UNIVERSIDADE FEDERAL FLUMINENSE

Programa de Mestrado e Doutorado em Engenharia de Produção

Forecasting

Lesson: MLR

Professor: Valdecy Pereira. D. Sc.

email: valdecy.pereira@gmail.com

Forecasting

1. Definition
2. MLR
3. Residuals
4. Bibliography

Forecasting

The purpose of **MLR** (Multiple Linear Regression) to analyze the relationship between metric (or binary) independent variables (predictors) and a metric the dependent variable (response variable), with the following formulation:

$$Y = B_0 + B_1 X_1 + \cdots + B_i \, X_i$$

Where:

Y= Dependent Variable;

X_i= Independent Variable;

B_0= Intercept;

B_i= *Slopes*.

Forecasting

What is the optimal number of predictors? The suggested rules are:

- Evan rule (conservative): $\frac{n}{k} \geq 10 \rightarrow$ at least 10 observations (n) for predictor (k)

- Doane rule (relaxed): $\frac{n}{k} \geq 5 \rightarrow$ at least 5 observations (n) for predictor (k)

Categorical variables can be included as dummy variables (1 = one belongs to category; 0 = it does not belong to category). It is not necessary to encode all categories because the last one is identified when all the others have a zero value. This method prevents the occurrence of collinearity, and allows the design matrix to be invertible. Dummy variables have the same statistical treatment of the independent variables.

Forecasting

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \end{bmatrix} \quad ; \quad X(Design\ Matrix) = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1j} \\ 1 & x_{21} & \cdots & x_{2j} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i1} & \cdots & x_{ij} \end{bmatrix}$$

Forecasting

Observation	Y	X1	X2
1	9.6	2	52
2	9.95	2	50
3	10.3	1	585
4	11.66	2	360
5	14.38	2	375
6	16.86	4	200
7	17.08	4	412
8	17.89	4	400
9	21.15	5	400
10	21.65	4	205
11	22.13	6	100
12	24.35	9	100
13	24.45	8	110
14	25.02	8	295
15	27.5	8	300
16	31.75	11	120
17	34.93	10	540
18	35	10	550
19	37	11	400
20	41.95	12	500
21	44.88	15	290
22	46.59	15	250
23	54.12	16	510
24	56.63	17	590
25	69	20	600

In order to explain a **MLR** approach, the following dataset will be used: The simulated dataset of 25 observations and 2 independent Variables X_1 and X_2.

Forecasting

The B matrix can be calculated as:

$$B = (X'X)^{-1}X'Y$$

Where:

$'$ = Transposed Matrix

$^{-1}$ = Inverse Matrix

Forecasting

X'	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	2	2	1	2	2	4	4	4	5	4	6	9	8	8	8	11	10	10	11	12	15	15	16	17	20
	52	50	585	360	375	200	412	400	400	205	100	100	110	295	300	120	540	550	400	500	290	250	510	590	600

X'X	25	206	8294
	206	2396	77177
	8294	77177	3531848

(X'X)-1	0.214652617	-0.007490914	-0.000340389
	-0.007490914	0.001670763	-1.89178E-05
	-0.000340389	-1.89178E-05	1.49588E-06

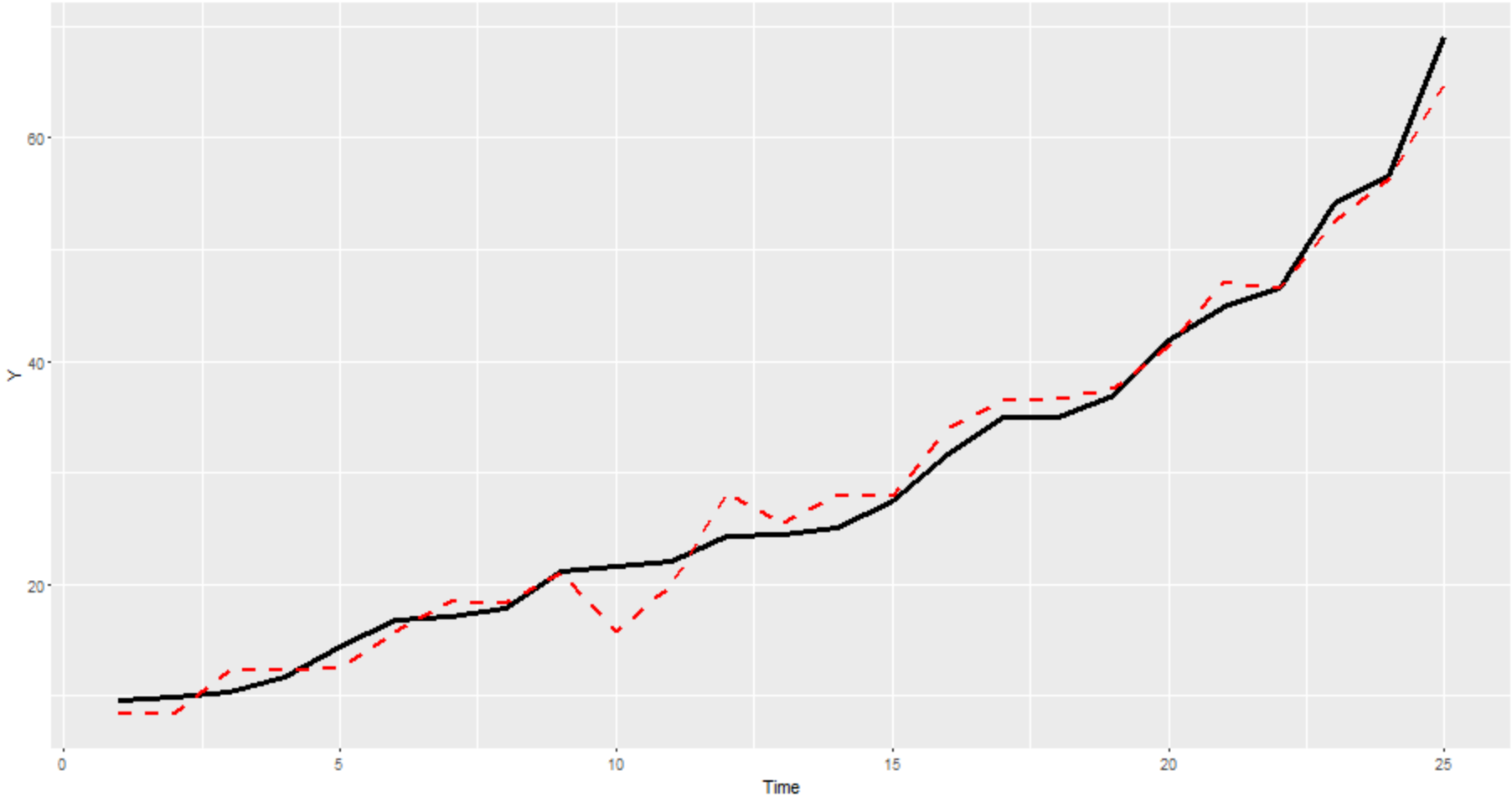
B = (X'X)-1X'Y	2.263791434
	2.744269643
	0.012527811

$$Y = B_0 + B_1 X_1 + B_2 X_2$$

$$Y = 2.263791434 + 2.744269643 X_1 + 0.012527811 X_2$$

Forecasting

Observation	Y	X1	X2	$\hat{y}$
1	9.6	2	52	8.40
2	9.95	2	50	8.38
3	10.3	1	585	12.34
4	11.66	2	360	12.26
5	14.38	2	375	12.45
6	16.86	4	200	15.75
7	17.08	4	412	18.40
8	17.89	4	400	18.25
9	21.15	5	400	21.00
10	21.65	4	205	15.81
11	22.13	6	100	19.98
12	24.35	9	100	28.21
13	24.45	8	110	25.60
14	25.02	8	295	27.91
15	27.5	8	300	27.98
16	31.75	11	120	33.95
17	34.93	10	540	36.47
18	35	10	550	36.60
19	37	11	400	37.46
20	41.95	12	500	41.46
21	44.88	15	290	47.06
22	46.59	15	250	46.56
23	54.12	16	510	52.56
24	56.63	17	590	56.31
25	69	20	600	64.67


60
40
20
Y
0
5
10
15
20
25
Time

Forecasting

The model standard error is calculated by.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - (k+1)}; \hat{\sigma}^2 = 5.235$$

$$\hat{\sigma} = 2.288$$

Where:

k = Total number of independent variables;

$\hat{y}$ = Estimated value of the dependent variable;

Forecasting

Variance-Covariance Matrix (C):

$$C = \hat{\sigma}^2 (X'X)^{-1}$$

	B0	B1	B2	SE(Bi)
B0	**1.123740429**	-0.039216122	-0.001781991	1.060066238
B1	-0.039216122	**0.008746709**	-9.90378E-05	0.093523844
B2	-0.001781991	-9.90378E-05	**7.83115E-06**	0.002798419

Forecasting

Variance-Covariance Matrix (C):

$$C = \hat{\sigma}^2 (X'X)^{-1}$$

Variance

	B0	B1	B2	SE(Bi)
B0	**1.123740429**	-0.039216122	-0.001781991	1.060066238
B1	-0.039216122	**0.008746709**	-9.90378E-05	0.093523844
B2	-0.001781991	-9.90378E-05	**7.83115E-06**	0.002798419

Forecasting

Variance-Covariance Matrix (C):

$$C = \hat{\sigma}^2 (X'X)^{-1}$$

Covariance

	B0	B1	B2	SE(Bi)
B0	**1.123740429**	-0.039216122	-0.001781991	1.060066238
B1	-0.039216122	**0.008746709**	-9.90378E-05	0.093523844
B2	-0.001781991	-9.90378E-05	**7.83115E-06**	0.002798419

Forecasting

Variance-Covariance Matrix (C):

$$C = \hat{\sigma}^2 (X'X)^{-1}$$

St. Deviation

	B0	B1	B2	SE(Bi)
B0	**1.123740429**	-0.039216122	-0.001781991	1.060066238
B1	-0.039216122	**0.008746709**	-9.90378E-05	0.093523844
B2	-0.001781991	-9.90378E-05	**7.83115E-06**	0.002798419

Forecasting

The confidence interval $(1-\alpha)$ of each B_i is calculated as:

$$B_i \pm t_{n-(k+1);\alpha/2} \times SE_{Bi}$$

For a 95% confidence interval:

T-Student Bilateral Table

GL	α			GL	α			GL	α		
	10%	5%	1%		10%	5%	1%		10%	5%	1%
2	2.920	4.303	9.925	*16*	1.74588	2.11991	2.92078	*30*	1.69726	2.04227	2.75
3	2.353	3.182	5.841	*17*	1.73961	2.10982	2.89823	*31*	1.69552	2.03951	2.74404
4	2.132	2.776	4.604	*18*	1.73406	2.10092	2.87844	*32*	1.69389	2.03693	2.73848
5	2.015	2.571	4.032	*19*	1.72913	2.09302	2.86093	*33*	1.69236	2.03452	2.73328
6	1.943	2.447	3.707	*20*	1.72472	2.08596	2.84534	*34*	1.69092	2.03224	2.72839
7	1.895	2.365	3.499	*21*	1.72074	2.07961	2.83136	*35*	1.68957	2.03011	2.72381
8	1.860	2.306	3.355	*22*	1.71714	2.07387	2.81876	*36*	1.6883	2.02809	2.71948
9	1.833	2.262	3.250	*23*	1.71387	2.06866	2.80734	*37*	1.68709	2.02619	2.71541
10	1.812	2.228	3.169	*24*	1.71088	2.0639	2.79694	*38*	1.68595	2.02439	2.71156
11	1.796	2.201	3.106	*25*	1.70814	2.05954	2.78744	*39*	1.68488	2.02269	2.70791
12	1.782	2.179	3.055	*26*	1.70562	2.05553	2.77871	*40*	1.68385	2.02108	2.70446
13	1.771	2.160	3.012	*27*	1.70329	2.05183	2.77068	*41*	1.68288	2.01954	2.70118
14	1.761	2.145	2.977	*28*	1.70113	2.04841	2.76326	*42*	1.68195	2.01808	2.69807
15	1.753	2.131	2.947	*29*	1.69913	2.04523	2.75639	*43*	1.68107	2.01669	2.6951

T-Student Bilateral Table

GL	α			GL	α			GL	α		
	10%	5%	1%		10%	5%	1%		10%	5%	1%
2	2.920	4.303	9.925	*16*	1.74588	2.11991	2.92078	*30*	1.69726	2.04227	2.75
3	2.353	3.182	5.841	*17*	1.73961	2.10982	2.89823	*31*	1.69552	2.03951	2.74404
4	2.132	2.776	4.604	*18*	1.73406	2.10092	2.87844	*32*	1.69389	2.03693	2.73848
5	2.015	2.571	4.032	*19*	1.72913	2.09302	2.86093	*33*	1.69236	2.03452	2.73328
6	1.943	2.447	3.707	*20*	1.72472	2.08596	2.84534	*34*	1.69092	2.03224	2.72839
7	1.895	2.365	3.499	*21*	1.72074	[illegible]	2.83136	*35*	1.68957	2.03011	2.72381
8	1.860	2.306	3.355	*22*	1.7171[illegible]	2.07387	[illegible].81876	*36*	1.6883	2.02809	2.71948
9	1.833	2.262	3.250	*23*	1.71387	[illegible]	2.80734	*37*	1.68709	2.02619	2.71541
10	1.812	2.228	3.169	*24*	1.71088	2.0639	2.79694	*38*	1.68595	2.02439	2.71156
11	1.796	2.201	3.106	*25*	1.70814	2.05954	2.78744	*39*	1.68488	2.02269	2.70791
12	1.782	2.179	3.055	*26*	1.70562	2.05553	2.77871	*40*	1.68385	2.02108	2.70446
13	1.771	2.160	3.012	*27*	1.70329	2.05183	2.77068	*41*	1.68288	2.01954	2.70118
14	1.761	2.145	2.977	*28*	1.70113	2.04841	2.76326	*42*	1.68195	2.01808	2.69807
15	1.753	2.131	2.947	*29*	1.69913	2.04523	2.75639	*43*	1.68107	2.01669	2.6951

Forecasting

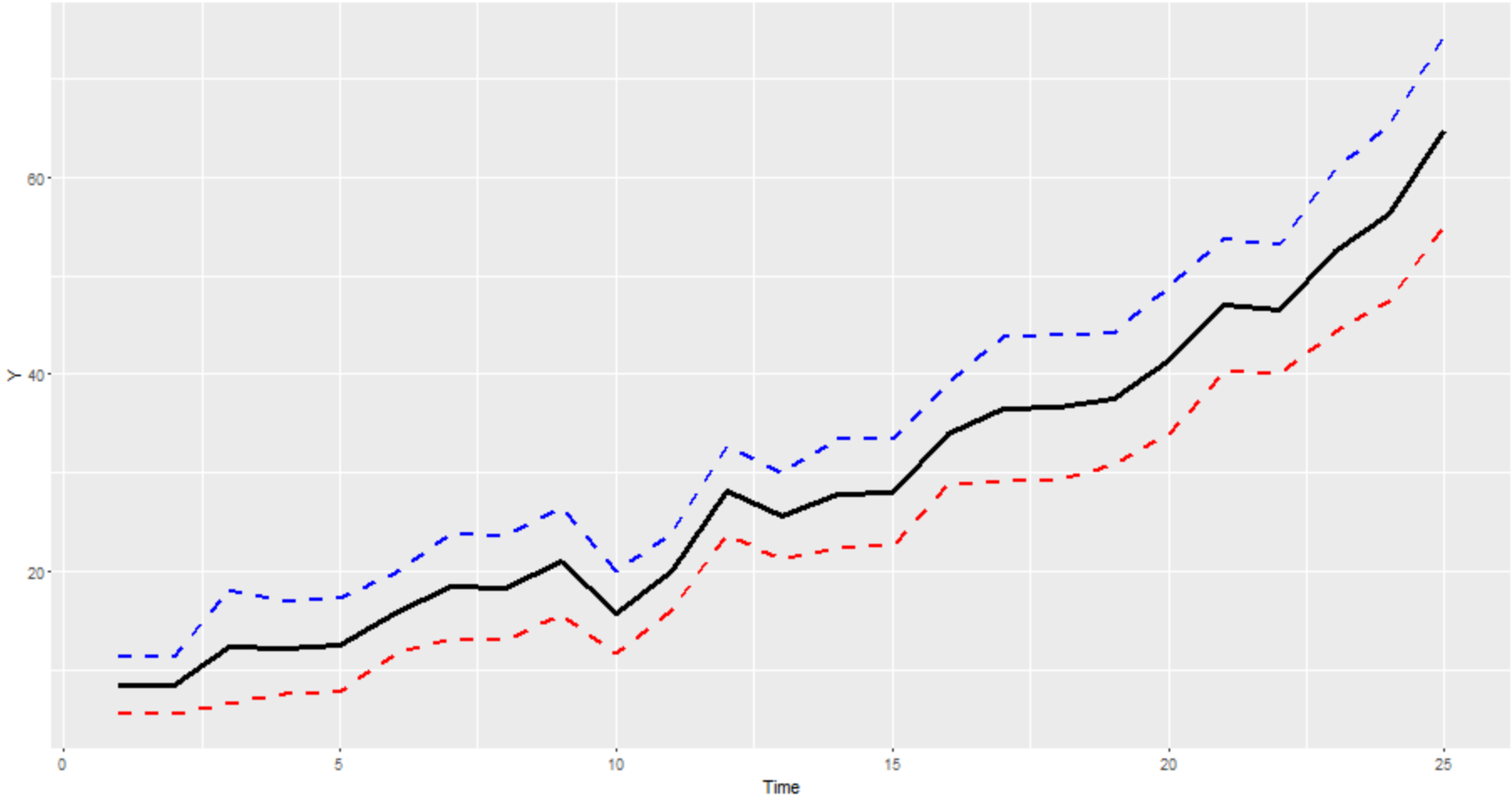
$$B_0 \pm t_{25-(2+1);2.5\%} \times SE_{B0} = 2.264 \pm 2.07387 \times 1.060 = [0.065; 4.462]$$

$$B_1 \pm t_{25-(2+1);2.5\%} \times SE_{B1} = 2.744 \pm 2.07387 \times 0.0935 = [2.550; 2.938]$$

$$B_2 \pm t_{25-(2+1);2.5\%} \times SE_{B2} = 0.012 \pm 2.07387 \times 0.003 = [0.007; 0.018]$$

Forecasting

Observation	Y	X1	X2	95%L	$\hat{y}$	95%U
1	9.6	2	52	5.52	8.40	11.29
2	9.95	2	50	5.50	8.38	11.26
3	10.3	1	585	6.55	12.34	18.12
4	11.66	2	360	7.59	12.26	16.94
5	14.38	2	375	7.69	12.45	17.21
6	16.86	4	200	11.61	15.75	19.88
7	17.08	4	412	13.04	18.40	23.77
8	17.89	4	400	12.96	18.25	23.55
9	21.15	5	400	15.51	21.00	26.49
10	21.65	4	205	11.65	15.81	19.97
11	22.13	6	100	16.04	19.98	23.92
12	24.35	9	100	23.69	28.21	32.74
13	24.45	8	110	21.21	25.60	29.98
14	25.02	8	295	22.45	27.91	33.38
15	27.5	8	300	22.49	27.98	33.47
16	31.75	11	120	28.93	33.95	38.98
17	34.93	10	540	29.20	36.47	43.74
18	35	10	550	29.27	36.60	43.93
19	37	11	400	30.81	37.46	44.12
20	41.95	12	500	34.03	41.46	48.89
21	44.88	15	290	40.27	47.06	53.85
22	46.59	15	250	40.00	46.56	53.12
23	54.12	16	510	44.30	52.56	60.82
24	56.63	17	590	47.39	56.31	65.23
25	69	20	600	55.11	64.67	74.23


60
40
20
Y
0
5
10
15
20
25
Time

Forecasting

The following hypothesis test for B_i can then be done:

$$H_0: B_i = 0\ (There\ is\ not\ a\ linear\ relation\ between\ x_i\ e\ y)$$
$$H_1: B_i \neq 0\ (There\ is\ a\ linear\ relation\ between\ x_i\ e\ y)$$

$$t_{test} = \frac{B_i}{SE_{Bi}}$$

$$t_{critical} = t_{n-(k+1);\alpha/2}$$

Reject the null hypothesis H_0 if $t_{test} > t_{critical}$ or $t_{test} < -t_{critical}$

Forecasting

Therefore:

$$t_{test} = \frac{B_0}{SE_{B_0}} = \frac{2.264}{1.060} = 2.135;\ t_{critical} = t_{n-(k+1);\alpha/2} = 2.07387$$

Reject the null hypothesis.

$$t_{test} = \frac{B_1}{SE_{B_1}} = \frac{2.744}{0.093} = 29.343;\ t_{critical} = t_{n-(k+1);\alpha/2} = 2.07387$$

Reject the null hypothesis.

$$t_{test} = \frac{B_2}{SE_{B_2}} = \frac{0.012}{0.003} = 4.477;\ t_{critical} = t_{n-(k+1);\alpha/2} = 2.07387$$

Reject the null hypothesis.

Forecasting

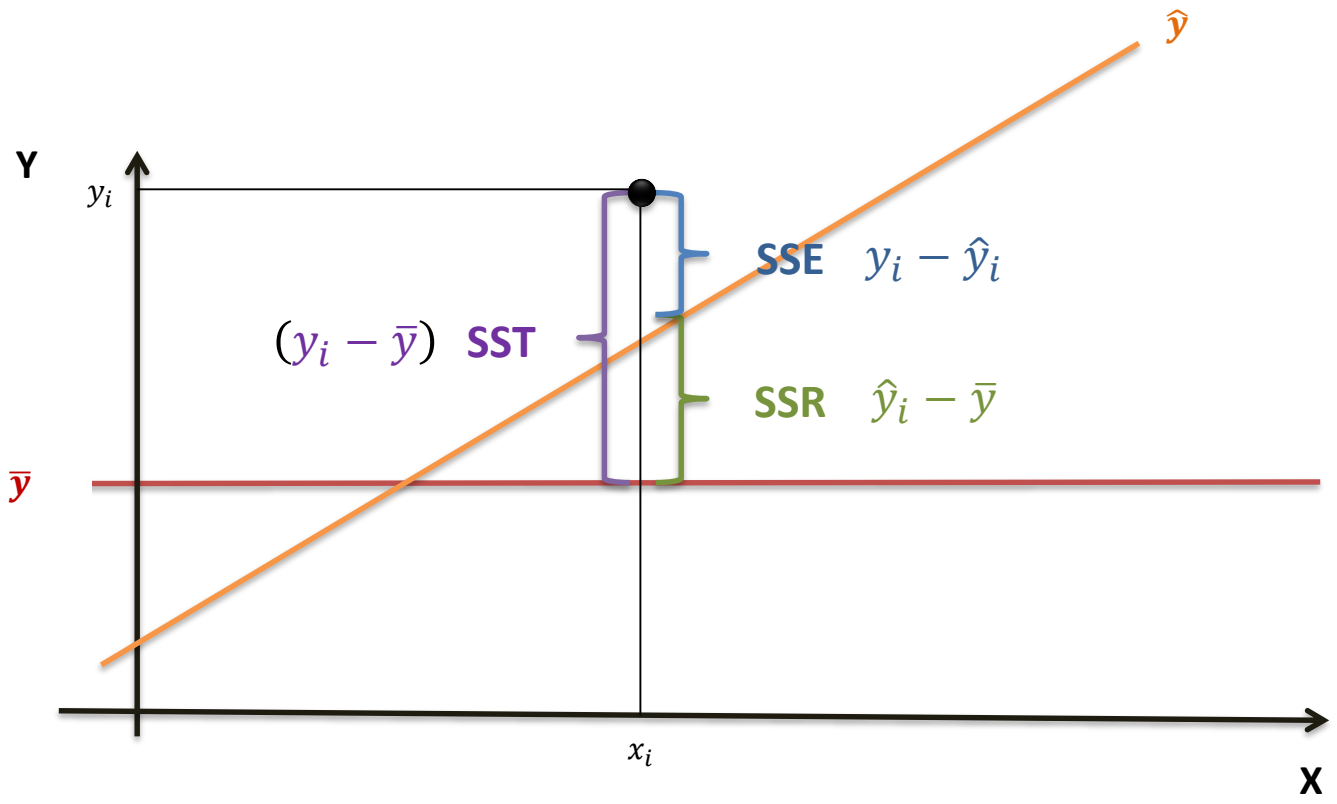
To evaluate the adequacy of the model with the data, we need to find the total variance sum of squares (SST) that is formed by explained variance or the regression sum of squares (SSR) and unexplained variance or error sum os squares (SSE):

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SST = SSR + SSE$$
$$6105.944 = 5990.771 + 115.173$$

- SST = measures the variation of y_i values around their average $\bar{y}$;
- SSR = Variation given by the relationship between X and Y;
- SSE = Variation of Y attributed to other factors than X.

Forecasting



Forecasting

The following hypothesis is made to test the model adequacy:

$$H_0: The\ model\ is\ not\ adequate$$
$$H_1: The\ model\ is\ adequate$$

$$F_{test} = \frac{[\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2]/(k)}{[\sum_{i=1}^{n}(y_i - \hat{y}_i)^2]/[n-(k+1)]}$$

Reject the null hypothesis H_0 if $F_{test} > F_{k;n-(k+1);\alpha}$

Snedecor's F Distribution Table

					α = 5,0%					
V2	V1									
	1	2	3	4	5	6	7	8	9	10
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
31	4,16	3,30	2,91	2,68	2,52	2,41	2,32	2,25	2,20	2,15
32	4,15	3,29	2,90	2,67	2,51	2,40	2,31	2,24	2,19	2,14
33	4,14	3,28	2,89	2,66	2,50	2,39	2,30	2,23	2,18	2,13
34	4,13	3,28	2,88	2,65	2,49	2,38	2,29	2,23	2,17	2,12
35	4,12	3,27	2,87	2,64	2,49	2,37	2,29	2,22	2,16	2,11

Snedecor's F Distribution Table

					α = 5,0%					
V2	V1									
	1	2	3	4	5	6	7	8	9	10
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32		3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4	3,44	05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28		3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
31	4,16	3,30	2,91	2,68	2,52	2,41	2,32	2,25	2,20	2,15
32	4,15	3,29	2,90	2,67	2,51	2,40	2,31	2,24	2,19	2,14
33	4,14	3,28	2,89	2,66	2,50	2,39	2,30	2,23	2,18	2,13
34	4,13	3,28	2,88	2,65	2,49	2,38	2,29	2,23	2,17	2,12
35	4,12	3,27	2,87	2,64	2,49	2,37	2,29	2,22	2,16	2,11

Forecasting

Therefore:

$$F_{k;n-(k+1);\alpha} = F_{2;25-(2+1);5\%} = 3.44$$

$$F_{test} = \frac{[\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2]/(k)}{[\sum_{i=1}^{n}(y_i - \hat{y}_i)^2]/[n-(k+1)]}$$

$$F_{test} = \frac{5990.771/(2)}{115.173/[25-(2+1)]} = \frac{2995.386}{5.235} = 572.167$$

Reject the null hypothesis.

Forecasting

The r^2 (r-squared or coefficient of determination) measures the strength of the relationship, indicating that the model explains a percentage of the variance of the dependent variable. For example, a r^2 equal to 0.80 means that 80% of the variance of the dependent variable comes from its relation with the independent variables. The $(r^2)_a$ - adjusted r^2 - is na indicator that adjusts the r^2 based on the number of k independent variables and it is useful to penalize models that uses too many independent variables.

Forecasting

The covariance (ρ) is a measure of how changes in one variable are associated with changes in a second variable. Specifically, covariance measures the degree to which two variables are linearly associated. The correlation (r) is the standardized measure of covariance ranging between -1 and 1.

The r^2 (r-squared or coefficient of determination) measures the strength of the relationship, indicating that the model explains a percentage of the variance of the dependent variable. For example, a r^2 equal to 0.80 means that 80% of the variance of the dependent variable comes from its relation with the independent variables.

Forecasting

Both indicators can be calculated as:

$$r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{115.173}{6105.944} = 0.981$$

$$r = \sqrt{r^2} = 0.990$$

The $(r^2)_a$ - adjusted r^2 - is na indicator that adjusts the r^2 based on the number of k independent variables:

$$(r^2)_a = 1 - (1 - r^2)\frac{n-1}{n-k-1}$$

$$(r^2)_a = 1 - (1 - 0.981)\frac{25-1}{25-2-1} = 0.979$$

The $(r^2)_a$ is useful to penalize models that uses too many independent variables.

Forecasting

The significance of each correlation can be tested by the following hypothesis:

$$H_0: The\ correlation\ is\ not\ significant$$
$$H_1: The\ correlation\ is\ significant$$

$$t_{test} = \frac{r}{\sqrt{\frac{1-(r)^2}{n-2}}}$$

$$t_{critical} = t_{n-2;\alpha/2}$$

Reject the null hypothesis H_0 if $t_{test} > t_{critical}$ or $t_{test} < -t_{critical}$

Forecasting

Therefore for a 95% confidence interval:

$$t_{test} = \frac{r}{\sqrt{\frac{1-(r)^2}{n-2}}} = \frac{0.990}{\sqrt{\frac{1-(0.990\ )^2}{25-2}}} = 34.588$$

$$t_{critical} = t_{25-2;2.5\%} = 2.06866$$

Reject the null hypothesis.

Forecasting

The collinearity (correlation between two predictors) or multicollinearity (correlation between multiple predictors) can be harmful for the model because:

- Estimates may be unstable;
- Standard errors can be unreliable;
- Confidence intervals can become very large;
- The coefficient of determination can be high, even though the T-tests are insignificant.

To detect this problem is necessary to calculate the **VIF** (***Variance Inflator Factor***) for each predictor. If its value is > 5 the predictor is highly correlated indicating collinearity:

$$VIF_j = \frac{1}{1 - (r^2)_j}$$

Where:

$(r^2)_j$ = Coefficient of determination between the $j\text{-}th$ predictor and all other predictors.

$1 - (r^2)_j$ = Tolerance between the $j\text{-}th$ predictor and all other predictors. The higher the tolerance, the less likely to occur collinearity or multicollinearity.

Forecasting

Therefore, $VIF_{x_1x_2} = VIF_{x_2x_1}$:

$$VIF_{x_1x_2} = \frac{1}{1-(r^2)_{x_1x_2}} = \frac{1}{1-0.143} = 1.167$$

There is no collineartiy between the variables.

Suggest interpretation:

$(r^2)_j$	$VIF_j = \frac{1}{1-(r^2)_j}$	**Interpretation**
0.00	$VIF_j = 1.00$	Insignifiant
0.50	$VIF_j = 2.00$	Medium
0.90	$VIF_j = 10.00$	Strong
0.99	$VIF_j = 100.00$	Severe

Forecasting

Misinterpretations of determination and correlation coefficients:

- "A high correlation coefficient indicates that useful predictions can be made".

This is not necessarily correct because if margins of the confidence intervals are large, the model is not very accurate.

- "A high correlation coefficient indicates that the estimated regression equation is well fitted to the data and a coefficient close to zero indicates that the variables are uncorrelated".

This is not necessarily correct because the correlation considers a linear relation and if a non-linear relation occurs, wrong interpretations can be made.

Residuals

Forecasting

Once verified the adequacy of the estimated model, it is also necessary to validate the errors (residuals) of the model. Supposedly, the residuals must be:

- Normally distributed;
- Homoscedastic;
- Independent (uncorrelated within a series of time).

The studentized residuals are most indicated approach to proceed with the evaluation. The studentized residual is the quotient resulting from the division of a residual by an estimate of its standard deviation, usually ranging from -3 to +3.

Forecasting

- Normally Distributed

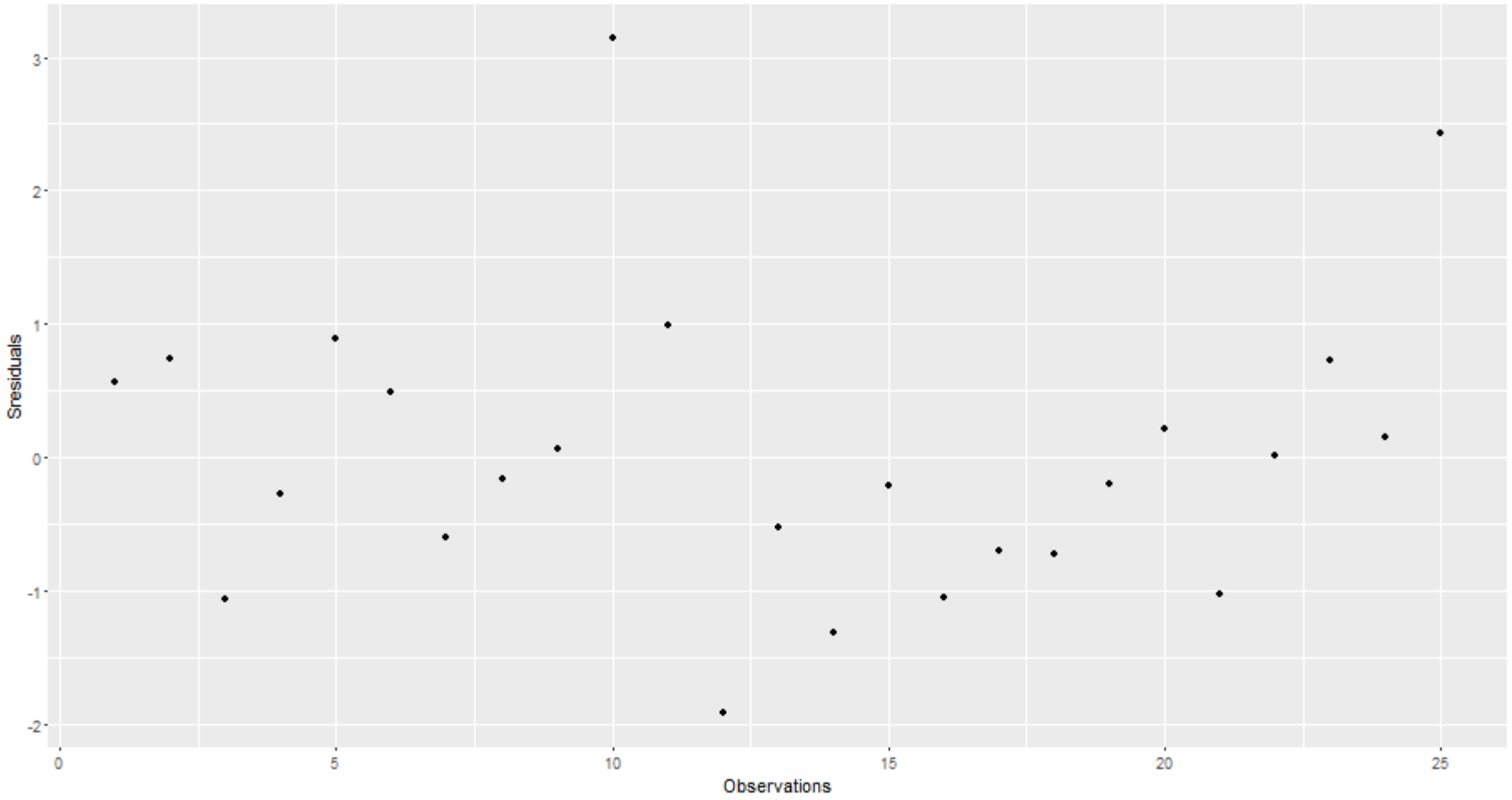
Violation of this assumption is considered mild and can make the confidence intervals unreliable. Large samples, logarithmic transformations in the dependent and independent variables or removal of outliers, can avoid this violation.

Forecasting

- Homocedastic

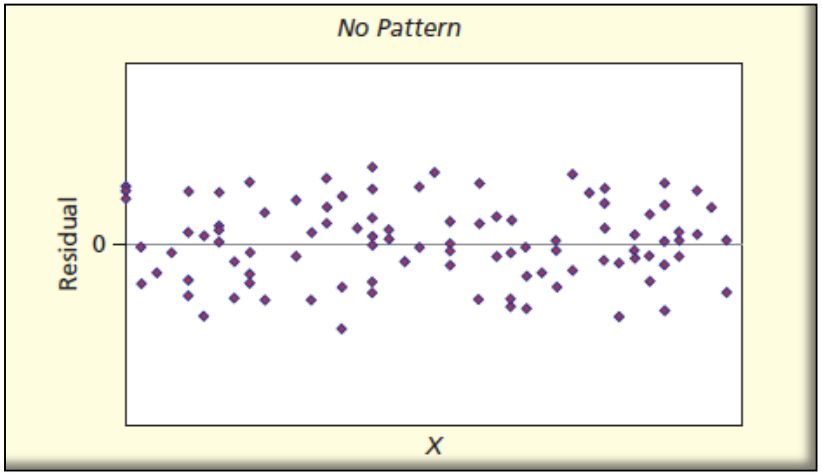
Violation of this assumption is considered severe and can increase the range of confidence intervals and make the model unfeasible. Large samples, logarithmic transformations in the dependent and independent variables or removal of outliers, can avoid this violation. To verify that the errors are homoscedastic (constant variance) the residual graphic, ideally, will have no patterns.

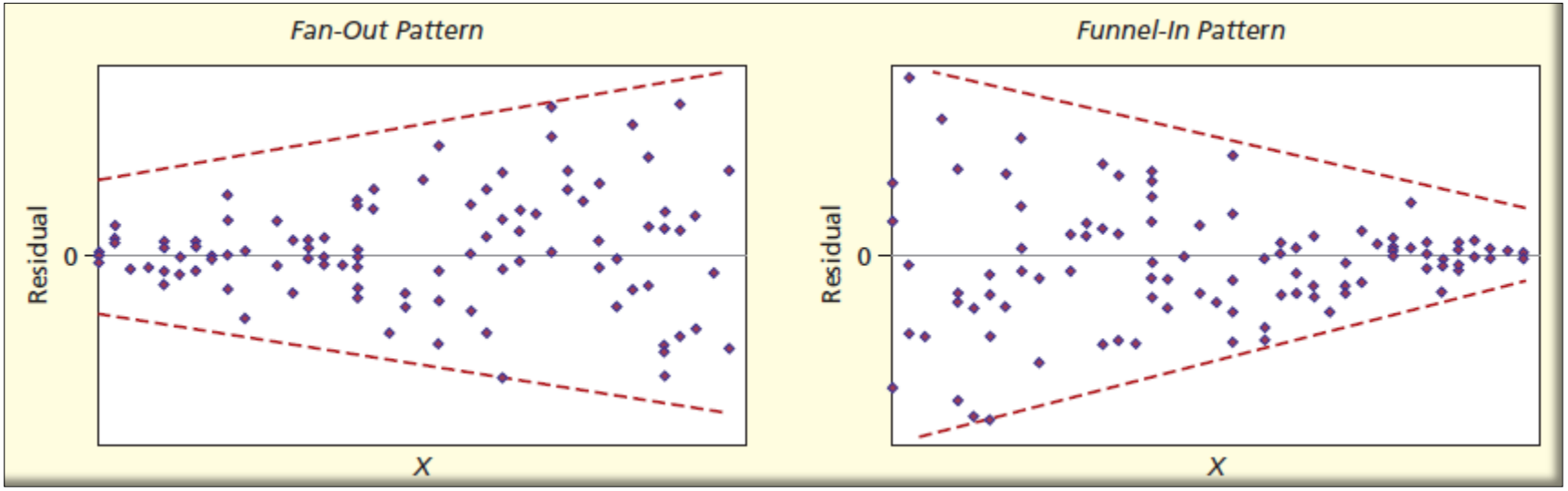
Forecasting



Forecasting

To verify that the errors are homoscedastic (constant variance) the residual graphic, ideally, will have no patterns.





Forecasting

The Breusch-Pagan Test can be done to verify the homoscedasticity. It is performed in three steps:

1. Perform multiple linear regression and keep the residuals.
2. Perform multiple linear regression using as the dependent variable squared residuals (found in step 1) keeping the same set of independent variables.
3. Test the hypothesis that the model is homoscedastic through traditional F test, but with the following hypothesis:

$$H_0: The\ model\ is\ homoscedastic$$
$$H_1: The\ model\ is\ not\ homoscedastic$$

Forecasting

Observation	Y	X1	X2	$\hat{y}$	e	e^2
1	9.6	2	52	8.40	1.196	1.431
2	9.95	2	50	8.38	1.571	2.469
3	10.3	1	585	12.34	-2.037	4.149
4	11.66	2	360	12.26	-0.602	0.363
5	14.38	2	375	12.45	1.930	3.724
6	16.86	4	200	15.75	1.114	1.240
7	17.08	4	412	18.40	-1.322	1.749
8	17.89	4	400	18.25	-0.362	0.131
9	21.15	5	400	21.00	0.154	0.024
10	21.65	4	205	15.81	5.841	34.116
11	22.13	6	100	19.98	2.148	4.613
12	24.35	9	100	28.21	-3.865	14.938
13	24.45	8	110	25.60	-1.146	1.313
14	25.02	8	295	27.91	-2.894	8.373
15	27.5	8	300	27.98	-0.476	0.227
16	31.75	11	120	33.95	-2.204	4.858
17	34.93	10	540	36.47	-1.542	2.376
18	35	10	550	36.60	-1.597	2.550
19	37	11	400	37.46	-0.462	0.213
20	41.95	12	500	41.46	0.491	0.241
21	44.88	15	290	47.06	-2.181	4.756
22	46.59	15	250	46.56	0.030	0.001
23	54.12	16	510	52.56	1.559	2.430
24	56.63	17	590	56.31	0.322	0.104
25	69	20	600	64.67	4.334	18.785

$$p\text{-}value = 0.773$$

Accept the null hypothesis.

Forecasting

The White Test can be done to verify the homoscedasticity. It is performed in three steps:

1. Perform multiple linear regression and keep the residuals.
2. Perform multiple linear regression using as the dependent variable squared residuals (found in step 1) use as independent variables the predicted value and the squared predicted value.
3. Test the hypothesis that the model is homoscedastic through traditional F test, but with the following hypothesis:

$$H_0: The\ model\ is\ homoscedastic$$
$$H_1: The\ model\ is\ not\ homoscedastic$$

Forecasting

Observation	Y	X1	X2	$\hat{y}$	$\hat{y}^2$
1	9.6	2	52	8.40	70.62
2	9.95	2	50	8.38	70.20
3	10.3	1	585	12.34	152.20
4	11.66	2	360	12.26	150.37
5	14.38	2	375	12.45	155.01
6	16.86	4	200	15.75	247.95
7	17.08	4	412	18.40	338.65
8	17.89	4	400	18.25	333.14
9	21.15	5	400	21.00	440.84
10	21.65	4	205	15.81	249.93
11	22.13	6	100	19.98	399.29
12	24.35	9	100	28.21	796.09
13	24.45	8	110	25.60	655.16
14	25.02	8	295	27.91	779.17
15	27.5	8	300	27.98	782.67
16	31.75	11	120	33.95	1152.88
17	34.93	10	540	36.47	1330.17
18	35	10	550	36.60	1339.32
19	37	11	400	37.46	1403.39
20	41.95	12	500	41.46	1718.84
21	44.88	15	290	47.06	2214.73
22	46.59	15	250	46.56	2167.81
23	54.12	16	510	52.56	2762.69
24	56.63	17	590	56.31	3170.57
25	69	20	600	64.67	4181.67

$$p\text{-}value = 0.692$$

Accept the null hypothesis.

Forecasting

- Independency

Violation of this assumption is considered mild and can increase the range of the confidence intervals. This assumption is only considered in time series.

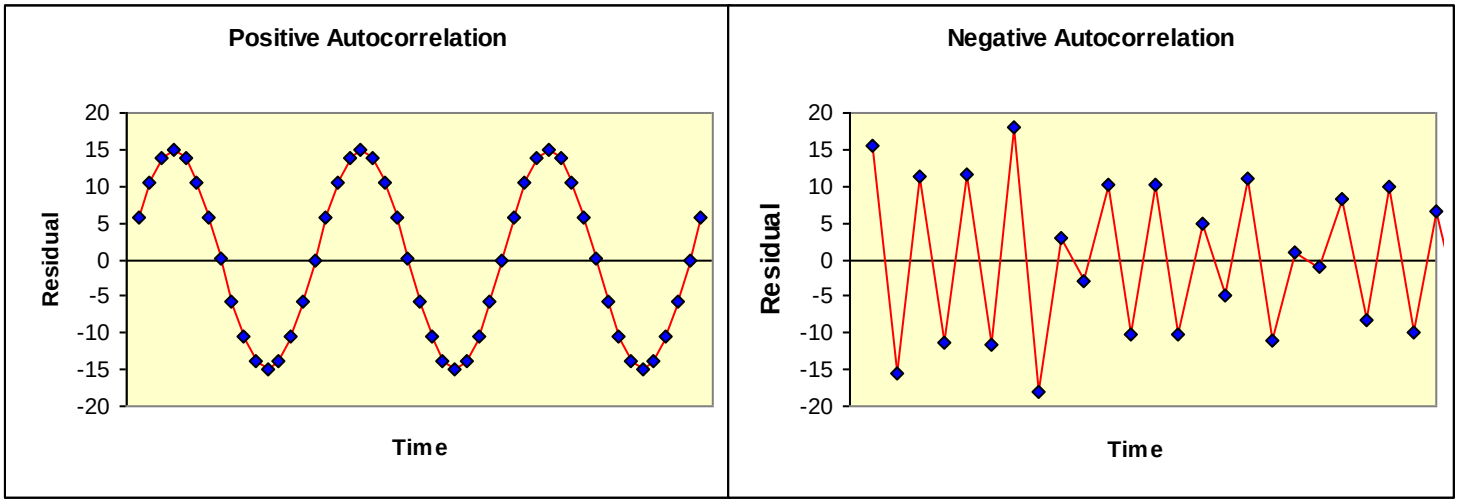
A differentiation of the first order or the removal of outliers, can avoid this violation.

Forecasting

To verify the independency of the errors (not autocorrelated), the Durbin-Watson test (ranging from 0 to 4) can be done:

- $DW < 2 \rightarrow$ Indicates positive autocorrelation (Common)
- $DW \cong 2 \rightarrow$ No autocorrelation
- $DW > 2 \rightarrow$ Indicates negative autocorrelation (Rare)

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n}(e_i)^2}$$



Forecasting

Alternatively the following hypothesis tes can be performe:

$$H_0: The\ residuals\ are\ not\ independent$$
$$H_1: The\ residuals\ are\ independent$$

$$DW_{test} = DW$$

$$DW_{critical} = (d_l; d_u)$$

- Reject the null hypothesis H_0 if $DW_{test} < d_l$ or $DW_{test} > 4 - d_l$
- Accept the null hypothesis H_0 if $d_u < DW_{test} < 4 - d_u$
- The test is inconclusive if $d_l < DW_{test} < d_u$ or $4 - d_l < DW_{test} < 4 - d_u$

Forecasting

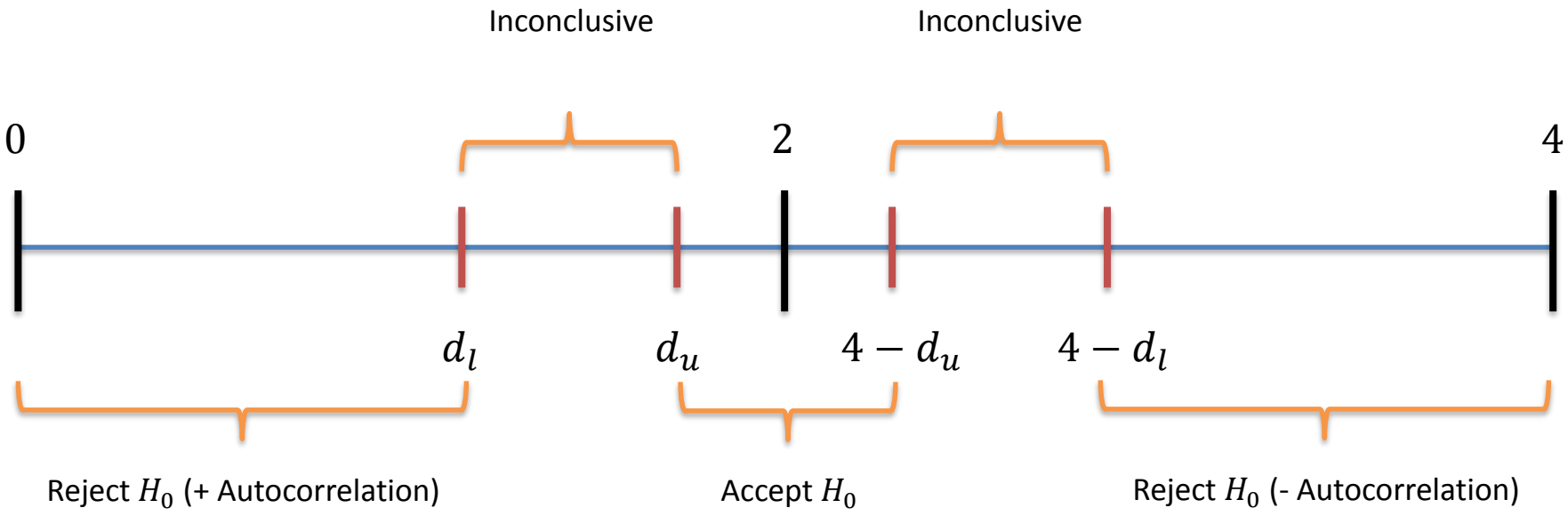


Tabela Durbin-Watson (99%)

	k' =1		k'=2		k'=3		k'=4		k'=5		k'=6		k'=7		k'=8		k'=9		k'=10	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
6	0.390	1.142	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
7	0.435	1.036	0.294	1.676	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
8	0.497	1.003	0.345	1.489	0.229	2.102	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
9	0.554	0.998	0.408	1.389	0.279	1.875	0.183	2.433	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
10	0.604	1.001	0.466	1.333	0.340	1.733	0.230	2.193	0.150	2.690	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
11	0.653	1.010	0.519	1.297	0.396	1.640	0.286	2.030	0.193	2.453	0.124	2.892	-----	-----	-----	-----	-----	-----	-----	-----
12	0.697	1.023	0.569	1.274	0.449	1.575	0.339	1.913	0.244	2.280	0.164	2.665	0.105	3.053	-----	-----	-----	-----	-----	-----
13	0.738	1.038	0.616	1.261	0.499	1.526	0.391	1.826	0.294	2.150	0.211	2.490	0.140	2.838	0.090	3.182	-----	-----	-----	-----
14	0.776	1.054	0.660	1.254	0.547	1.490	0.441	1.757	0.343	2.049	0.257	2.354	0.183	2.667	0.122	2.981	0.078	3.287	-----	-----
15	0.811	1.070	0.700	1.252	0.591	1.465	0.487	1.705	0.390	1.967	0.303	2.244	0.226	2.530	0.161	2.817	0.107	3.101	0.068	3.374
16	0.844	1.086	0.738	1.253	0.633	1.447	0.532	1.664	0.437	1.901	0.349	2.153	0.269	2.416	0.200	2.681	0.142	2.944	0.094	3.201
17	0.873	1.102	0.773	1.255	0.672	1.432	0.574	1.631	0.481	1.847	0.393	2.078	0.313	2.319	0.241	2.566	0.179	2.811	0.127	3.053
18	0.902	1.118	0.805	1.259	0.708	1.422	0.614	1.604	0.522	1.803	0.435	2.015	0.355	2.238	0.282	2.467	0.216	2.697	0.160	2.925
19	0.928	1.133	0.835	1.264	0.742	1.416	0.650	1.583	0.561	1.767	0.476	1.963	0.396	2.169	0.322	2.381	0.255	2.597	0.196	2.813
20	0.952	1.147	0.862	1.270	0.774	1.410	0.684	1.567	0.598	1.736	0.515	1.918	0.436	2.110	0.362	2.308	0.294	2.510	0.232	2.174
21	0.975	1.161	0.889	1.276	0.803	1.408	0.718	1.554	0.634	1.712	0.552	1.881	0.474	2.059	0.400	2.244	0.331	2.434	0.268	2.625
22	0.997	1.174	0.915	1.284	0.832	1.407	0.748	1.543	0.666	1.691	0.587	1.849	0.510	2.015	0.437	2.188	0.368	2.367	0.304	2.548
23	1.017	1.186	0.938	1.290	0.858	1.407	0.777	1.535	0.699	1.674	0.620	1.821	0.545	1.977	0.473	2.140	0.404	2.308	0.340	2.479
24	1.037	1.199	0.959	1.298	0.881	1.407	0.805	1.527	0.728	1.659	0.652	1.797	0.578	1.944	0.507	2.097	0.439	2.255	0.375	2.417
25	1.055	1.210	0.981	1.305	0.906	1.408	0.832	1.521	0.756	1.645	0.682	1.776	0.610	1.915	0.540	2.059	0.473	2.209	0.409	2.362

Forecasting

Therefore considering a 99% confidence interval:

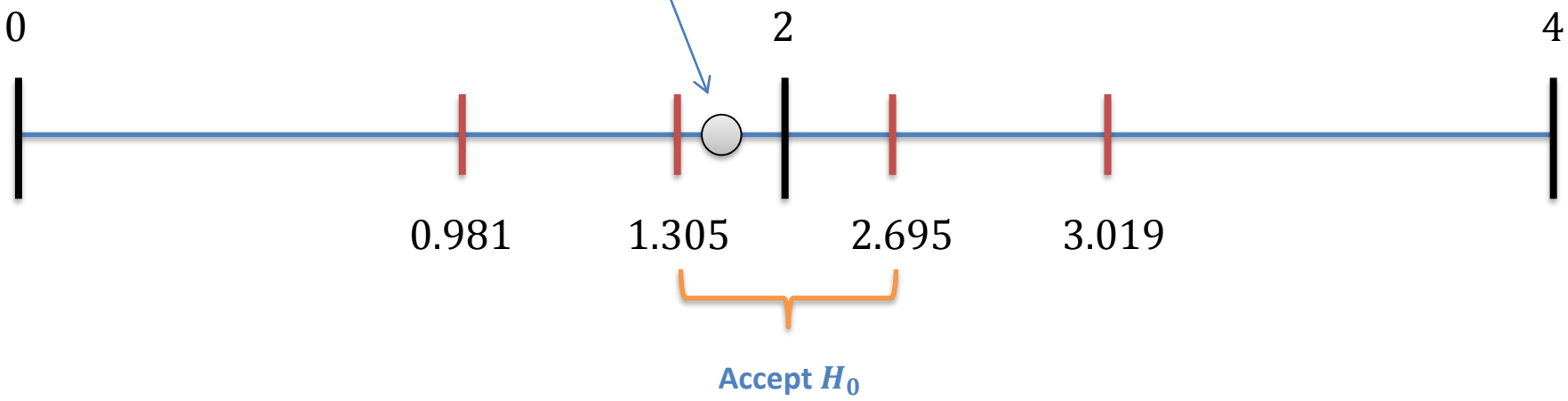
$$DW_{test} = DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n}(e_i)^2} = \frac{165.457}{115.173} = 1.436$$

Tabela Durbin-Watson (99%)

	k' =1		k'=2		k'=3		k'=4		k'=5		k'=6		k'=7		k'=8		k'=9		k'=10	
n	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
6	0.390	1.142	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
7	0.435	1.036	0.294	1.676	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
8	0.497	1.003	0.345	1.489	0.229	2.102	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
9	0.554	0.998	0.408	1.389	0.279	1.875	0.183	2.433	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
10	0.604	1.001	0.466	1.333	0.340	1.733	0.230	2.193	0.150	2.690	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
11	0.653	1.010	0.519	1.297	0.396	1.640	0.286	2.030	0.193	2.453	0.124	2.892	-----	-----	-----	-----	-----	-----	-----	-----
12	0.697	1.023	0.569	1.274	0.449	1.575	0.339	1.913	0.244	2.280	0.164	2.665	0.105	3.053	-----	-----	-----	-----	-----	-----
13	0.738	1.038	0.616	1.261	0.499	1.526	0.391	1.826	0.294	2.150	0.211	2.490	0.140	2.838	0.090	3.182	-----	-----	-----	-----
14	0.776	1.054	0.660	1.254	0.547	1.490	0.441	1.757	0.343	2.049	0.257	2.354	0.183	2.667	0.122	2.981	0.078	3.287	-----	-----
15	0.811	1.070	0.700	1.252	0.591	1.465	0.487	1.705	0.390	1.967	0.303	2.244	0.226	2.530	0.161	2.817	0.107	3.101	0.068	3.374
16	0.844	1.086	0.738	1.253	0.633	1.447	0.532	1.664	0.437	1.901	0.349	2.153	0.269	2.416	0.200	2.681	0.142	2.944	0.094	3.201
17	0.873	1.102	0.773	1.255	0.672	1.432	0.574	1.631	0.481	1.847	0.393	2.078	0.313	2.319	0.241	2.566	0.179	2.811	0.127	3.053
18	0.902	1.118	0.805	1.259	0.708	1.422	0.614	1.604	0.522	1.803	0.435	2.015	0.355	2.238	0.282	2.467	0.216	2.697	0.160	2.925
19	0.928	1.133	0.835	1.264	0.742	1.416	0.650	1.583	0.561	1.767	0.476	1.963	0.396	2.169	0.322	2.381	0.255	2.597	0.196	2.813
20	0.952	1.147	0.862	1.270	0.774	1.410	0.684	1.567	0.598	1.736	0.515	1.918	0.436	2.110	0.362	2.308	0.294	2.510	0.232	2.174
21	0.975	1.161	0.889	1.276	0.803	1.408	0.718	1.554	0.634	1.712	0.552	1.881	0.474	2.059	0.400	2.244	0.331	2.434	0.268	2.625
22	0.997	1.174	0.915	1.284	0.832	1.407	0.748	1.543	0.666	1.691	0.587	1.849	0.510	2.015	0.437	2.188	0.368	2.367	0.304	2.548
23	1.017	1.186	0.938	1.290	0.858	1.407	0.777	1.535	0.699	1.674	0.620	1.821	0.545	1.977	0.473	2.140	0.404	2.308	0.340	2.479
24	1.037	1.199	[illegible]	[illegible]	0.881	1.407	0.805	1.527	0.728	1.659	0.652	1.797	0.578	1.944	0.507	2.097	0.439	2.255	0.375	2.417
25	1.055	1.210	0.981	1.305	0.906	1.408	0.832	1.521	0.756	1.645	0.682	1.776	0.610	1.915	0.540	2.059	0.473	2.209	0.409	2.362

Forecasting

$DW = 1.436;\ d_l = 0.981;\ d_u = 1.305$



Bibliography

CORRAR, L.J.; PAULO, E.; DIAS FILHO, J. M. **Análise Multivariada para Cursos de Administração, Ciências Contábeis e Economia**. ATLAS, 2009.

FÁVERO, L. P.; BELFIORE, P.; SILVA, F. L.; CHAN, B. **Análise de Dados: Modelagem Multivariada para Tomada de Decisões**. CAMPUS, 2009.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise Multivariada de Dados**. BOOKMAN, 2009.

LATTIN, J.; CARROLL, J. D.; GREEN, P. E. **Análise de Dados Multivariados**. CENGAGE Learning, 2011.

LEVINE, D. M.; STEPHAN, D. F.; KREHBIEL, T. C.; BERENSON, M. L. **Estatística - Teoria e Aplicações - Usando Microsoft Excel**. LTC, 2012.