S_X}$$

$$|r_{XY}| \leq 1$$

Συντελεστής R^2

- Η συσχέτιση των τιμών που προκύπτουν από την εξίσωση της ευθείας παλινδρόμησης και των πραγματικών τιμών συμβολίζεται με R. Στην πράξη η συσχέτιση αυτή χρησιμοποιείται στην τετραγωνική της μορφή και ως εκ τούτου είναι ένας συντελεστής πάντα θετικός ($0< R^2 <1$).
- Αντιπροσωπεύει το ποσοστό της διακύμανσης της μεταβλητής Y που ερμηνεύεται από την ευθεία της γραμμικής παλινδρόμησης.

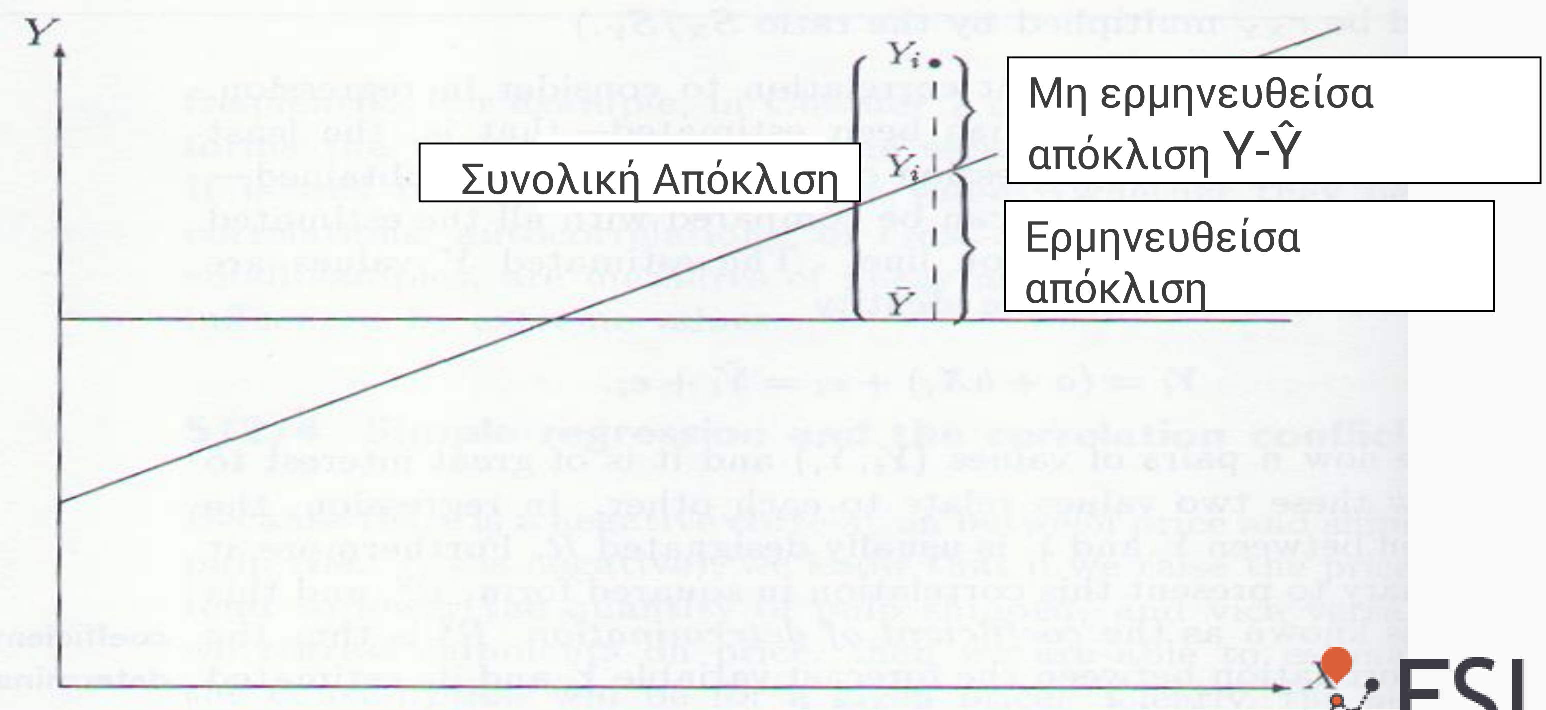
$$R^2 = \frac{\text{διακύμανση των τιμών } \hat{Y}}{\text{διακύμανση των τιμών } Y}$$

$$R^2 = \frac{\text{ερμηνευθείσα διακύμανση της } Y}{\text{συνολική διακύμανση της } Y}$$

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = r_{XY}^2$$

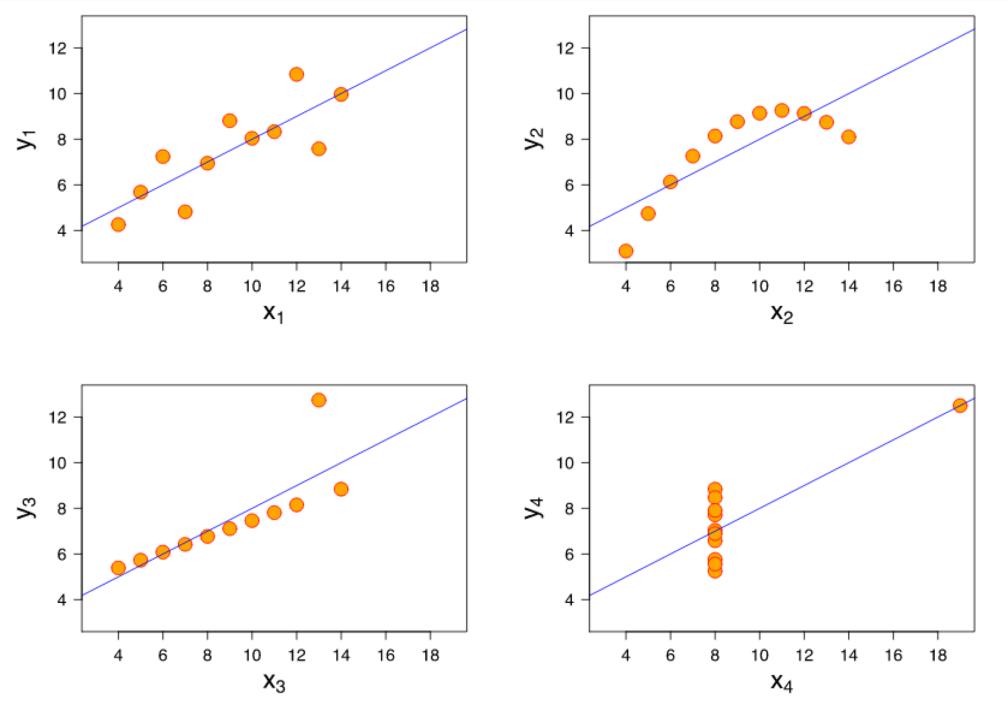
Συντελεστής R^2

Επιχειρηματικές Προβλέψεις: Μέθοδοι & Τεχνικές



Συντελεστής R^2

Επιχειρηματικές Προβλέψεις: Μέθοδοι & Τεχνικές



Στατιστικοί Δείκτες

- Θεωρώντας την εξίσωση παλινδρόμησης ως στατιστικό μοντέλο, υπολογίζονται κάποιοι στατιστικοί δείκτες οι οποίοι επιτρέπουν την εκτίμηση
  - Της πιθανότητας οι μελλοντικές τιμές της εξαρτημένης μεταβλητής να διαφέρουν από τις προβλεπόμενες κατά συγκεκριμένη ποσότητα
  - Της αξιοπιστίας του υπολογισμού της ευθείας παλινδρόμησης
  - Της ακρίβειας των συντελεστών α και b

Ο στατιστικός δείκτης F

Ο στατιστικός δείκτης F επιτρέπει την εκτίμηση της σημαντικότητας της εξίσωσης παλινδρόμησης, δηλαδή δίνει απάντηση στο ερώτημα αν υπάρχει σημαντική σχέση ανάμεσα στις μεταβλητές X και Y.

$$F = \frac{\dfrac{\sum\left(\hat{Y}_i - \bar{Y}\right)^2}{k-1}}{\dfrac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-k}}$$

$$F = \frac{\dfrac{R^2}{k-1}}{\dfrac{1-R^2}{n-k}}$$

Στατιστικοί πίνακες F

denom. df_2	p	numerator df_1 1	2	3	4	5	6	7	8	9	10
1	0.10	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2
	0.05	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
	0.01	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5	6055.8
2	0.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
	0.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
	0.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
3	0.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
	0.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
4	0.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
	0.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	0.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
	0.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	0.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
	0.01	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	0.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
	0.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	0.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
	0.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	0.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
	0.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
	0.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	0.10	3.28	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
	0.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	0.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85

Ο στατιστικός δείκτης F επιτρέπει την εκτίμηση της σημαντικότητας της εξίσωσης παλινδρόμησης, δηλαδή δίνει απάντηση στο ερώτημα αν υπάρχει σημαντική σχέση ανάμεσα στις μεταβλητές X και Y.

$$F = \frac{\frac{R^2}{k-1}}{\frac{1-R^2}{n-k}}$$

Οι στατιστικοί δείκτες t

Οι στατιστικοί δείκτες t επιτρέπει την εκτίμηση της σημαντικότητας των συντελεστών a και b της εξίσωσης παλινδρόμησης, και ειδικότερα αν αυτοί είναι σημαντικά διάφοροι υποθετικών τιμών.

Τυπική απόκλιση σφαλμάτων:

$$\hat{\sigma}_e = \sqrt{\frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-k}}$$

Τυπικό σφάλμα συντελεστών:

$$\mathrm{SE}_\alpha = \hat{\sigma}_e \cdot \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

$$\mathrm{SE}_b = \hat{\sigma}_e \cdot \sqrt{\frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

$$t_a = \frac{a - a'}{SE(a)}$$

$$t_b = \frac{b - b'}{SE(b)}$$

Στατιστικοί πίνακες t

df	Tail probability p 0.1	0.05	0.025	0.01	0.005
1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.34	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79
30	1.31	1.70	2.04	2.46	2.75
40	1.30	1.68	2.02	2.42	2.70
50	1.30	1.68	2.01	2.40	2.68
60	1.30	1.67	2.00	2.39	2.66
70	1.29	1.67	1.99	2.38	2.65
80	1.29	1.66	1.99	2.37	2.64
90	1.29	1.66	1.99	2.37	2.63
100	1.29	1.66	1.98	2.36	2.63
∞	1.28	1.64	1.96	2.33	2.58
	80%	90%	95%	98%	99%
	Confidence level				

Οι στατιστικοί δείκτες t επιτρέπει την εκτίμηση της σημαντικότητας των συντελεστών a και b της εξίσωσης παλινδρόμησης, και ειδικότερα αν αυτοί είναι σημαντικά διάφοροι υποθετικών τιμών.

$$t_a = \frac{a - a'}{SE(a)} \qquad t_b = \frac{b - b'}{SE(b)}$$

Πρόβλεψη

Έχοντας υπολογίσει τους συντελεστές της εξίσωσης παλινδρόμησης, μπορούμε, για κάθε νέα τιμή της μεταβλητής Χ, να καθορίσουμε μια συγκεκριμένη τιμή για τη μεταβλητή Υ και το διάστημα εμπιστοσύνης μέσα στο οποίο αυτή θα κυμαίνεται

Προβλεπόμενη τιμή: $\hat{Y}_0 = a + bX_0$

Τυπικό Σφάλμα για την προβλεπόμενη τιμή:

$$SE(\hat{Y}_0) = \hat{\sigma}_e \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum(X_i - \bar{X})^2}} \qquad \hat{\sigma}_e = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n-k}}$$

Τελική πρόβλεψη: $Y_0 = \hat{Y}_0 \pm t \cdot SE(\hat{Y}_0)$

FSU

Παράδειγμα

	X	Y	Numerator			Denominator				LRL
	Period	Sales	X-Mean(X)=A	Y-Mean(Y)=B	A*B	(X-Mean(X))^2	(Y-Mean(Y))^2	(Yf-Mean(Y))^2	(Y-Yf)^2	Forecast
	1	30	-4,5	-12	54	20,25	144	233,478	10,758	26,72
	2	20	-3,5	-22	77	12,25	484	141,372	102,212	30,11
	3	45	-2,5	3	-7,5	6,25	9	72,250	132,250	33,5
	4	35	-1,5	-7	10,5	2,25	49	26,112	3,572	36,89
	5	30	-0,5	-12	6	0,25	144	2,958	105,678	40,28
	6	60	0,5	18	9	0,25	324	2,789	266,669	43,67
	7	40	1,5	-2	-3	2,25	4	25,604	49,844	47,06
	8	50	2,5	8	20	6,25	64	71,403	0,203	50,45
	9	45	3,5	3	10,5	12,25	9	140,186	78,146	53,84
	10	65	4,5	23	103,5	20,25	529	231,953	60,373	57,23
	11									60,62
	12									64,01
	13									67,4
	1	2			3	4	5	6	7	
Average	5,5	42		Sum	280	82,5	1760	948,105	809,705	41,975

Παράδειγμα

Επιχειρηματικές Προβλέψεις: Μέθοδοι & Τεχνικές

	X	**Y**	**Numerator**			**Denominator**				**LRL**
	Period	**Sales**	**X-Mean(X)=A**	**Y-Mean(Y)=B**	**A*B**	**(X-Mean(X))^2**	**(Y-Mean(Y))^2**	**(Yf-Mean(Y))^2**	**(Y-Yf)^2**	**Forecast**
	1	30	-4,5	-12	54	20,25	144	233,478	10,758	26,72
	2	20	-3,5	-22	77	12,25	484	141,372	102,212	30,11
	3	45	-2,5	3	-7,5	6,25	9	72,250	132,250	33,5
	4	35	-1,5	-7	10,5	2,25	49	26,112	3,572	36,89
	5	30	-0,5	-12	6	0,25	144	2,958	105,678	40,28
	6	60	0,5	18	9	0,25	324	2,789	266,669	43,67
	7	40	1,5	-2	-3	2,25	4	25,604	49,844	47,06
	8	50	2,5	8	20	6,25	64	71,403	0,203	50,45
	9	45	3,5	3	10,5	12,25	9	140,186	78,146	53,84
	10	65	4,5	23	103,5	20,25	529	231,953	60,373	57,23
	11									60,62
	12									64,01
	13									67,4
	1	**2**			**3**	**4**	**5**	**6**	**7**	
Average	5,5	42		Sum	280	82,5	1760	948,105	809,705	41,975

$$a = \bar{Y} - b\bar{X} = 23{,}33$$

$$b = \frac{\frac{\sum X_i Y_i}{n} - \bar{X}\bar{Y}}{\frac{\sum X_i^{2}}{n} - \bar{X}^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = 3{,}39$$

Παράδειγμα

Επιχειρηματικές Προβλέψεις: Μέθοδοι & Τεχνικές

	X	Y	Numerator			Denominator				LRL
	Period	Sales	X-Mean(X)=A	Y-Mean(Y)=B	A*B	(X-Mean(X))^2	(Y-Mean(Y))^2	(Yf-Mean(Y))^2	(Y-Yf)^2	Forecast
	1	30	-4,5	-12	54	20,25	144	233,478	10,758	26,72
	2	20	-3,5	-22	77	12,25	484	141,372	102,212	30,11
	3	45	-2,5	3	-7,5	6,25	9	72,250	132,250	33,5
	4	35	-1,5	-7	10,5	2,25	49	26,112	3,572	36,89
	5	30	-0,5	-12	6	0,25	144	2,958	105,678	40,28
	6	60	0,5	18	9	0,25	324	2,789	266,669	43,67
	7	40	1,5	-2	-3	2,25	4	25,604	49,844	47,06
	8	50	2,5	8	20	6,25	64	71,403	0,203	50,45
	9	45	3,5	3	10,5	12,25	9	140,186	78,146	53,84
	10	65	4,5	23	103,5	20,25	529	231,953	60,373	57,23
	11									60,62
	12									64,01
	13									67,4
	1	2			3	4	5	6	7	
Average	5,5	42		Sum	280	82,5	1760	948,105	809,705	41,975

$$Cov_{XX} = \frac{\sum(X_i - \bar{X})^2}{n} = Var_X = S_X^2 = 82,5$$

$$r_{XY} = \frac{Cov_{XY}}{\sqrt{Cov_{YY} \cdot Cov_{XX}}} = \frac{Cov_{XY}}{S_Y S_X} = 0,735$$

$$Cov_{YY} = \frac{\sum(Y_i - \bar{Y})^2}{n} = Var_Y = S_Y^2 = 176$$

$$Cov_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n} = 28$$

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = r_{XY}^2 = 0,539$$

Παράδειγμα

Επιχειρηματικές Προβλέψεις: Μέθοδοι & Τεχνικές

	X	Y	Numerator			Denominator				LRL
	Period	Sales	X-Mean(X)=A	Y-Mean(Y)=B	A*B	(X-Mean(X))^2	(Y-Mean(Y))^2	(Yf-Mean(Y))^2	(Y-Yf)^2	Forecast
	1	30	-4,5	-12	54	20,25	144	233,478	10,758	26,72
	2	20	-3,5	-22	77	12,25	484	141,372	102,212	30,11
	3	45	-2,5	3	-7,5	6,25	9	72,250	132,250	33,5
	4	35	-1,5	-7	10,5	2,25	49	26,112	3,572	36,89
	5	30	-0,5	-12	6	0,25	144	2,958	105,678	40,28
	6	60	0,5	18	9	0,25	324	2,789	266,669	43,67
	7	40	1,5	-2	-3	2,25	4	25,604	49,844	47,06
	8	50	2,5	8	20	6,25	64	71,403	0,203	50,45
	9	45	3,5	3	10,5	12,25	9	140,186	78,146	53,84
	10	65	4,5	23	103,5	20,25	529	231,953	60,373	57,23
	11									60,62
	12									64,01
	13									67,4
	1	2			3	4	5	6	7	
Average	5,5	42		Sum	280	82,5	1760	948,105	809,705	41,975

$$F = \frac{\dfrac{\sum(\hat{Y}_i - \bar{Y})^2}{k-1}}{\dfrac{\sum(Y_i - \hat{Y}_i)^2}{n-k}} = 9{,}367$$

$$\hat{\sigma}_e = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n-k}} = 10{,}060$$

$$SE(a) = \hat{\sigma}_e \sqrt{\frac{\sum X_i^2}{n\sum(X_i - \bar{X})^2}} = 6{,}873$$

$$SE(b) = \hat{\sigma}_e \sqrt{\frac{1}{\sum(X_i - \bar{X})^2}} = 1{,}108$$

$$t_a = \frac{a - a'}{SE(a)} = 3{,}395$$

$$t_b = \frac{b - b'}{SE(b)} = 3{,}064$$

FSU

Παράδειγμα

	X	Y	Numerator			Denominator				LRL
	Period	Sales	X-Mean(X)=A	Y-Mean(Y)=B	A*B	(X-Mean(X))^2	(Y-Mean(Y))^2	(Yf-Mean(Y))^2	(Y-Yf)^2	Forecast
	1	30	-4,5	-12	54	20,25	144	233,478	10,758	26,72
	2	20	-3,5	-22	77	12,25	484	141,372	102,212	30,11
	3	45	-2,5	3	-7,5	6,25	9	72,250	132,250	33,5
	4	35	-1,5	-7	10,5	2,25	49	26,112	3,572	36,89
	5	30	-0,5	-12	6	0,25	144	2,958	105,678	40,28
	6	60	0,5	18	9	0,25	324	2,789	266,669	43,67
	7	40	1,5	-2	-3	2,25	4	25,604	49,844	47,06
	8	50	2,5	8	20	6,25	64	71,403	0,203	50,45
	9	45	3,5	3	10,5	12,25	9	140,186	78,146	53,84
	10	65	4,5	23	103,5	20,25	529	231,953	60,373	57,23
	11									60,62
	12									64,01
	13									67,4
	1	2			3	4	5	6	7	
Average	5,5	42		Sum	280	82,5	1760	948,105	809,705	41,975

$$SE(\hat{Y}_0) = \hat{\sigma}_e \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}} \xRightarrow{X_0 = 11} SE(\hat{Y}_0) = 12{,}287$$

$$SE(\hat{Y}_0) = \hat{\sigma}_e \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}} \xRightarrow{X_0 = 12} SE(\hat{Y}_0) = 12{,}872$$

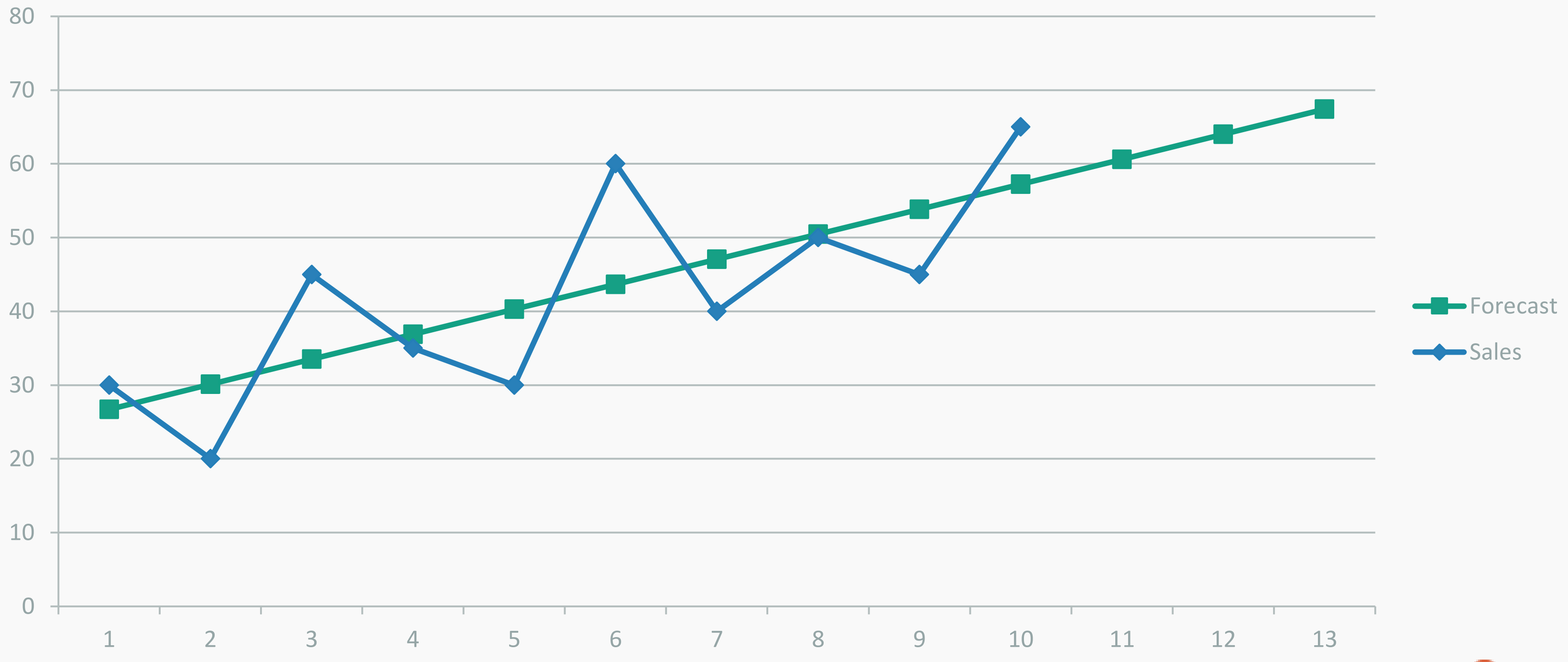
$$SE(\hat{Y}_0) = \hat{\sigma}_e \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}} \xRightarrow{X_0 = 13} SE(\hat{Y}_0) = 13{,}523$$

$$\hat{Y}_0 = a + bX_0$$

$$Y_0 = \hat{Y}_0 \pm t \cdot SE(\hat{Y}_0)$$

Παράδειγμα

Επιχειρηματικές Προβλέψεις: Μέθοδοι & Τεχνικές



Πολλαπλή Παλινδρόμηση

Επιχειρηματικές Προβλέψεις: Μέθοδοι & Τεχνικές

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \cdots + b_k \cdot X_k + e$$

Πολλαπλή Παλινδρόμηση, Σε περιπτώσεις που απαιτούνται περισσότερες από μία ανεξάρτητες μεταβλητές, το μοντέλο της απλής παλινδρόμησης μπορεί να γενικευθεί μέσω της τεχνικής της πολλαπλής παλινδρόμησης ώστε να συμπεριλάβει όλες τις μεταβλητές που επηρεάζουν την τιμή της μεταβλητής πρόβλεψης. Στην πολλαπλή παλινδρόμηση υπάρχει μια εξαρτημένη μεταβλητή της οποίας η τιμή πρέπει να προβλεφθεί βάσει των τιμών δύο ή περισσοτέρων ανεξάρτητων μεταβλητών. Η γενική μορφή της πολλαπλής παλινδρόμησης είναι:

Υπολογισμός Συντελεστών

01

$$Y_i = b_0 + b_1 \cdot X_{1i} + b_2 \cdot X_{2i} + e_i = \hat{Y}_i + e_i$$

02

$$e_i = Y_i - \hat{Y}_i$$

03

$$(b_0, b_1, b_2) \mid \min\left[\sum_{i=1}^{n} e_i^{\,2}\right]$$

04

$$\sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

$$= \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_{1,i} - b_2 X_{2,i}\right)^2$$

Προκειμένου να βρούμε τους άγνωστους συντελεστές b_0, b_1 και b_2 οι οποίοι ελαχιστοποιούν την παραπάνω ποσότητα, αρκεί να υπολογίσουμε τις μερικές παραγώγους αυτής για κάθε έναν από τους συντελεστές, να θέσουμε τις υπολογισμένες παραγώγους ίσες με το μηδέν και να λύσουμε ένα γραμμικό σύστημα τριών εξισώσεων με τρεις αγνώστους. Η παραπάνω διαδικασία μπορεί να γενικευθεί σε οποιοδήποτε μοντέλο πολλαπλής παλινδρόμησης, με περισσότερες από δύο ανεξάρτητες μεταβλητές.

Συντελεστής R^2

Για τον υπολογισμό του συντελεστή R^2 χρησιμοποιείται η ίδια εξίσωση που χρησιμοποιήθηκε και στην περίπτωση της απλής παλινδρόμησης:

$$R^2 = \frac{\text{ερμηνευθείσα διακύμανση των τιμών Y}}{\text{συνολική διακύμανση των τιμών Y}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Όμως, στην προηγούμενη εξίσωση δε λαμβάνονται υπόψη ο αριθμός των ανεξάρτητων μεταβλητών και ο αριθμός του συνόλου των παρατηρήσεων. Για να ξεπεραστεί το πρόβλημα αυτό, υπολογίζεται ένας «διορθωμένος» συντελεστής R^2 από την εξίσωση:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Ο συντελεστής εκφράζει το ποσοστό της διασποράς της μεταβλητής Y που αιτιολογείται από τις ανεξάρτητες μεταβλητές $X_1, X_2, \ldots, X_k$. Η διαφορά $(n-1)$ εκφράζει τους συνολικούς βαθμούς ελευθερίας της συνολικής διακύμανσης του μοντέλου, ενώ η παράσταση $(n-k-1)$ εκφράζει τους βαθμούς ελευθερίας της ερμηνευθείσας διακύμανσης.

F test

Ο στατιστικός δείκτης F, ο οποίος αποτελεί ένα μέτρο της σημαντικότητας του μοντέλου παλινδρόμησης, υπολογίζεται από αντίστοιχες εξισώσεις όπως στην απλή παλινδρόμηση:

$$F = \frac{\frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{k}}{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-k-1}} \qquad F = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}}$$

Αξίζει να σημειωθεί ότι η τιμή του δείκτη *F* εξαρτάται από τα μεγέθη του αριθμητή και του παρονομαστή. Αν η μη ερμηνευθείσα διακύμανση (διακύμανση των σφαλμάτων) είναι μεγάλη, τότε ο παρονομαστής είναι μεγάλος και ο δείκτης *F* γίνεται μικρότερος, γεγονός που σημαίνει ότι το μοντέλο παλινδρόμησης δεν είναι επιτυχημένο. Αντίθετα, αν η ερμηνευθείσα διακύμανση (αριθμητής) είναι σχετικά μεγαλύτερη, τότε και ο δείκτης *F* είναι μεγαλύτερος. Όπως έχει ήδη αναφερθεί στην περίπτωση της απλής παλινδρόμησης, υπάρχει στενή σχέση ανάμεσα στο συντελεστή R^2 και στο στατιστικό δείκτη *F*.

t-test

- Αφού εξεταστεί η συνολική σημαντικότητα του μοντέλου παλινδρόμησης, είναι μερικές φορές χρήσιμο να εξεταστεί η σημαντικότητα καθενός από τους συντελεστές παλινδρόμησης.
- Στην περίπτωση της πολλαπλής παλινδρόμησης, ο στατιστικός δείκτης t για ένα συγκεκριμένο συντελεστή αποτελεί εκτίμηση της σημαντικότητας του συντελεστή αυτού με την παρουσία όλων των άλλων ανεξάρτητων μεταβλητών.
- Για κάθε συντελεστή παλινδρόμησης bj μπορεί να οριστεί ένα τυπικό σφάλμα (ένα μέτρο της σταθερότητας του συντελεστή) και, με βάση την υπόθεση της κανονικότητας του μοντέλου παλινδρόμησης, ο δείκτης t, ο οποίος δίνεται από την ακόλουθη εξίσωση, ακολουθεί την t-κατανομή με (n-k-1) βαθμούς ελευθερίας.

$$t_{b_j} = \frac{b_j}{SE_{b_j}}$$

t-test

•Χρησιμοποιώντας την εξίσωση του δείκτη *t* για κάθε συντελεστή του μοντέλου παλινδρόμησης, υπολογίζεται η σημαντικότητά του, μέσα από τη σύγκριση της τιμής του συντελεστή αυτού με την τιμή 0, τιμή για την οποία η αντίστοιχη ανεξάρτητη μεταβλητή δε συνεισφέρει στην πρόβλεψη του *Y*, με δεδομένη την παρουσία των άλλων ανεξάρτητων μεταβλητών.

$$t_{b_j} = \frac{b_j}{SE_{b_j}}$$

• Ένα σημαντικό θέμα της πολλαπλής παλινδρόμησης είναι η σταθερότητα των συντελεστών παλινδρόμησης εξαρτάται από τη συσχέτιση των ανεξάρτητων μεταβλητών. Για δύο ανεξάρτητες μεταβλητές X_1 και X_2, όσο μεγαλύτερη είναι η μεταξύ τους συσχέτιση τόσο πιο ασταθείς θα είναι οι δύο συντελεστές (b_1 και b_2) που θα υπολογιστούν για τις μεταβλητές αυτές.

Residual Errors

- Η μελέτη των υπολοίπων σφαλμάτων (residual errors, δηλαδή σφάλματα προσαρμογής του μοντέλου στα πραγματικά δεδομένα) είναι πολύ σημαντική για να αποφασισθεί η καταλληλότητα ενός μοντέλου πρόβλεψης. Αν τα σφάλματα είναι επαρκώς τυχαία, τότε το μοντέλο μπορεί να θεωρηθεί ικανοποιητικό. Αν τα σφάλματα ακολουθούν οποιοδήποτε πρότυπο, τότε το μοντέλο δεν εκμεταλλεύεται όλη τη συστηματική πληροφορία που εμπεριέχεται στα δεδομένα. Μερικές από τις πιο πιθανές αναλύσεις των σφαλμάτων είναι οι ακόλουθες:

- διαγραμματική αναπαράσταση των σφαλμάτων για οπτική επισκόπηση και εύρεση της κατανομής που ακολουθούν
- μελέτη της αυτοσυσχέτισης των υπολοίπων σφαλμάτων
- υπολογισμός του στατιστικού δείκτη Durbin-Watson

Durbin-Watson

Ο στατιστικός δείκτης *DW* δίνεται από την εξίσωση:

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

Σε κάθε συνδυασμό αριθμού παρατηρήσεων, αριθμού συντελεστών παλινδρόμησης και επιπέδου εμπιστοσύνης, αντιστοιχεί ένα ζευγάρι αριθμητικών τιμών DW_L και DW_U. Ανάλογα με την υπολογισμένη τιμή του στατιστικού δείκτη, τα σφάλματα του εκάστοτε μοντέλου παλινδρόμησης χαρακτηρίζονται ως:

- Σημαντικά θετικά συσχετισμένα, αν $DW \le DW_L$
- Ασυσχέτιστα, αν $DW_U \le DW \le 4 - DW_U$
- Σημαντικά αρνητικά συσχετισμένα, αν $DW \ge 4 - DW_L$

Αν $DW_L \le DW \le DW_U$ ή $4 - DW_U \le DW \le 4 - DW_L$ τότε δεν μπορεί να εξαχθεί ασφαλές συμπέρασμα από το στατιστικό δείκτη *Durbin-Watson* σχετικά με την τυχαιότητα των σφαλμάτων.

Measure of predictive accuracy

Επιχειρηματικές Προβλέψεις: Μέθοδοι & Τεχνικές

- Adjusted R^2
- Cross-validation
- Akaike's Information Criterion
- Corrected Akaike's Information Criterion
- Schwarz Bayesian Information Criterion
- # To obtain all these measures in R, use CV(fit)

Adjusted R^2

$$\bar{R}^2 = 1-(1-\bar{R}^2)\frac{N-1}{N-k-1}$$

N: πλήθος παρατηρήσεων
K: είναι το πλήθος των predictors

Μεγιστοποίηση του $\bar{R}^2$ ταυτίζεται με ελαχιστοποίηση της διακύμανση των σφαλμάτων πρόβλεψης.

Cross - Validation

- Αποκρύπτεις τη παρατήρηση i από το data set, και εφάρμοσε το μοντέλο στα υπόλοιπα δεδομένα. Έπειτα υπολόγισε το σφάλμα $e_i^* = y_i - \hat{y}_i$ για την παρατήρηση που απέκρυψες.
- Επανάλαβε το πρώτο βήμα για i= 1, … , N
- Υπολόγισε το MSE από $e_1^*, \dots, e_N^*$

Akaike's Information Criterion

N: πλήθος παρατηρήσεων
k: πλήθος predictors

$$AIC = Nlog\left(\frac{SSE}{N}\right) + 2(k+2)$$

$$SSE = \sum_{i=1}^{N} e_i^2$$

Corrected Akaike's Information Criterion

$$AIC_c = AIC + \frac{2(k+2)(k+3)}{N-k-3}$$

Schwarz Bayesian Information Criterion

$$BIC = Nlog\left(\frac{SSE}{N}\right) + 2(k+2)\log(N)$$

Βασικές υποθέσεις

Μακρυδάκη, Wheelright και Hyndman (1998)

- Η πρώτη υπόθεση αφορά την ύπαρξη γραμμικής σχέσης ανάμεσα στην εξαρτημένη και τις ανεξάρτητες μεταβλητές. Στις περιπτώσεις που δεν ικανοποιείται η υπόθεση αυτή, μετασχηματίζονται οι ανεξάρτητες μεταβλητές σε νέες μεταβλητές που εμφανίζουν γραμμική σχέση με την εξαρτημένη μεταβλητή *Y*.

- Η δεύτερη υπόθεση αφορά τη σταθερή διακύμανση των σφαλμάτων παλινδρόμησης, η οποία αναφέρεται συχνά με τον τεχνικό όρο ομοσκεδαστικότητα (homoscedesticity). Ο αντίστοιχος όρος για την έλλειψη σταθερής διακύμανσης είναι ετεροσκεδαστικότητα. Με άλλα λόγια, η υπόθεση αυτή δηλώνει ότι τα σφάλματα πρόβλεψης θα πρέπει να είναι σταθερά για όλο το εύρος των παρατηρήσεων.

Βασικές υποθέσεις

Μακρυδάκη, Wheelright και Hyndman (1998)

- Η τρίτη υπόθεση είναι ότι τα υπόλοιπα σφάλματα είναι ανεξάρτητα το ένα από το άλλο. Αυτό σημαίνει ότι η τιμή του κάθε υπολοίπου είναι ανεξάρτητη από τις τιμές των προηγούμενων και των επόμενων.
- Όταν η υπόθεση αυτή δεν ικανοποιείται, υπάρχει σειριακή συσχέτιση (ή αυτοσυσχέτιση) ανάμεσα σε διαδοχικές τιμές των υπολοίπων σφαλμάτων. Εναλλακτικοί τρόποι αναγνώρισης της ανεξαρτησίας των υπολοίπων είναι η γραφική αναπαράσταση των τιμών τους, η εξέταση του προσήμου τους ή ο υπολογισμός του στατιστικού δείκτη Durbin-Watson.
- Όταν τα υπόλοιπα δεν είναι ανεξάρτητα, μπορεί να έχει παραλειφθεί κάποια σημαντική ανεξάρτητη μεταβλητή ή μπορεί να μην υπάρχει γραμμική σχέση ανάμεσα στις μεταβλητές της εξίσωσης παλινδρόμησης.
- Στην περίπτωση αυτή, η εξίσωση δεν αποδίδει πλήρως το βασικό λανθάνον πρότυπο (underlying pattern) των δεδομένων και τα υπόλοιπα σφάλματα, τα οποία δεν είναι τυχαία σφάλματα, αντιπροσωπεύουν κάποιο τμήμα του βασικού προτύπου.

Βασικές υποθέσεις

Μακρυδάκη, Wheelright και Hyndman (1998)

- Η τέταρτη υπόθεση είναι ότι, αν οι τιμές των υπολοίπων σφαλμάτων παρασταθούν γραφικά, θα πρέπει να εμφανίζουν μια σχεδόν κανονική διασπορά. Αυτή η υπόθεση δεν είναι γενικά δεσμευτική, καθώς τα υπόλοιπα αντιπροσωπεύουν την επίδραση (σχετικά ασήμαντη) ενός μεγάλου αριθμού παραγόντων στην τιμή της εξαρτημένης μεταβλητής.

- Τέλος, ένα σημαντικό θέμα στην πολλαπλή παλινδρόμηση είναι η πιθανότητα πολυσυγγραμικότητας. Η πολυσυγγραμι-κότητα δημιουργείται όταν δύο ή περισσότερες ανεξάρτητες μεταβλητές είναι ισχυρά συσχετισμένες και αποτελεί συχνό πρόβλημα σε οικονομικά και επιχειρησιακά δεδομένα, εξαιτίας του υψηλού βαθμού συσχέτισης που υπάρχει ανάμεσα στους διάφορους παράγοντες. Το γεγονός αυτό θα πρέπει να ληφθεί υπόψη κατά την επιλογή των ανεξάρτητων μεταβλητών και κατά τη συλλογή των δεδομένων. Ο στόχος είναι η χρησιμοποίηση ανεξάρτητων μεταβλητών οι οποίες δεν είναι ισχυρά συσχετισμένες (ένας εμπειρικός κανόνας είναι ότι η συσχέτιση δε θα πρέπει να υπερβαίνει την τιμή +0,7 ή να είναι μικρότερη από -0,7). Αν οι ανεξάρτητες μεταβλητές είναι ισχυρά συσχετισμένες, παρέχουν πλεονάζουσα πληροφορία, η οποία δε βελτιώνει την ερμηνευτική δύναμη της παλινδρόμησης.

Εφαρμογή στην πράξη

1. Διατύπωση του Προβλήματος
2. Επιλογή Οικονομικών & Άλλων Σχετικών Δεικτών
3. Αρχική Δοκιμαστική Εφαρμογή της Πολλαπλής Παλινδρόμησης
4. Μελέτη του Πίνακα Απλών Συσχετίσεων
5. Επιλογή της Εξίσωσης Παλινδρόμησης
6. Παρατηρώντας την Τιμή του R^2
7. Έλεγχος της Εγκυρότητας των Υποθέσεων για την Παλινδρόμηση
8. Προετοιμασία του μοντέλου για πρόβλεψη/εκτίμηση

Παλινδρόμηση και Δεδομένα Χρονοσειρών

- Scenario based forecasting
  - forecaster assumes possible scenarios for the predictor variable that are of interest

- Ex-ante versus ex-post forecasts
  - *Ex ante forecasts* are those that are made using only the information that is available in advance
  - *Ex post forecasts* are those that are made using later information on the predictors.

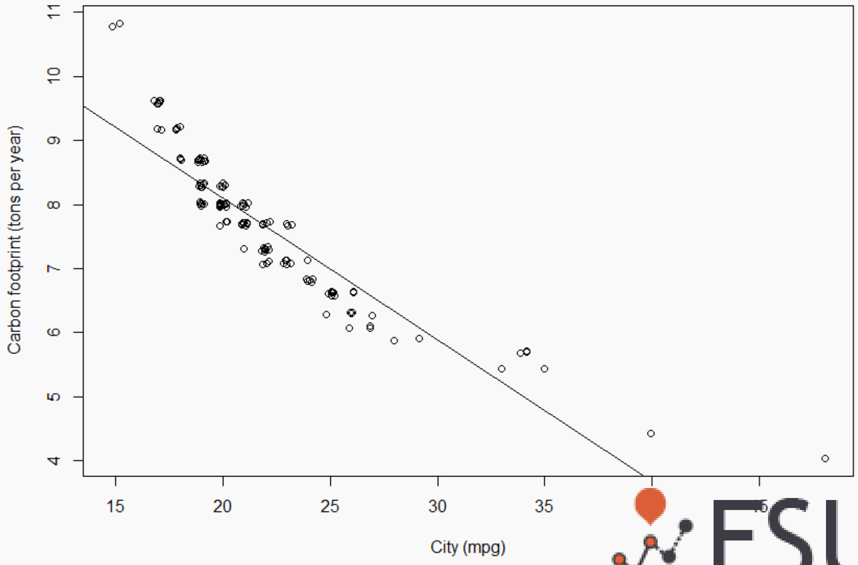
Using R


```
plot(jitter(Carbon) ~ jitter(City),xlab="City (mpg)",ylab="Carbon footprint (tons per year)",data=fuel)

fit <- lm(Carbon ~ City, data=fuel)
```

```
Coefficients:
(Intercept)        City
12.526             -0.221
```

```
abline(fit)
```



Using R

Example Linear Regression

- R Code

```
fit.ex4 <- tslm(austa ~ trend)
f <- forecast(fit.ex4, h=5,level=c(80,95))
plot(f, ylab="International tourist arrivals to Australia (millions)",
xlab="t")
lines(fitted(fit.ex4),col="blue")
summary(fit.ex4)
```

- Output

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.337535   0.100366   3.363  0.00218 **
trend       0.176075   0.005475  32.157  < 2e-16 ***
```

FSU

Using R

Example Linear Regression

Διαστήματα Εμπιστοσύνης

```
> confint(fit,level=0.95)
                 2.5 %      97.5 %
(Intercept) 12.1315464 12.9197478
City        -0.2385315 -0.2034092
```

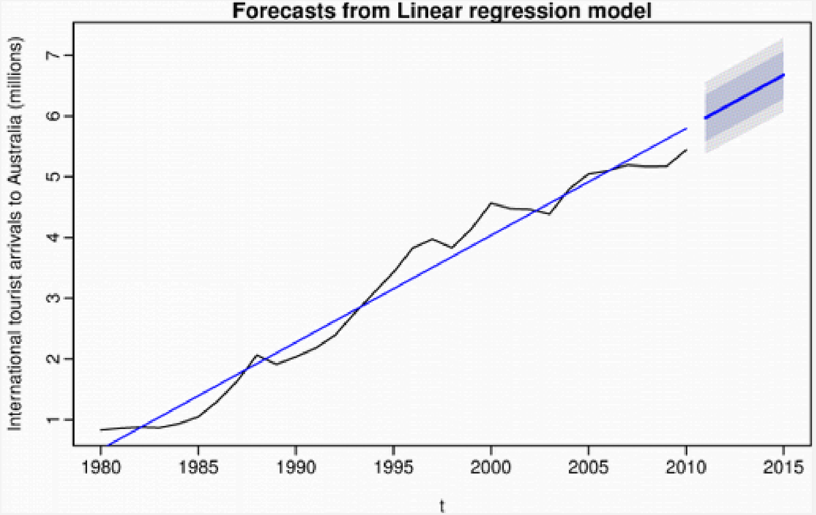
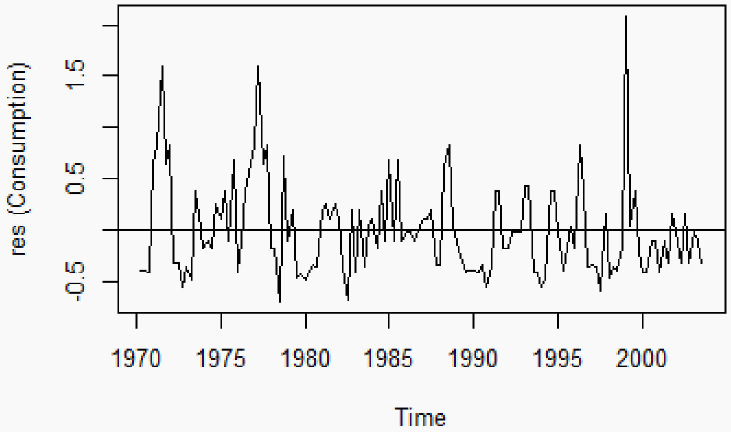


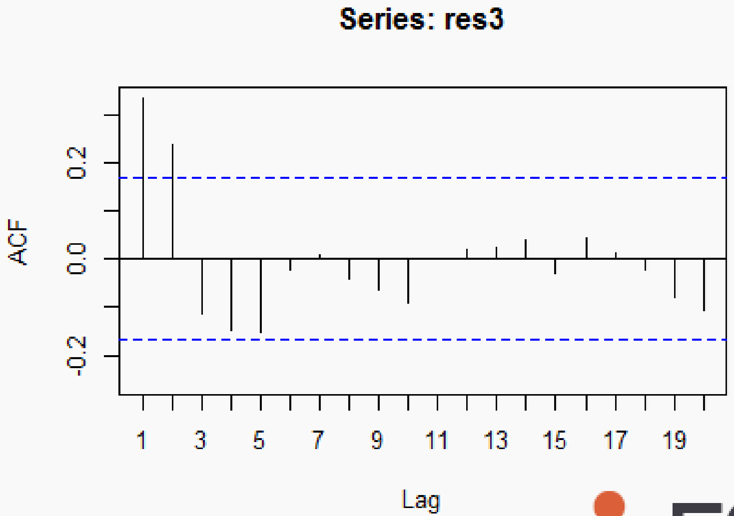
Figure 4.12: Forecasting international tourist arrivals to Australia for the period 2011-2015 using a linear trend. 80% and 95% forecast intervals are shown.

Using R

Example Linear Regression

```
par(mfrow=c(2,2))
res3 <- ts(resid(fit),s=1970.25,f=4)
plot.ts(res3,ylab="res (Consumption)")
abline(0,0)
Acf(res3)
```





FSU

Using R


```
summary(fit)
```

Call:
lm(formula = Carbon ~ City, data = fuel)
Residuals:

Min	1Q	Median	3Q	Max
-0.7014	-0.3643	-0.1062	0.1938	2.0809

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.525647	0.199232	62.87	<2e-16 ***
City	-0.220970	0.008878	-24.89	<2e-16 ***

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

Residual standard error: 0.4703 on 132 degrees of freedom
Multiple R-squared: 0.8244, Adjusted R-squared: 0.823
F-statistic: 619.5 on 1 and 132 DF, p-value: < 2.2e-16

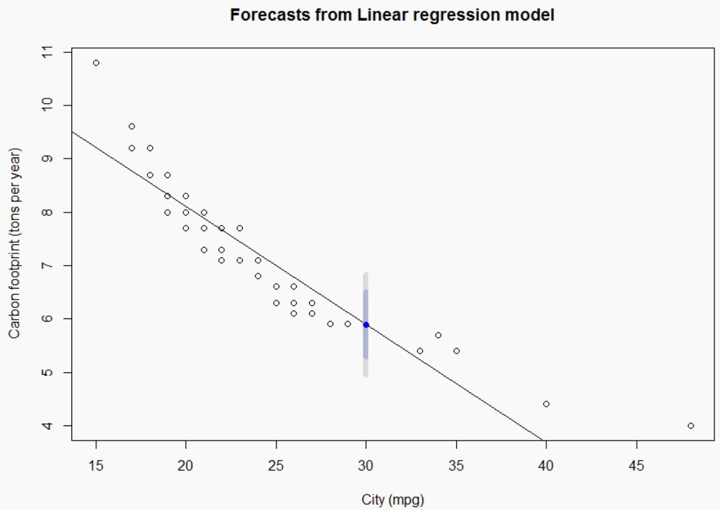
Using R

Forecasting with Linear Regression

```
fitted(fit)[1]
20 7.001388

fcast <- forecast(fit, newdata=data.frame(City=30))
   Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
1    5.896537      5.281691  6.511383   4.95226  6.840813

plot(fcast, xlab="City (mpg)",
 ylab="Carbon footprint
(tons per year)")
```



Using R

Forecasting with Linear Regression

- Scenario based forecasting
  - possible scenarios for the predictor variable that are of interest.
  - ```
    fcast <- forecast(fit.ex3, newdata=data.frame(income=c(-1,1)))
    plot(fcast, ylab="% change in consumption", xlab="% change in income")
    ```
- Ex-ante versus ex-post forecasts
  - *Ex ante forecasts* are those that are made using only the information that is available in advance.
  - *Ex post forecasts* are those that are made using later information on the predictors.
    ```

Using R

Example Linear Trend

```
fit.ex4 <- tslm(austa ~ trend)
f <- forecast(fit.ex4, h=5,level=c(80,95))
plot(f, ylab="International tourist arrivals to Australia
(millions)",
  xlab="t")
lines(fitted(fit.ex4),col="blue")
summary(fit.ex4)
```

Άσκηση

Ο παρακάτω πίνακας απεικονίζει τις winning times (in seconds) για τους άντρες που αγωνίστηκαν στα 400m στους Ολυμπιακούς Αγώνες από το 1896 έως το 2012 (data set `olympic`).

1896	**54.2**	**1928**	**47.8**	**1964**	**45.1**	**1992**	**43.50**
1900	49.4	1932	46.2	1968	43.8	1996	43.49
1904	49.2	1936	46.5	1972	44.66		
1908	50.0	1948	46.2	1976	44.27		
1912	48.2	1952	45.9	1980	44.60		
1920	49.6	1956	46.7	1984	44.27		
1924	47.6	1960	44.9	1988	43.87		

1. Ανανεώστε το data set `olympic` έτσι ώστε να περιλαμβάνει και τους αντίστοιχους χρόνους των τελευταίων Ολυμπιακών Αγώνων.
2. Σχεδιάστε το διάγραμμα και αναφέρετε τα κυριότερα ποιοτικά χαρακτηριστικά των χρονοσειρών που αναγνωρίζετε.
3. Σχεδιάστε/ Υπολογίστε την ευθεία γραμμικής παλινδρόμησης στα δεδομένα. Προφανώς οι «winning times» έχουν μειωθεί, αλλά με τι «μέσο» ρυθμό ανά χρόνο;
4. Σχεδιάστε τα residuals στο χρόνο. Τι συμπεράσματα προκύπτουν για την καταλληλόλητα του μοντέλου από το διάγραμμα;
5. Προβλέψτε τις «winning time» για τα 400m ανδρών στους τελικούς Ολυμπιακούς των 2000, 2004, 2008 and 2012.Δώστε κατάλληλα διαστήματα εμπιστοσύνης για κάθε πρόβλεψή σας. Τι υποθέσεις έχετε κάνει σε αυτούς τους υπολογισμούς;
6. Βρείτε τις πραγματικές τιμές για τις «winning times» για τους προαναφερόμενους Ολυμπιακού αγώνες. (seewww.databaseolympics.com). Πόσο καλές ήταν οι προβλέψεις σας και τα διαστήματα εμπιστοσύνης που υπολογίσατε. Αξιολογείστε.

Feel free to say hi!

We are friendly and social

Ηρώων Πολυτεχνείου 9, Ζωγράφος
Αττική, 15780, Ελλάδα
E-mail: info(at)fsu.gr
Τηλέφωνο: 2107723637 Fax: 2107723740

Κτίριο της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
2ος όροφος - 2.2.1 Εργαστήριο

@FSU NTUA

Μονάδα Προβλέψεων και Στρατηγικής ΕΜΠ

lesson@fsu.gr

FSU