University of Applied Sciences Cologne
Faculty of Economics, Business Administration and Law
Prof. Dr. Arrenberg
Room 221, Tel. 39 14
jutta.arrenberg@fh-koeln.de

Exercises Quantitative Methods

Worksheet: Linear Regression

Exercise 6.1 (Cortinhas, C.; Black, K.; page 20)
As most people suspect, the United States is number one consumer of oil in the world, followed by China, Japan, India, Russian Federation, Germany, South Korea and Canada. China however, is the world's largest consumer of coal, with the United States coming in second, followed by India, Japan and Russian Federation. The annual oil consumption figures for eight of the top total energy-consuming nations in the world, according to figures released by the *BP Statistical Review of World Energy* for the year 2010, are as follows:

Country	Oil Consumption (millions tons)	Coal Consumption (millions tons oil equivalent)
United States	850.0	524.6
China	428.6	1 713.5
Japan	201.6	123.7
India	155.5	277.6
Russian Federation	147.6	93.8
Germany	115.1	76.5
South Korea	105.6	76.0
Canada	102.3	23.4

Is there a way to graphically display oil and coal figures together so that readers can visually compare countries on their consumption of the two different energy sources?

Reference: Cortinhas, Carlos; Black, Ken: Statistics for Business and Economics, Wiley & Sons, 2012

Exercise 6.2 (Simple Linear Regression, Anderson et. al., page 556)

As part of a study on transportation safety, the US Department of Transportation collected data on the percentage of licenced drivers under the age of 21 in a city and the number of fatal accidents per 1000 licences over a one-year period in the city in a sample of 42 cities. Data collected follow.

Percentage under 21	Fatal accidents per 1000 licences	Percentage under 21	Fatal accidents per 1000 licences
13	2.962	17	4.100
12	0.708	8	2.190
8	0.885	16	3.623
12	1.652	15	2.623
11	2.091	9	0.835
17	2.627	8	0.820
18	3.830	14	2.890
8	0.368	8	1.267
13	1.142	15	3.224
8	0.645	10	1.014
9	1.028	10	0.493
16	2.801	14	1.443
12	1.405	18	3.614
9	1.433	10	1.926
10	0.039	14	1.643
9	0.338	16	2.943
11	1.849	12	1.913
12	2.246	15	2.814
14	2.855	13	2.634
14	2.352	9	0.926
11	1.294	17	3.256

Fatal_Accidents.sav

a) Develop a graphical summary of the data.

b) Use regression analysis to investigate the relationship between the percentage of drivers under the age of 21 and the number of fatal accidents.

c) What conclusion and recommendations can you derive from your analysis?

Reference: Anderson, Sweeney, Williams, Freeman, Shoesmith: Statistics for Business and Economics, Thompson Learning, London, 2007

Exercise 6.3 (Multiple Linear Regression, Anderson et. al., page 564)
As an illustration of multiple regression analysis we will consider a problem faced by Eurodistributor Company, an independent company in the Netherlands. A major portion of Eurodistributor's business involves deliveries throughout its local area. To develop better work schedules, the managers want to estimate the total daily travel time for their drivers.

Initially the managers believed that the total daily travel time would be closely related to the distance travelled in making the daily deliveries. The number of deliveries could also contribute to the total travel time. A simple random sample of ten driving assignments provided the data shown in the following table:

Driving assignment	x_1=Distance travelled (kilometres)	x_2=Number of deliveries	y=Travel time (hours)
1	100	4	9.30
2	50	3	4.80
3	100	4	8.90
4	100	2	6.50
5	50	2	4.20
6	80	2	6.20
7	75	3	7.40
8	65	4	6.0
9	90	3	7.60
10	90	2	6.10

Eurodistributor.sav

a) Consider the model with only the distance as an independent variable and comment the regression coefficient.

b) Consider the model that includes the number of deliveries as a second independent variable and comment the regression coefficients.

Exercise 6.4
Costs are high when the salesman makes a call on his clients. You have to check whether the visits affect the sales. Please consider a linear relationship between the variable Y=sales and the variables X_1=the total of the annual visits, X_2=the price of sale, X_3=the number of the outdoor staff, X_4=the costs for sales promotion:

$$\text{Linear Model: } Y \approx b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + b_4 \cdot x_4$$

Open the file *sales37.sav*
The table shows the values of the variables Y, X_1, X_2, X_3, X_4 of 37 cases.

a) Please calculate the multiple linear correlation coefficient. What is the interpretation of the received value?

b) Check the acceptability of the model with a test.

   1. What are the assumptions of this test?
   2. Check whether the assumptions of this test are violated or not.

c) How can we check whether each of the variables X_1=the total of the annual visits, X_2=the price of sale, X_3=the number of the outdoor staff, X_4=the costs for sales promotion affects the sales or not?

d) Do we have heteroscedasticity or homoscedasticity? Why?

Exercise 6.5 (Anderson et. al., page 568)
The owner of Toulon Theatres would like to estimate weekly gross revenue as a function of advertising expenditures. Historical data for a sample of eight weeks follow:

Weekly gross revenue (€ 000s)	Television advertising (€ 000s)	Newspaper advertising (€ 000s)
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.3
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

Advertising.sav

a) Develop a linear regression model with the amount of television advertising as the independent variable.

b) Develop a linear regression model with both the television advertising and newspaper advertising as the independent variables.

c) Is the estimated regression coefficient for television advertising expenditures the same in part a) and in part b)? Comment the coefficient in each case.

d) What is the estimate of the weekly gross revenue for a week when € 3500 is spent on television advertising and € 1800 is spent on newspaper advertising?

e) Is the estimated value in part d) reliable?

University of Applied Sciences Cologne
Faculty of Economics, Business Administration and Law
Prof. Dr. Arrenberg
Room 221, Tel. 39 14
jutta.arrenberg@fh-koeln.de

Quantitative Methods

Summary: Multiple Linear Regression

Multiple Linear Regression Model
$Y \approx b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + \ldots b_p \cdot x_p$

Use: To predict values

Statistical Inference:

1. The regression coefficients $b_0, b_1, b_2, \ldots, b_p$ are computed by the method of the least squares.
2. Multiple correlation coefficient $R \in [0; 1]$ weak or medium $\Rightarrow$ bad model
3. Correlation coefficient of Bravais-Pearson $r \in [-1; 1]$ of two variables only
4. Heteroscedasticity $\Leftrightarrow$ V-formation in the scatter plot between predicted values and residuals $\Rightarrow b_0, b_1, b_2, \ldots b_p$ are incorrect
5. Multicollinearity $\Leftrightarrow$ VIF $\geq 10 \Rightarrow b_0, b_1, b_2, \ldots b_p$ are incorrect
6. Extrapolation $\Leftrightarrow x \notin [x_{min}; x_{max}] \Rightarrow$ predicted value is not reliable
7. Interpolation $\Leftrightarrow x \in [x_{min}; x_{max}]$
8. Interpolation and strong correlation $R \Rightarrow$ predicted value is reliable
9. Leverage value $\Rightarrow$ low or high impact of a single point to the shape of the regression line
10. R_a^2 is increasing indicates that the additional independent variable should be added to the model
11. Residuals = observed value minus predicted value
12. In the chart „Histogram“ the bell curve fits the blocks $\Rightarrow$ the residuals have normal distribution
13. In the chart „Plot-Point Diagram“ the points are closed to the 45-degree line $\Rightarrow$ the residuals have normal distribution

14. p-value of $X_i \leq 0.05 \Leftrightarrow X_i$ has a significant influence on Y
15. p-value ANOVA $\leq 0.05 \Leftrightarrow$ the model has sense