Saint Joseph University - Year 2023-2024

Data Science License - Statistical analysis of data

TD3 Sheet – Multiole Linear Regression

EXERCISE 1

The data presented in the table below concerns 9 companies in the chemical industry. We seek to establish a relationship between production y_i, working hours x_{i1} and capital employed x_{i2}.

Tableau - Production, travail et capital

Entreprise i	Travail (heures) x_{i1}	Capital (machines/heures) x_{i2}	Production (100 tonnes) y_i
1	1 100	300	60
2	1 200	400	120
3	1 430	420	190
4	1 500	400	250
5	1 520	510	300
6	1 620	590	360
7	1 800	600	380
8	1 820	630	430
9	1 800	610	440

The model is a multiple linear regression model with two explanatory variables, i.e. for all $i = 1 \ldots 9$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

1. Graph Y versus X1 and X2.
2. Test if the ε_i are normally distributed.
3. Determine a regression equation by specifying what are the explanatory variables and the explained variable?
4. Estimate parameters β_0, β_1 and β_2
5. Construct a 95% confidence interval for parameters β_0, β_1 and β_2.
6. Establish the ANOVA table associated with this regression.

7. Test the hypothesis $H_0: \beta_j = 0$ (against $H_1: \beta_j \neq 0$) with a significance threshold α = 5% for j = 1, 2.

EXERCISE 2

The data provided in this exercise presents the rate of death from heart attacks among men aged 55 to 59 in different countries. The variables are as follows:

- Y: 100 [log (number of deaths by heart attack per 100,000 men aged 55 to 59 years) -2].
- X_1: telephones per thousand inhabitants.
- X_2: fat calories as a percentage of total calories.
- X_3: calories from animal protein as a percentage of total calories.

1. Regress Y on X_1 and test the significance of this simple linear regression.
2. Find the equation of the multiple linear regression of Y on X_1 and X_2.
3. Establish the ANOVA table associated with this regression.
4. Test whether the addition of the variable X_2 to the equation found in question 1.
5. Construct the multiple linear regression of Y on X_1, X_2 and X_3.
6. Give the limits of the 95% confidence interval for β_3 in this equation.
7. Give the limits of the 95% confidence interval for $\hat{Y}$ at the point X_1 = 221, X_2 = 39 and X_3 = 7.
8. Test whether X_2 and X_3 together add anything to the simple linear regression of Y on X_1.
9. Regress X_1 on X_2 and X_3.
10. Give the limits of the 95% confidence interval for the coefficients of the linear regression of X_1 on X_3.

Observation i	Pays	X_1 x_{1i}	X_2 x_{2i}	X_3 x_{3i}	Y x_{4i}
1	Australie	124	33	8	81
2	Autriche	49	31	6	55
3	Canada	181	38	8	80
4	Ceylan	4	17	2	24
5	Chili	22	20	4	78
6	Danemark	152	39	6	52
7	Finlande	75	30	7	88
8	France	54	29	7	45
9	Allemagne	43	35	6	50
10	Irlande	41	31	5	69
11	Israel	17	23	4	66
12	Italie	22	21	3	45
13	Japon	16	8	3	24
14	Mexique	10	23	3	43
15	Pays-Bas	63	37	6	38
16	Nouvelle-Zélande	170	40	8	72
17	Norvège	125	38	6	41
18	Portugal	12	25	4	38
19	Suède	221	39	7	52
20	Suisse	171	33	7	52
21	Grande-Bretagne	97	38	6	66
22	États-Unis	254	39	8	89

EXERCISE 3

We use the multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

1. Complete the following analysis of variance table:

Source of variation	Sum of squares	df	Mean squares	F_{obs}
Regression	1504.4			
Residual			19.6	
Total	1680.8			

2. Test the hypothesis $H_0: \beta_1 = \beta_2 = 0$ (against H_1: at least one of $\beta \neq 0$) with a significance threshold α = 5%.
3. What is the coefficient of determination R^2 of the model?
4. Give an estimate of the variance of ε.

EXERCISE 4

A human resources consulting firm carried out a study on the anxiety level Y measured on a scale of 1 to 50 of business executives over a period of two weeks. We want to examine whether the following factors can influence the level of anxiety of executives:

- X_1: systolic blood pressure
- X_2: test assessing managerial abilities
- X_3: level of satisfaction with the position occupied.

The analysis of variance table indicates the contribution of each variable introduced in the order indicated and this for 22 frames.

Source de variation	Somme des carrés	*ddl*
Régression due à X_1	981, 326	1
Régression due à X_2	190, 232	1
Régression due à X_3	129, 431	1
Résiduelle	442, 292	18
Totale	1743, 281	21

1. What is the sum of squares due to the regression for all three explanatory variables?
2. What proportion of the variation in anxiety level is explained by the three explanatory variables?
3. Can we conclude that overall the three explanatory variables have a significant effect on the level of anxiety? Use a significance level of α = 5%. Specify the hypotheses we want to test.
4. If we only take into account the explanatory variable X_1, what would then be the corresponding analysis of variance table?

Source of variation	Sum of squares	df
Regression due to X_1	981.326	
Residual		
Total		

5. Test the following null hypotheses, at significance level = 5%, using an appropriate F ratio:
   $H_0: \beta_1 = 0$ in model $Y = \beta_0 + \beta_1 X_1 + \varepsilon$
   $H_0: \beta_2 = 0$ in model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
   $H_0: \beta_3 = 0$ in the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

6. What is the value of the coefficient of determination R^2 associated with the estimation of each model specified in question 5.?
7. Which of the three models seems best suited to explain fluctuations in the level of anxiety of business executives?

EXERCISE 5

The CITRON company manufactures a plastic material which is used in the manufacture of toys. The company's quality control department carried out a study aimed at establishing to what extent the breaking strength (in kg/cm2) of this plastic material could be affected by the thickness of the material as well as the density of this material. Twelve tests were carried out and the results are presented in the table below.

Essai numéro	Résistance à la rupture Y_i	Épaisseur du matériau X_{i_1}	Densité X_{i_2}
1	37,8	4	4,0
2	22,5	4	3,6
3	17,1	3	3,1
4	10,8	2	3,2
5	7,2	1	3,0
6	42,3	6	3,8
7	30,2	4	3,8
8	19,4	4	2,9
9	14,8	1	3,8
10	9,5	1	2,8
11	32,4	3	3,4
12	21,6	4	2,8

Various analyzes were carried out on computer and we summarize them in the following three tables:

Régression de la résistance à la rupture Y en fonction de l'épaisseur X_1

Coefficients $\hat{\beta}_j$	Erreurs-types $s(\hat{\beta}_j)$
$\hat{\beta}_0 = 3.523$	
$\hat{\beta}_1 = 6.036$	$s(\hat{\beta}_1) = 1.279$

Source de variation	Somme des carrés	*ddl*
Régression X_1	980.63	1
Résiduelle	440.03	110

Régression de la résistance à la rupture Y en fonction de la densité X_2

Coefficients $\hat{\beta}_j$	Erreurs-types $s(\hat{\beta}_j)$
$\hat{\beta}_0 = -36.373$	
$\hat{\beta}_2 = 17.464$	$s(\hat{\beta}_2) = 6.069$

Source de variation	Somme des carrés	*ddl*
Régression X_2	643.57	1
Résiduelle	777.10	10

Régression de la résistance à la rupture Y en fonction de l'épaisseur X_1 et de la densité X_2

Coefficients $\hat{\beta}_j$	Erreurs-types $s(\hat{\beta}_j)$
$\hat{\beta}_0 = -30.081$	
$\hat{\beta}_1 = 4.905$	$s(\hat{\beta}_1) = 1.014$
$\beta_2 = 11.072$	$s(\beta_2) = 3.621$

Source de variation	Somme des carrés	*ddl*
Régression (X_1, X_2)	1204.85	2
Résiduelle	215.81	9

1. What percentage of variation in breaking strength is explained by each of the regressions?
2. For each linear regression, complete the following table:

Source of variation	Residual mean squares	Residual standard deviations
Regression due to X_1		
Regression due to X_2		
Regression due to (X_1, X_2)		

3. Complete the following analysis of variance table for the regression including the two explanatory variables.

Source of variation	Sum of squares	df	Mean squares	F_{obs}
Regression due to (X_1 , X_2)				
Residual				
Total				

4. Test the hypothesis $H_0: \beta_1 = \beta_2 = 0$ (against H_1: at least one of $\beta \neq 0$) with a significance threshold α = 5%. What is your conclusion?
5. In the case of the linear regression model including only the thickness of the material as an explanatory variable, determine a 95% confidence interval for β_1.
6. With the confidence interval calculated in question 5., can we affirm, at the α = 5% significance level, that the linear regression is significant between the breaking strength and the thickness of the material? Justify your conclusion.
7. What is the marginal contribution of the variable X_2 when it is introduced following the variable X_1?
8. Is the marginal contribution of the "material density" variable, when it is introduced following the "material thickness" variable, significant at the α = 5% significance level? Use both equivalent ways to perform this test.

"F partiel"	F_c	T_{obs}	T_c

9. We want to obtain various estimates and predictions of breaking strength. What is, on average, the breaking strength of toys for which the thickness of the material used and the density of the material are those indicated in the following table?

Épaisseur X_1	Densité X_2	Estimation de la résistance moyenne	Écart-type de l'estimation
4	3, 8		2, 10
3	3, 4		1, 43
4	2, 9		2, 57

10. Between what values can the average breaking strength fall, for toys where the material thickness is X_1 = 4 and the material density is X_2 = 3.8, if the company uses a 95% confidence level?

11. What is the margin of error in the estimate made in question 10?
12. We want a prediction interval for the breaking strength for a toy having the material thickness and density specified in the question. What is this interval at the 95% confidence level?

1) $SSR = SST - SSE = 1743.281 - 442.292 = 1291.989$

2) $R^2 = \frac{SSR}{SST} = \frac{1291.989}{1743.281} = 0.741 = 74\%$

3) $y = \beta_0 + \beta_1 x + \beta_2 x + \beta_i x$

1) $H_0: \beta_1 = \beta_2 = \beta_3 = 0$

Where $\beta_1, \beta_2, \beta_3$ are the coefficients of the 3 explanatory variables.

H_1: At least one $\beta_i \neq 0$ for $i = 1, 2, 3$.

2) $F = \frac{MSR}{MSE} = \frac{\frac{SSR}{3}}{\frac{SSE}{n-1-3}} = \frac{\frac{1291.989}{3}}{\frac{442.292}{18}} = \frac{430.663}{24.57} = 17.52$

4) $\alpha = 0.05$

$\Rightarrow F = 3.1599$

~~F~~ $F > F_{table}$ (critical value) $\Rightarrow$ reject H_0 Keep H_1

5) H_0 is rejected $\Rightarrow$ the overall of the 3 explanatory variables have a significant effect on the level of anxiety.

4)

	Sum of squares	df
Regression due to X_1	981.326	1
Residual SSE	761.955	$n-2=20$
total	1743.281	$n-1=21$

↑ is always the same

Exercise 4 solution

	Sum of squares	df
Regression due to X_1	981.326	1
Regression due to X_2	190.232	1
Residual	571.723	$n-p-1 = n-2-1 = 19$
total	1743.281	21 $(n-1)$

1) $H_0 : B_1 = 0$ $\quad H_1 : B_1 \neq 0$

$H_0 : B_1 = B_2 = 0$ $\quad H_1 : B_i \neq 0$ for $i = 1, 2$

$H_0 : B_1 = B_2 = B_3 = 0$ $\quad H_1 : B_i \neq 0$ for $i = 1, 2, 3$

2) $f_1 = \frac{MSR}{MSE} = \frac{\frac{SSR}{3}}{\frac{SSE}{18}} = 17.52$

$f_2 = \frac{MSR}{MSE} = \frac{\frac{SSR}{2}}{\frac{SSE}{19}} = \frac{585.779}{30.0906} = 19.46$

$f_3 = \frac{MSR}{MSE} = 17.52$

3) $f_{1\,crit} = 4.3512 < f_1$ $\quad \Longrightarrow$ there is a relationship

$f_{2\,crit} = 3.5219 < f_2$

$f_{3\,crit} = 3.1599 < f_3$

6) $R_1^2 = \frac{SSR}{SST} = \frac{981.326}{1743.281} = 0.56 = 56\%$

$R_2^2 = \frac{SSR}{SST} = \frac{1171.558}{1743.281} = 0.67 = 67\%$

$R_3^2 = \frac{SSR}{SST} = 74\%$

7) The last Model seems to be more suitable $\Rightarrow SSR \nearrow \rightarrow SSE \searrow$