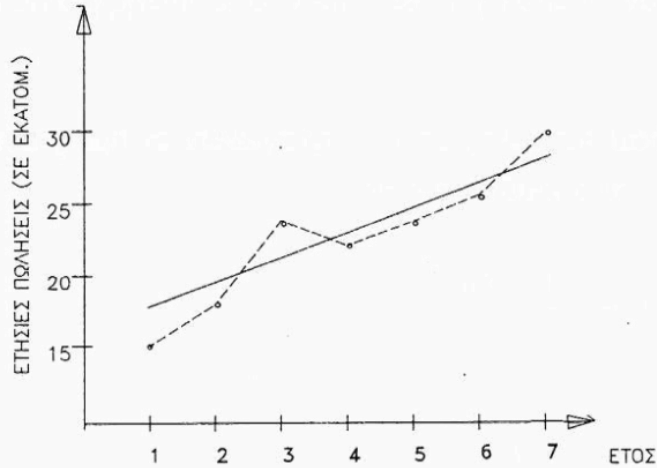


ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Μονάδα Προβλέψεων & Στρατηγικής
Forecasting & Strategy Unit

Επιχειρηματικές Προβλέψεις: Μέθοδοι & Τεχνικές
Γραμμική Παλινδρόμηση Διάλεξη 8



Απλή Γραμμική Παλινδρόμηση



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

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

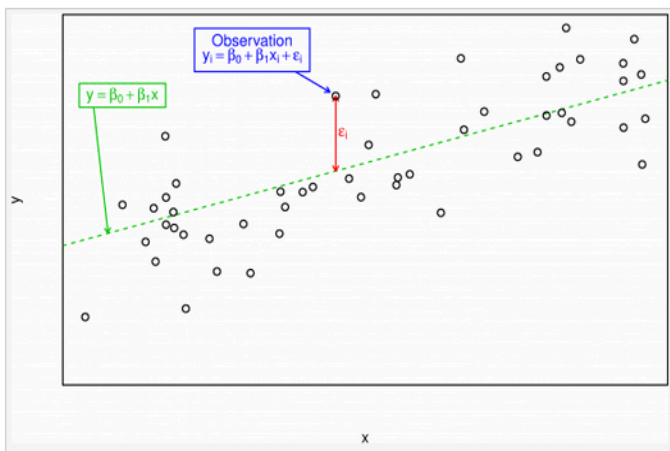
$$\hat{Y}_i = a + bX_i$$

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

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

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

Βασικές Υποθέσεις



Τα σφάλματα του μοντέλου

- 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.
- Δύο μεταβλητές ενδέχεται να συσχετίζονται χωρίς στην πραγματικότητα η τιμή της μίας να επηρεάζει ή να 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

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

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

$$r_{XY} = \frac{Cov_{XY}}{\sqrt{Cov_{YY} \cdot Cov_{XX}}} = \frac{Cov_{XY}}{S_Y S_X} \quad |r_{XY}| \leq 1$$

Συντελεστής R^2

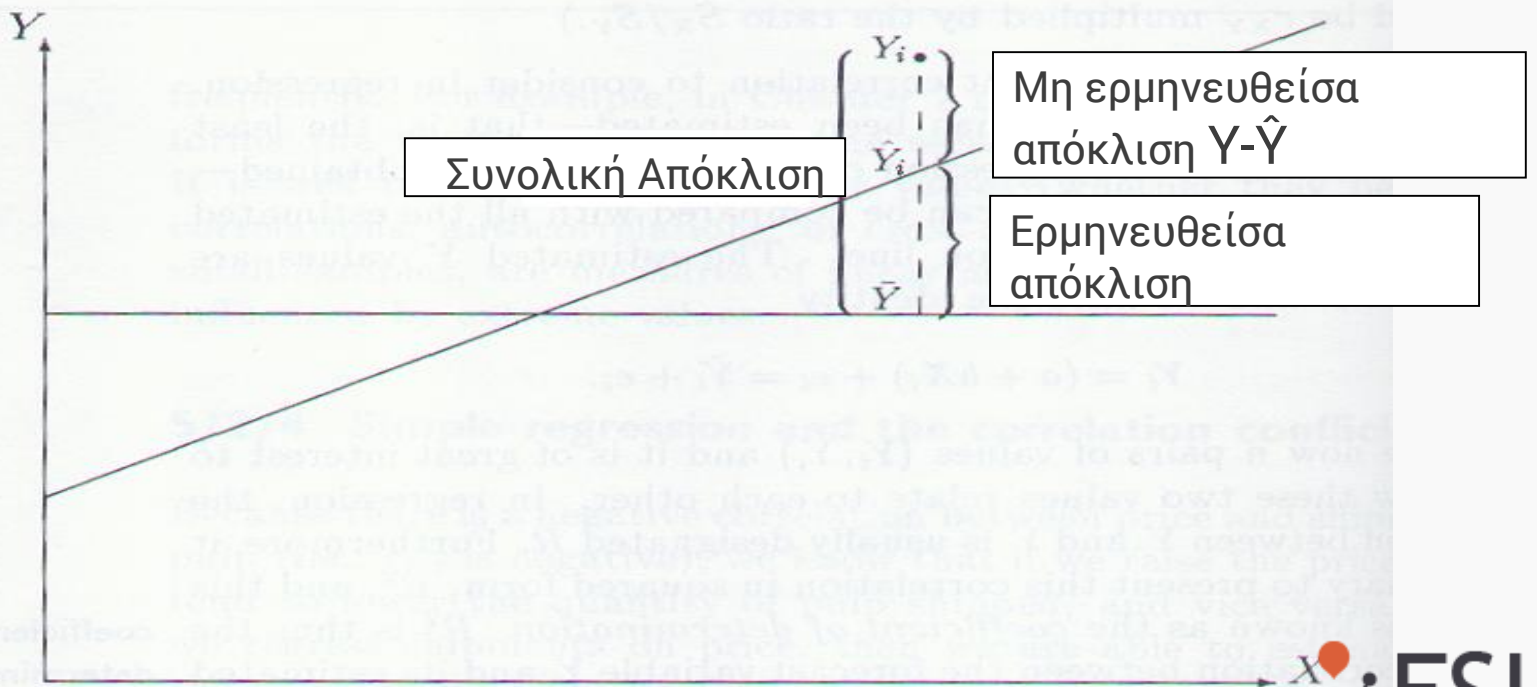
- Η συσχέτιση των τιμών που προκύπτουν από την εξίσωση της ευθείας παλινδρόμησης και των πραγματικών τιμών συμβολίζεται με R . Στην πράξη η συσχέτιση αυτή χρησιμοποιείται στην τετραγωνική της μορφή και ως εκ τούτου είναι ένας συντελεστής πάντα θετικός ($0 < R^2 < 1$).
- Αντιπροσωπεύει το ποσοστό της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

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

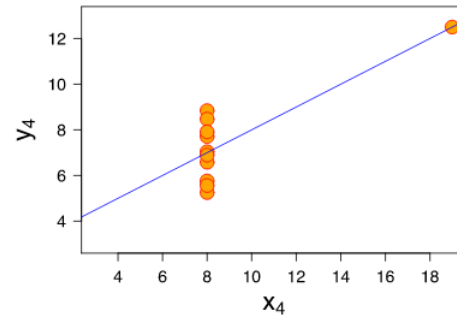
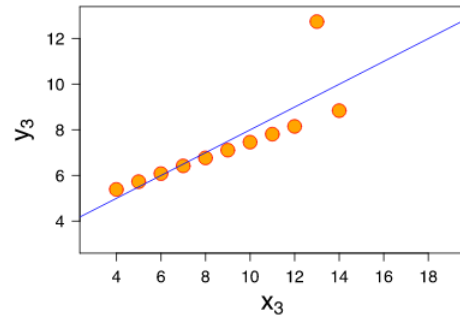
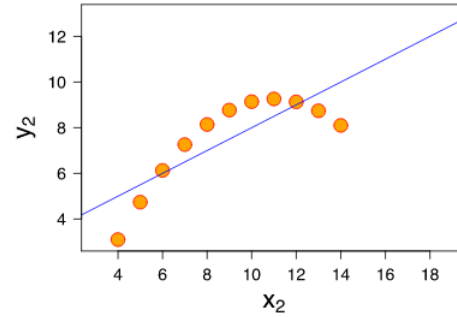
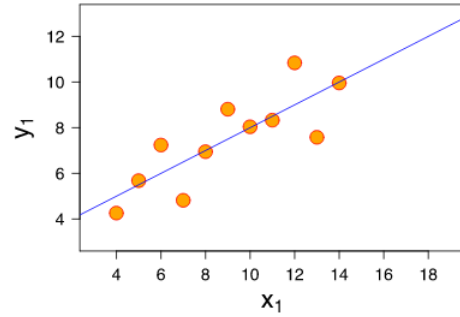
$$R^2 = \frac{\text{ερμηνευθείσα διακύμανση της } Y}{\text{συνολική διακύμανση της } Y}$$

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

Συντελεστής R^2



Συντελεστής R^2



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 a και b

Ο στατιστικός δείκτης F

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

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

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

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 επιτρέπουν την εκτίμηση της σημαντικότητας των συντελεστών a και b της εξίσωσης παλινδρόμησης, και απαντούν στο ερώτημα αν αυτοί είναι σημαντικά διάφοροι υποθετικών τιμών.

Τυπική απόκλιση σφαλμάτων:

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

Τυπικό σφάλμα
συντελεστών:

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

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

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

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

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				

Στατιστικοί πίνακες t

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

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

Πρόβλεψη

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

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

Τυπικό Σφάλμα για την προβλεπόμενη τιμή:

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

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

Παράδειγμα

$$Y=a*X+b$$

<u>X</u>	<u>Y</u>	<u>Numerator</u>			<u>Denominator</u>				<u>LRL</u>
<u>Period</u>	<u>Sales</u>	<u>X-Mean(X)=A</u>	<u>Y-Mean(Y)=B</u>	<u>A*B</u>	<u>A^2</u>				<u>Forecast</u>
1	30	-4,5	-12	54	20,25				26,73
2	20	-3,5	-22	77	12,25				30,12
3	45	-2,5	3	-7,5	6,25				33,52
4	35	-1,5	-7	10,5	2,25				36,91
5	30	-0,5	-12	6	0,25				40,30
6	60	0,5	18	9	0,25				43,70
7	40	1,5	-2	-3	2,25				47,09
8	50	2,5	8	20	6,25				50,48
9	45	3,5	3	10,5	12,25				53,88
10	65	4,5	23	103,5	20,25				57,27
11									60,67
12									64,06
13									67,45

	<u>Mean(X)</u>	<u>Mean(Y)</u>
Average	5,5	42

Sum	280	82,5				
-----	-----	------	--	--	--	--

$$b=A*B/A^2=3,39$$

$$a=Mean(Y)-b*Mean(X)=23,33$$

Παράδειγμα

X	Y	Numerator			Denominator			LRL
Period	Sales	X-Mean(X)=A	Y-Mean(Y)=B	A*B	A^2	B^2	Forecast	
1	30	-4,5	-12	54	20,25	144	26,73	
2	20	-3,5	-22	77	12,25	484	30,12	
3	45	-2,5	3	-7,5	6,25	9	33,52	
4	35	-1,5	-7	10,5	2,25	49	36,91	
5	30	-0,5	-12	6	0,25	144	40,30	
6	60	0,5	18	9	0,25	324	43,70	
7	40	1,5	-2	-3	2,25	4	47,09	
8	50	2,5	8	20	6,25	64	50,48	
9	45	3,5	3	10,5	12,25	9	53,88	
10	65	4,5	23	103,5	20,25	529	57,27	
11							60,67	
12							64,06	
13							67,45	

	Mean(X)	Mean(Y)	Sum	280	82,5	1760		
Average	5,5	42						

$$Cov_{XX} = \frac{\sum(X_i - \bar{X})^2}{n} = Var_X = S_X^2 = 8,25$$

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

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

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

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = r_{XY}^2 = 0,540$$



Παράδειγμα

X	Y	Numerator			Denominator			LRL	
Period	Sales	X-Mean(X)=A	Y-Mean(Y)=B	A*B	A^2	B^2	(Yf-Mean(Y))^2	(Y-Yf)^2	Forecast
1	30	-4,5	-12	54	20,25	144	233,26	10,71	26,73
2	20	-3,5	-22	77	12,25	484	141,11	102,44	30,12
3	45	-2,5	3	-7,5	6,25	9	71,99	131,90	33,52
4	35	-1,5	-7	10,5	2,25	49	25,92	3,64	36,91
5	30	-0,5	-12	6	0,25	144	2,88	106,15	40,30
6	60	0,5	18	9	0,25	324	2,88	265,79	43,70
7	40	1,5	-2	-3	2,25	4	25,92	50,28	47,09
8	50	2,5	8	20	6,25	64	71,99	0,24	50,48
9	45	3,5	3	10,5	12,25	9	141,11	78,83	53,88
10	65	4,5	23	103,5	20,25	529	233,26	59,71	57,27
11									60,67
12									64,06
13									67,45

	Mean(X)	Mean(Y)						
Average	5,5	42	Sum	280	82,5	1760	950,30	809,70

$$F = \frac{\frac{\sum(\hat{Y}_i - \bar{Y})^2}{k-1}}{\frac{\sum(Y_i - \hat{Y}_i)^2}{n-k}} = 9,389$$

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

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

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

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

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



Παράδειγμα

X	Y	Numerator			Denominator				LRL
Period	Sales	X-Mean(X)=A	Y-Mean(Y)=B	A*B	A^2	B^2	(Yf-Mean(Y))^2	(Y-Yf)^2	Forecast
1	30	-4,5	-12	54	20,25	144	233,26	10,71	26,73
2	20	-3,5	-22	77	12,25	484	141,11	102,44	30,12
3	45	-2,5	3	-7,5	6,25	9	71,99	131,90	33,52
4	35	-1,5	-7	10,5	2,25	49	25,92	3,64	36,91
5	30	-0,5	-12	6	0,25	144	2,88	106,15	40,30
6	60	0,5	18	9	0,25	324	2,88	265,79	43,70
7	40	1,5	-2	-3	2,25	4	25,92	50,28	47,09
8	50	2,5	8	20	6,25	64	71,99	0,24	50,48
9	45	3,5	3	10,5	12,25	9	141,11	78,83	53,88
10	65	4,5	23	103,5	20,25	529	233,26	59,71	57,27
11									60,67
12									64,06
13									67,45

	Mean(X)	Mean(Y)						
Average	5,5	42	Sum	280	82,5	1760	950,30	809,70

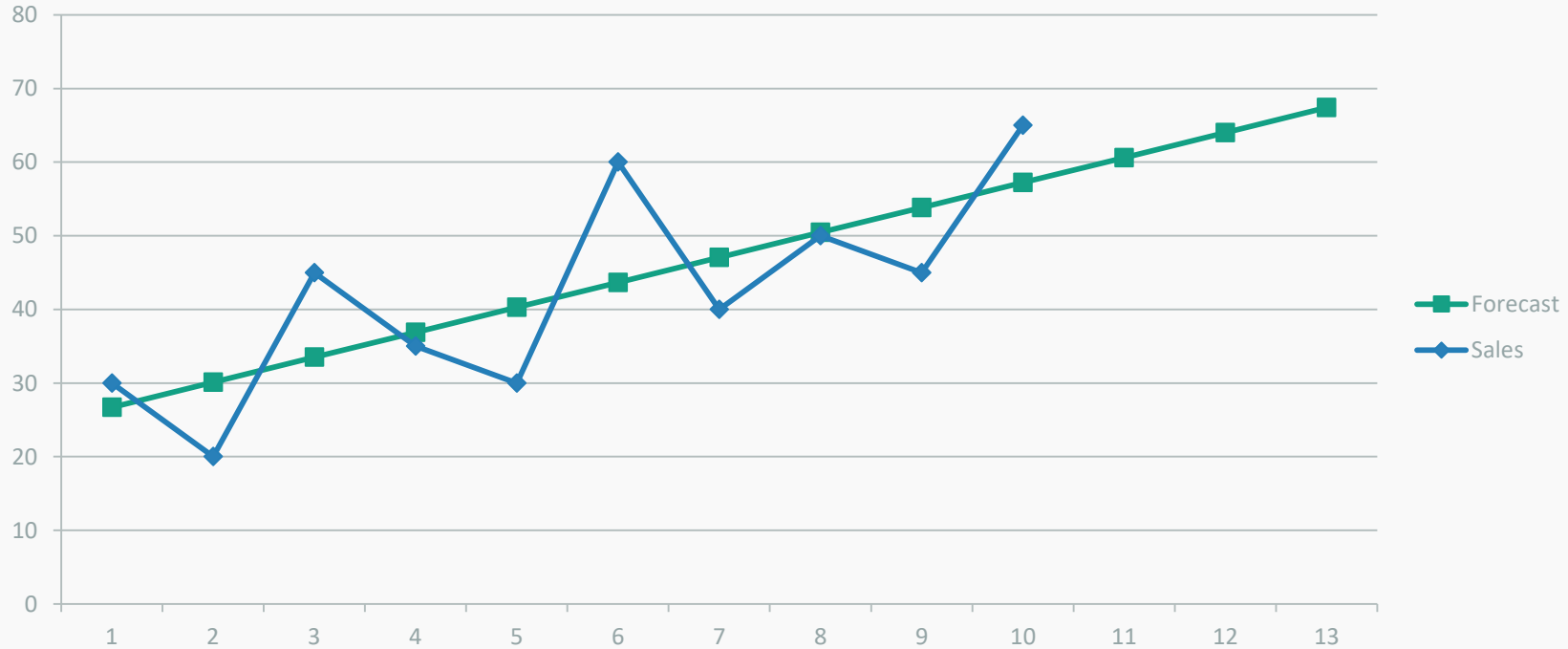
$$\hat{Y}_0 = a + bX_0$$

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

$$SE(\hat{Y}_0) = \hat{\sigma}_e \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum(X_i - \bar{X})^2}} \xrightarrow{x_0=11} SE(\hat{Y}_0) = 12,18$$



Παράδειγμα



Πολλαπλή Παλινδρόμηση

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 περισσότερων ανεξάρτητων μεταβλητών. Η γενική μορφή της πολλαπλής παλινδρόμησης είναι:

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

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

$$\begin{aligned} \sum_{i=1}^n e_i^2 &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^n (Y_i - b_0 - b_1 X_{1,i} - b_2 X_{2,i})^2 \end{aligned}$$

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, \dots, 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:

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

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 b_j 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.

• 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.

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

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 $DW \leq DW_L$
- Ασυσχέτιστα, αν $DW_U \leq DW \leq 4 - DW_U$
- Σημαντικά αρνητικά συσχετισμένα, αν $DW \geq 4 - DW_L$

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

Measure of predictive accuracy

- 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 R²

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

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 $i = 1, \dots, N$
- Υπολόγισε το MSE από e_1^*, \dots, e_N^*

Akaike's Information Criterion

N: πλήθος παρατηρήσεων

k: πλήθος predictors

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

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

Corrected Akaike's Information Criterion

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

Schwarz Bayesian Information Criterion

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

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

Using R

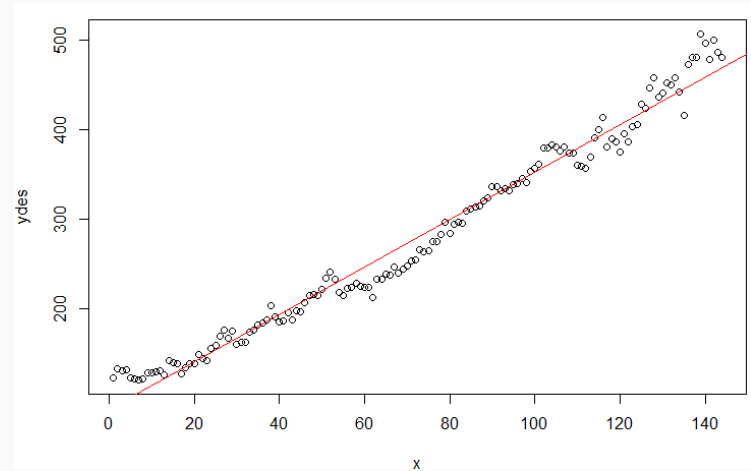
```
library(fpp)

y <- AirPassengers
#Create train and test set
insample <- window(y, start=1949, end= c(1959,12))
outsample <- window(y, start=1960)

#Deseasonalize data
seasonal_indices <- decompose(insample,
type="multiplicative")$seasonal
ydes <- insample/seasonal_indices
x <- c(1:length(insample))
trainset <- data.frame(ydes,x)

plot(x, ydes)
fit <- lm(ydes~x, data=trainset)

abline(fit, col="red")
```



Using R

Call:

```
lm(formula = ydes ~ x, data = trainset)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-38.975	-10.449	0.397	9.569	41.223

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	92.49411	2.63328	35.12	<2e-16	***
x	2.55389	0.03436	74.33	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.04 on 130 degrees of freedom

Multiple R-squared: 0.977, Adjusted R-squared: 0.9768

F-statistic: 5525 on 1 and 130 DF, p-value: < 2.2e-16

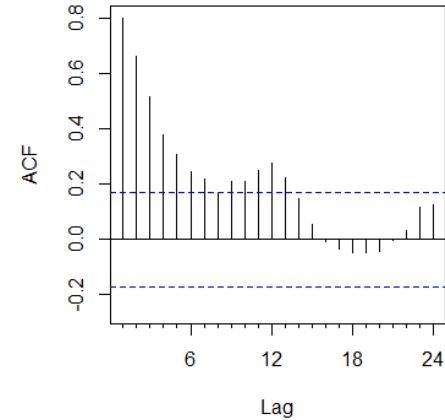
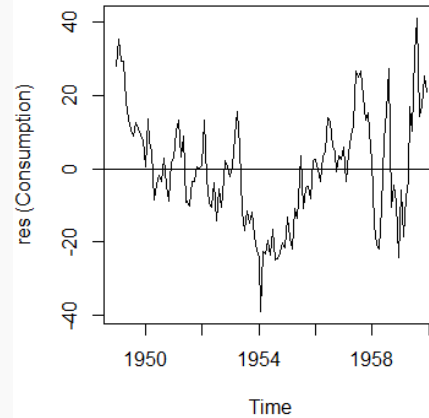
Using R

```
#Compute confidence intervals  
confint(fit,level=0.95)
```

```
                2.5 %      97.5 %  
(Intercept) 87.284480 97.703739  
x            2.485913  2.621859
```

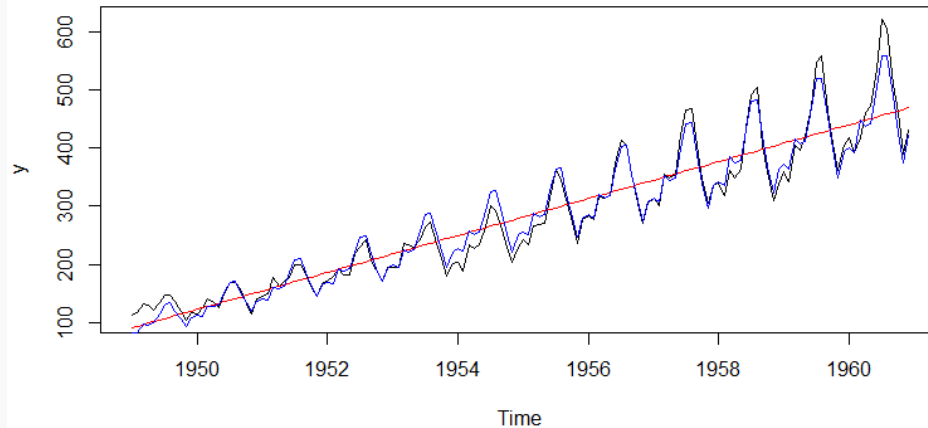
Using R

```
par(mfrow=c(1,2))
res<-ts(resid(fit),s=1949,f=12)
plot.ts(res,ylab="res (Consumption)")
abline(0,0)
Acf(res)
```



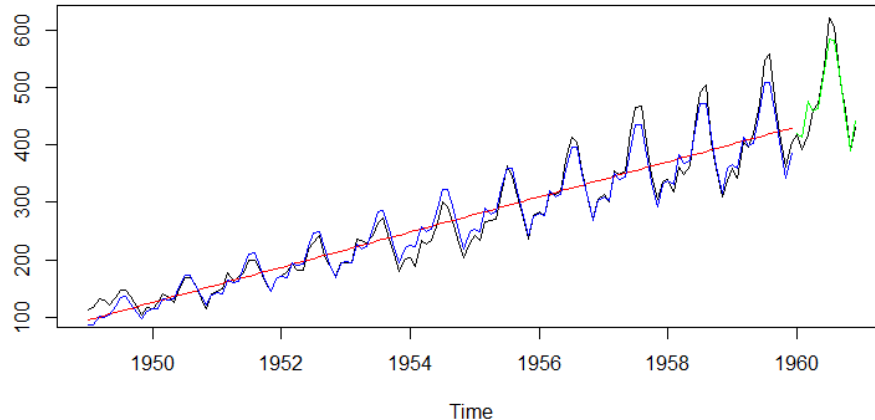
Using R

```
plot(y)
#Original in-sample forecasts
frc <- ts(predict(fit), frequency = 12, start = start(y))
lines(frc, col="red")
#Re-seasonalized in-sample forecasts
frc_s <- ts(predict(fit), frequency = 12, start = start(y))*seasonal_indices
lines(frc_s, col="blue")
```



Using R

```
#Out of sample forecasts
x <- max(x)+c(1:12)
frc_new <- ts(predict(fit, newdata = data.frame(x)), start = start(outsample),
frequency = 12)
#Re-seasonalize
si_new <- as.numeric(head(tail(seasonal_indices,24),12))
lines(frc_new*si_new, col="green")
```





Feel free to say hi!

We are friendly and social

Ηρώων Πολυτεχνείου 9, Ζωγράφος
Αττική, 15780, Ελλάδα
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



spiliotis@fsu.gr

