**The Elements of Statistical Learning, WS 2019/2020**
Jilles Vreeken and Tobias Marschall
Exercise Sheet #2: *Regression 1*

**Deadline:** Thursday, November 14. 2019, 10:00 a.m.
Please read and adhere to the following requirements to generate a valid submission. This problem set is worth 50 points. You may submit your solutions in groups of two students. The solutions to the theoretical problems should be submitted either digitally (in PDF format) to esl-ta@mpi-inf.mpg.de or as a hard copy before the lecture. **Label your hard copy submissions with your name(s), Matriculation number, and Exercise group.**
Solutions to programming problems, including resulting plots and answers to the questions, need to be submitted in digital format (PDF). For the programming problems you have to submit an R script (that means it needs to run with `Rscript your_file.R`). Write comments within the script to explain what your code is supposed to do.
For digital submissions the subject line of your email should have the following format:

[SL][problem set 2][*day of ex. group*]_lastname1_firstname1_lastname2_firstname2

where *day of ex. group* is either “Mo” or “We” depending on your assigned time slot. Please include the numbers of the problems you submitted solutions to (both digitally and analogously) in the emailâs body. **Please make sure that all the files are attached to the email.** The attached files should only include an executable version of your code as .R file and one .pdf file with all the other solutions.

**Problem 1** (T, 10 Points). (Exercise 3.3 in ESL, cf. ESL, Section 3.3.2, p.51)
Consider all estimates $\tilde{\theta}$ of the linear combination of the parameters $\theta = a^T\beta$ that are unbiased, i.e.

$$\mathbb{E}\left(\tilde{\theta}\right) = \theta.$$

Prove the **Gauss-Markov theorem**: The least squares estimate $\hat{\theta} = a^T\hat{\beta}$ has variance no bigger than that of any other linear unbiased estimate of $\theta$ that has the form $\tilde{\theta} = \mathbf{c}^T\mathbf{y}$. Is it also the *best* (in terms of test error) linear unbiased estimate (argue with the bias-variance tradeoff)?

*Hint:* Consider $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ and the least squares estimator $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. Assume an arbitrary linear estimator $\tilde{\theta} = \mathbf{c}^T\mathbf{y}$ is unbiased for parameter $\theta = a^T\beta$ and calculate its variance: $Var(\tilde{\theta}) = Var(\hat{\theta} + (\tilde{\theta} - \hat{\theta}))$.

**Problem 2** (T, 10 Points). The $R^2$ statistic is a common measure of model fit corresponding to the fraction of variance in the data that is explained by the model. In general, $R^2$ is given by the formula

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

Show that for univariate regression, $R^2 = Cor(X, Y)^2$ holds.
*Bonus:* Show that in the univariate case, $R^2 = Cor(Y, \hat{Y})^2$ holds.

**Problem 3** (T, 10 Points). In this exercise we will investigate the so-called curse of dimensionality.

1. (2P, Exercise 4.7.4 in ISLR)
   When the number of features $p$ is large, there tends to be a deterioration in the performance of KNN and other local approaches that perform prediction using only observations that are near the test observation for which a prediction must be made. This phenomenon is known as the *curse of dimensionality*, and it ties into the fact that non-parametric approaches often perform poorly when $p$ is large.

   (a) Suppose that we have a set of observations, each with measurements on $p = 1$ feature, $X$. We assume that $X$ is uniformly distributed in [0,1]. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only observations that are within 10% of the range of $X$ closest to that test observation. For instance, in order to predict the response for a test observation with $X$=0.6, we will use observations in the range [0.55, 0.65]. On average, what fraction of the available observations will we use to make the prediction?

   (b) Now suppose that we have a set of observations, each with measurements on $p$=2 features, $X_1$ and $X_2$. We assume that $(X_1, X_2)$ are uniformly distributed on [0,1]×[0,1]. We wish to predict a test observation's response using only observations that are within 10% of the range of $X_1$ and $X_2$ closest to that test observation. On average, what fraction of the available oversations will we use to make the prediciton?

   (c) Now suppose we have a set of observations on $p$=m features. In the same scenario as before, we wish to predict a test observation's response using only observations that are within 10% of each feature's range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?

2. (8P, Exercise 2.3 in ESL)
   Consider $N$ data points uniformly distributed in a $p$-dimensional unit ball centered at the origin. Suppose we consider a nearest-neighbor estimate at the origin. Show that the median distance from the origin to the closest data point is given by the expression:

   $$d\,(p,N) = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p}$$

   What does this mean for the k-nearest neighbor algorithm?

   *Hint*: Consider that the volume of a $p$-dimensional sphere with radius $r$ is given by $V(r,p) = G(p)r^p$, with $G(p)$ a dimension-dependent constant. The probability that a point falls into a sphere of radius $r$ is proportional to the sphere's volume since the points are uniformly distributed.

**Problem 4** (P, 20 Points). The book provides a practical guide for linear regression. Go through **3.6 Lab: Linear Regression** (ISLR p. 109–119), doing this lab will make it easier to solve the following programming exercise. This exercise uses the *Auto* data set which is contained in the R package *ISLR*. Install the R package *ISLR*.

1. Create scatterplots between all the variables. Is the relationship between those variables linear? Describe the connection between the variables. (Exclude the *name* variable, which is qualitative.)

2. Detect the variables in the scatterplots that appear to be most highly correlated and anti-correlated, respectively. Justify your choice using the *cor()* function.

3. Perform simple linear regression with *mpg* as the response using the variables *cylinders, displacement, horsepower* and *year*, respectively, as features. Which predictors appear to have a statistically significant relationship to the outcome and how good are the resulting models (measured using $R^2$)?

4. Use the **lm()** function to perform a multiple linear regression with *mpg* as the response and all other variables except *name* as the predictors. Use the **summary()** function to print the results. Compare the full model to those generated in 3) in terms of their model fit. What can you observe in the different models concerning the significance of the relationship between response and individual predictors? What does the sign of the coefficient tell you about the relationship between the predictor and the response?

5. Use the **plot()** function to produce diagnostic plots of the linear regression fit. Does the residual plot suggest any non-linearity in the data? Does the residual plot suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

6. Generate three linear models that are based on all pairwise interaction terms $(X_1 X_2)$ for *cylinders*, *weight*, and *year* as well as on the non-linear transformations $\log(X)$, $\sqrt{X}$, $X^2$ for the *displacement* variable (one per linear model). Comment on your findings.