Fachbereich AIAS
Universität Salzburg
Ferndinand Ferber, BSc.
Univ.-Prof. Dr. Wolfgang Trutschnig

Salzburg, May 2024
SS 23/24

Exercise Sheet 07 @'Applied AI Using R'

We are working again with linear regressions. As mentioned last time, exercises like these do prepare you for participating in projects and for internships in companies.

Aufgabe 28.
Download the *indonesian nutrition dataset* from `https://ufile.io/04mcf2kp` and parse it into `R`.

1. Set up a linear regression model to predict the number of calories from the `proteins`, `fat` and `carbohydrate` variables.
2. Extract the model coefficients and pick of one of them. How can you interpret this coefficient? Can we also interpret the intercept?
3. What is the R^2 value for the model? Is it promising or discouraging?

Aufgabe 29.
How sensitive is `lm` and R^2 to outliers?

1. Proceed in the same way as in Aufgabe 27 to generate data from a linear model.
2. To each generated dataset add an outlier (x, y), rerun `lm` on the new dataset and compare it to the model estimated without the outlier. What can be observed? How much do the coefficients change, what happens to R^2?
3. Calculate the leverage of each point, in particular the leverage of the outlier.

Aufgabe 30.
Work on your own through the slides 18-24 of the lecture (Percy II) and then go through the following tasks:

1. Take the model you trained in Exercise 1 or 2 and create the *residual diagnostics plots*.
2. Visually check if your model fulfills the homoscedasticity property. Which plot(s) do you consult?
3. Verbalize what heteroscedasticity would mean for your model.
4. Visually check if your model's residuals are normally distributed. If not, what can you say about the tails of the distribution (heavy/light tailed) compared to the normal distribution?
5. Visually check if there are outliers in the training data with respect to your model. What plot(s) do you consult?

Aufgabe 31.
Install the `{textrecipes}` package. We use the same nutrition dataset as in Aufgabe 28.

1. Define the following preprocessor:

```
preproc <- recipe(calories ~ proteins + fat + carbohydrate + name,
           data = nutrition) |>
step_tokenize(name) |>
step_tokenfilter(name, min_times = 40) |>
step_tf(name, weight_scheme = "binary")
```

2. What does `step_tokenize()` do? What does `step_tokenfilter()` do? And what does `step_tf()` do? Look up the documentation.

3. Train a second linear model utilizing this preprocessor. Does it perform better with the additional information?

4. What's the interpretation of the coefficients for the new variables?

5. Not all of the new variables have a significant p-value. What does that mean?

Aufgabe 32.
Answer the following questions (with or without coding):

1. Are the coefficients in linear regression interpreted as the change in the dependent variable for a one-unit change in the independent variable, holding all other variables constant?

2. Is it appropriate to interpret p-values in linear regression as a measure of the strength of the relationship between the independent and dependent variables?

3. Is there a direct mathematical relationship between the coefficient of determination R^2 and the Pearson correlation coefficient?

4. Is it reasonable/valid to use linear regression when the dependent variable is categorical?

5. When all independent variables are centered around zero, can you interpret the intercept term of a linear model?