Statistics II — Exercises for Chapter 5, 2010/11

Exercise 1. Consider the four datasets provided in the transparencies for Chapter 5 (section 5.1)

(a) Check that all four datasets generate exactly the same LS linear regression equation.

(b) Apply to dataset # 1 the methods for detecting the presence of specification error discussed in class and comment the results.

(c) Apply to dataset # 2 the methods for detecting the presence of specification error discussed in class and comment the results.

(d) Apply to dataset # 3 the methods for detecting the presence of specification error discussed in class and comment the results.

(e) Find the outlier in dataset # 3. Obtain the LS regression line after eliminating this data point and comment on the results.

Exercise 2. Using a sample of 30 observations, the following linear regression model was estimated $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$, con $\widehat{\beta}_0 = 10.1$ y $\widehat{\beta}_1 = 8.4$. The sum of squared deviations to the mean for the response variable, due to the model, is $\sum_i (\widehat{y}_i - \bar{y})^2 = 128$, while the sum of squared residuals is $\sum_i e_i^2 = 286$.

(a) Compute the Coefficient of Determination and interpret it.

(b) What can you say regarding the correlation coefficient between the values x_i and y_i?

(c) Build the corresponding ANOVA table using these data.

(d) Test, at the 5% significance level, the hypothesis that the response variable y does not depend on x. Repeat this test at the 1% significance level.

(e) Provide an unbiased estimator for the variance of the error term.

Exercise 3. The manager of a car dealership is interested in finding the statistical relationship between the number of salesmen employed during weekends and the corresponding total number of cars sold. The following data were obtained over the course of 6 consecutive weekends:

	x_i (**# of salesmen**)	y_i (**# of cars sold**)
1	5	22
2	7	20
3	4	15
4	2	9
5	4	17
6	8	25

(a) Find the LS regression line of y (# of cars sold) over x (# of salesmen).

(b) Build the corresponding ANOVA and check the validity of the decomposition TSS = ESS + RSS.

(c) Compute the coefficient of determination and interpret it.

(d) Use the ANOVA table to test, at the 1% and 5% significance levels, the hypothesis that the number of salesmen employed during the weekend does not affect the corresponding total number of cars sold.

(e) Test the same hypotheses as in question (d), but using the procedure studied in Chapter 4. Check that the corresponding test statistic T and the test statistic F you used on question (d) satisfy the equality $F = T^2$.

Exercise 4. Using the data transformations seen in class, linearize the following nonlinear relationships:

(a) $y = \ln(5\sqrt{x})$.

(b) $y = \frac{2}{3}8^x$.

(c) $y = 1/(4-x)$.

(d) $y = \frac{5}{4}\sqrt{x}$.

Exercise 5. Assume we have obtained the following measurements for a response-variable y as a function of the explanatory variable x:

x_i	y_i
1	5.47
2	7.54
3	9.13
4	10.47
5	11.65
6	12.72

(a) Draw the scatterplot (x_i, y_i). Do you think a linear relationship can adequately describe this dataset?

(b) Assuming an adequate model is of the form $y = ax^b + u$, apply the correct transformations to the variables x and y, and obtain point estimates of the parameters a y b from a LS linear regression between the transformed variables.

(c) Build the ANOVA table for the transformed variables, obtain and interpret the corresponding coefficient of determination.

Exercise 6. For dataset # 1 from Exercise 1, obtain the LS estimators for the linear regression coefficients using vector-matrix notation.

Exercise 7. Using $n = 34$ observations the following multiple linear regression model was estimated $\widehat{y} = 2.50 + 6.8x_1 + 6.9x_2 - 7.2x_3$. The standard errors of the regression coefficient estimates for the explanatory variables are as follows $s(\widehat{\beta}_1) = 3.1$, $s(\widehat{\beta}_2) = 3.7$ y $s(\widehat{\beta}_3) = 3.2$. The corresponding coefficient of determination is $R^2 = 0.85$.

(a) Obtain confidence intervals at the 95% level for the regression coefficients corresponding to each explanatory variable.

(b) For each explanatory variable, test at the 5% significance level, the hypothesis that the response does not depend on that variable.

(c) Is there evidence, at the 1% significance level, that for each explanatory variable the corresponding regression coefficient is positive?

Exercise 8. Assume you have obtained estimates of the regression coefficients for the following model $y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u_i$. Test at the 5% significance level, the null hypotheses that the response does not depend on each one of the explanatory variables using the following partial ANOVA tables:

(a)

Source of variation	SS	D.F.	Mean	F ratio
Model	4500	3		
Residuals	500	26		
Total				

(b)

Source of variation	SS	D.F.	Mean	F ratio
Model	9780	6		
Residuals	2100	32		
Total				

(c)

Source of variation	SS	D.F.	Mean	F ratio
Model	460000	8		
Residuals	25000	27		
Total				

Exercise 9. For 10 single-family houses we obtained the price (in M€), the size of the built area (in m^2), the size of the surrounding terrain (in Has.), and the number of bathrooms

price (M€)	built area (m^2)	terrain (Has.)	# bathrooms
170	120,90	0,10	1
177	134,85	0,12	1,5
191	148,80	0,12	2
194	172,05	0,18	2
202	195,30	0,16	2
210	186,00	0,16	2,5
214	195,30	0,20	2
228	223,20	0,20	2,5
240	251,10	0,20	2,5
252	241,80	0,28	3

Below we give the Statgraphics output for the linear multiple regression of y (price) over x_1 (built area), x_2 (size of terrain), y x_3 (# of bathrooms).

(a) Obtain 95% confidence intervals for the regression coefficients of the linear model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$.

(b) Test at the 5% and 10% significance levels the hypotheses that the response variable does not depend on x_j, para $j = 1, 2, 3$.

(c) Obtain and interpret the value of the coefficient of determination. Obtain an estimate of the standard deviation of the error term.

```
Multiple Regression Analysis
----------------------------------------------------------------------------
Dependent variable: precio
----------------------------------------------------------------------------
                                    Standard          T
Parameter               Estimate      Error       Statistic        P-Value
```

CONSTANT	100,985	7,86246	12,844	0,0000
superfi	0,354243	0,0975193	3,63255	0,0109
superfTerreno	109,115	73,4594	1,48537	0,1880
WCs	10,3945	6,86311	1,51454	0,1807

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	6158,96	3	2052,99	73,92	0,0000
Residual	166,635	6	27,7726		
Total (Corr.)	6325,6	9			