Chair of Decision Sciences and Systems
TUM School of Computation, Information and Technology
Technical University of Munich

Business Analytics & Machine Learning
Homework sheet 2: Regression

Prof. Dr. Martin Bichler, Prof. Dr. Jalal Etesami
Julius Durmann, Markus Ewert, Johannes Knörr, Yutong Chao

Exercise H2.1 *Retail shop*

The following table displays customer demand for a retail shop.

t	0	1	2	3	4	5	6	7	8
Demand	28.20	37.65	47.28	59.76	73.44	86.19	100.31	112.58	121.63

Note: You can use Python to solve this exercise. Consider using the provided notebook as a template.

a) For the time series above, calculate the forecasted demand value for $t = 10$ using the simple linear regression and the formula below:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 \cdot t.$$

b) Calculate the RMSE and explain its meaning.

c) For the time series above, calculate the forecasted demand value for $t = 10$, assuming a biannual seasonal component of the following form: Starting from the first period $t = 0$, suppose after every second period a new year begins. Make use of the formula below:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 \cdot t + \hat{\beta}_2 \cdot Q_1.$$

d) Does the data reflect biannual data?

Exercise H2.2 *OLS implementation*

In this exercise, you will implement your own function for solving OLS regression problems in Python. The function takes the data samples in matrix-form (X, y) as inputs and returns the minimizing solution β as well as the remaining error $\mathcal{L}(\beta)$. You may want to use the provided Notebook as a template.

a) Implement the function. Use the provided template to get started.

b) For our provided toy data set (*ols-implementation-data.csv*), find the optimal regression parameters with the help of your implementation. Don't forget to add a variable for the intercept parameter!

c) Repeat b) with the aid of scikit-learn `LinearRegression` and verify your solution.

d) How much of the total variance can you explain with your model? Compute the R^2 measure. What happens if you forget about the intercept? How does the R^2 measure compare?

e) The computed R^2 value is not very good (even with the intercept). What could be the reason?

Exercise H2.3 *Determinants of Wages Data*

This exercise performs regression on the CPS1988 data set [1].
Note: Use Python and statsmodels to solve this exercise. Have a look at the provided template notebook.

a) Load the data set from the provided file (*CPS1988.csv*). Briefly describe the data set:
   i) Name the dependent variable and the independent variables.
   ii) Which scales of measurement do the variables belong to (e.g., nominal, ordinal, interval or ratio)?
   iii) Does the data set consist of cross-sectional, time-series or panel data?

b) Plot the dependent variable against each independent variable and transform the variables if necessary.
   i) Which transformations would you carry out and why?
   ii) Estimate the following model:

$$\begin{aligned}\ln(\widehat{\text{wage}}_i) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{education}_i + \hat{\beta}_2 \cdot \text{ethnicity}_i + \hat{\beta}_3 \cdot \text{experience}_i \\ + \hat{\beta}_4 \cdot \text{experience}_i^2.\end{aligned} \tag{MR1}$$

c) Interpret the model from above (Equation MR1):
   i) Which variables are statistically significant?
   ii) Is the entire model statistically significant?
   iii) What is the explanatory power of the model and why?
   iv) Interpret each regression coefficient.

d) Now consider the following alternative model:

$$\begin{aligned}\ln(\widehat{\text{wage}}_i) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{education}_i + \hat{\beta}_2 \cdot \text{ethnicity}_i + \hat{\beta}_3 \cdot \text{education}_i \cdot \text{ethnicity}_i \\ + \hat{\beta}_4 \cdot \text{experience}_i + \hat{\beta}_5 \cdot \text{experience}_i^2.\end{aligned} \tag{MR2}$$

   What is the difference between both models from above (Equation MR1 and Equation MR2)?

e) Repeat c) with the model from Equation MR2.

—

[1] Christian Kleiber and Achim Zeileis. *Applied Econometrics with R*. ISBN 978-0-387-77316-2. New York: Springer-Verlag, 2008. URL: https://CRAN.R-project.org/package=AER.