Exercise 1

Advanced Methods for Regression and Classification

October 16, 2025

General remark: Chapter 4 of the course notes could be helpful.
Load the data `College` from the package `ISLR`. This means that you first need to install the package with

```
install.packages("ISLR")
```

and then load the data with

```
data(College,package="ISLR")
```

Look at `?College` and at `str(College)` for more detailed information. Remove (if necessary) all observations which contain missings by using the command `na.omit()`.

In the following, we compute the linear regression model by `lm()` (always use an intercept). For the output you can use `coef()` to look at the estimated regression coefficients, `summary()` to obtain statistical inference, or `plot()` to obtain diagnostic plots.

1. Predict the response `Outstate` by using only `Expend` as predictor. Plot the data, and visualize the regression line (`abline()`). What do you conclude?

2. It seems that the model is somewhat biased. Can you come up with a "more appropriate" model that better follows the visually visible linear trend? Just ideas: transformation, higher order terms, outlier handling.

3. Now predict the response `Apps` by just using the binary variable `Private` as predictor. How does the regression model look like? Interpret the meaning of the regression coefficients.

4. Convert `Private` to a variable with levels ±1, and regress `Apps` on this variable. How does the regression model look like? Interpret the meaning of the regression coefficients.

For the following tasks, split the data randomly into training and test data (about 2/3 and 1/3), build the model with the training data, and evaluate the model using the RMSE as a criterion. The RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

where N is the number of observations to be considered (e.g. only training data, or only test data), y_i are the values of the response variable, and $\hat{y}_i$ are the estimated values of the response. You can report the RMSE always for the training and the test data.

5. Predict the response `Apps` by using all variables in the data frame that make sense content-wise. Inspect and interpret (as far as possible) the diagnostic plots. Are they supporting our model requirements?

6. As the regression coefficients depend on the scale of the explanatory variables, we cannot compare them directly. For a comparison, we need to scale the variables to variance 1, which can be done by `scale()`. Estimate the regression model as in 5. based on scaled variables. What can you conclude by inspecting the coefficients?

7. Compute for models 5. and 6. the RMSEs of training and test set. What do you conclude? Do the models 5. and 6. perform equally well? Do they even lead to the identical predictions? Can we see this based on the RMSEs?

8. The diagnostic plots may suggest that the response could be transformed. Repeat 5. for the log-transformed response. Is the model more appropriate? In which sense?

9. How could we identify if model 5. or 8. is performing better? Obviously, we cannot compare the RMSEs.