Computational Statistics
Summer 2023
Exercise Sheet 3

Fabio Miranda - Fabio.MalcherMiranda@hpi.de
Hendrik Raetz - Hendrik.Raetz@hpi.de
Pauline Hiort - Pauline.Hiort@hpi.de

May 15th 2023

This week's exercise focuses on linear regression, model diagnostics and experimental design. Problem 12 is a mathematical proof of the relationship between the R^2 statistic and correlation.

Problem 12 - RSS (paper & pen)
Recall from the lecture that R^2 is a measurement of the proportion of variance in Y that can be explained by a linear regression on X. It is defined by $R^2 = 1 - \frac{RSS}{TSS}$ with $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ being the *residual sum of squares* and $TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2$ being the *total sum of squares*.
In case of a simple linear regression of Y onto X, the R^2 statistic is equal to the square of the correlation between X and Y. Prove that this is the case.

Math tricks: Let a and b be constants and X, Y and Z be random variables.

$$Cov(a + X, cY) = c * Cov(X, Y)$$

$$Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$$

$$Var(X + a) = Var(X)$$

$$Var(aX) = a^2 Var(x)$$

Problem 13 - Regression
Barren Wuffett is seeking new investment opportunities. He collected some data from start-ups and wants to know how their business strategies (spendings in marketing, administration, and research and development) influence their profit. Help him find a model to predict profit by using a simple linear regression and the mentioned variables. **Hint:** in R use the commands `lm`, `predict`, `cor`.

A Download and parse the corresponding data set *ex03_startups.csv* from Moodle. Explore the pairwise correlations in the data and build 4 linear models in which you use the variables *Marketing.Spend*, *Administration*, *R.D.Spend* or all of them to predict *profit*.

B Examine your prediction results visually and by computing the correlation coefficients between the predicted and observed data. Which model performs best?[1] Based on the best model, what is the increase in profit when you increase the investments in R.D. by $100?

C Discuss the performance of the different models in light of the pairwise correlation of the variables in the data set.

Problem 14 - Regression by Linear Algebra
Revisit the startups data from the last problem and check if there is bug in the `lm` function in R. Compute the β coefficients by using matrix algebra in R or Python/numpy and compare them to the `lm` solution.

Problem 15 - Regression Diagnostics and Transformation
Your company caught a severe virus that prohibits accurate payments for this month because the salary data is now locked. All you have is an incomplete table that some intern started which only shows how many years a person has worked at the company and the salary they previously received based on their gender (gasp!). HR is desperate and asks you to build a model that predicts the salary based on seniority (years in the company) and gender.

Hint:
Gender is a categorical variable (also called factors or dummy variables), this influences the interpretation of the intercept of the linear model. With a 2-level (male and female) variable, we only need a single column to encode the gender, e.g. 0 for male and 1 for female. In R, either use the `factor` command on the gender column or use 'c(gender)' in the formula notation to correctly encode the categorical variable.

[1] Note: This is a simplified approach for model comparison, we will cover proper statistical methods in the next lectures.

A Download the corresponding data set *ex03_salary.csv* from Moodle. Load the data into R / Python and fit a linear model. Write down the linear model equation and interpret the coefficient estimates. Simplify the equations for both female and male employees in the form of $Y = \beta_0 + \beta_1 x_1$, with β_1 being the coefficient for seniority.

B Plot the data with regression line. Discuss visually the goodness of fit.

C Add a residual plot and a Q-Q plot as further diagnostic. Discuss your findings with regard to the plot in B.

D After evaluating the residual plot, describe the need for a transformation and decide for a transformation of the response variable to improve your model.

E Repeat your diagnostic plots and add a plot that shows the initial salary data and your new regression line (**Hint:** you have to use the inverse transformation on the predictions). Are we still looking at a linear model?

Problem 16 - Non-parametric Regression

Analyze the EHEC infection time series with a Nadaraya-Watson estimator with a Gaussian kernel and various bandwidths and compare it to a standard linear regression model. We aim at predicting the number of infections (second column) by the week of the year (first column).

A Download the data set *ex03_ehec.csv* from Moodle and import it into R/Python. Make a suitable plot and briefly describe the data. Run a linear regression model with an intercept and include it in the plot. Describe the quality of the least squares fit.

B Run a Nadaraya-Watson estimate with a gaussian kernel function (Recommended R function: `ksmooth`) with the bandwidths $h_1 = 1$, $h_2 = 2$, $h_3 = 5$, and $h_4 = 10$. Plot the original data and the resulting curves in a single plot and briefly describe the goodness of fit.

C From the model you created in part A, compare AIC and BIC values. Why would you want to take the lower values of AIC/BIC?

D In order to compare the prediction performance, run a leave-one-out cross validation. This means iterate through all observations and always exclude the observation from a single week from the data set and use all remaining observations to fit a curve and to obtain an estimate for the number of infections in the missing week.
Compute and compare the mean squared error (MSE) over all missing data points for all remaining bandwidths and the linear regression estimate. Which estimator performs best? Discuss and compare the MSE

and the visual fits of the estimators. Which line represents the infection development best?

These exciting real world problems are not graded, however, we believe these will enrich your experience in the Computational Statistics course and therefore, it is strongly recommended that you attempt to complete these problems. We will discuss the problems on Wednesday, **May 24, 2023**. Each student is *required* to present at least one solution during the semester. To ensure an interactive exercise experience solutions will be presented each week by either volunteers or randomly generated volunteers via R.