Data Analysis and Visualization Exercise 10

Matthias Heinig, Felix Brechtmann, Daniela Klaproth-Andrade, Julien Gagneur

Quizzes

The following quizzes will be solved orally by the students and the professor during the lecture.

1. After performing linear regression on a dataset containing variables x and y, the following model was obtained:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 \ , \ \hat{\beta}_0 = 0 \text{ and } \hat{\beta}_1 = 0$$

What is the R^2 of this model:

A. 0

B. 1

C. $-\infty$

D. $+\infty$

2. After performing linear regression on a dataset containing variables x and y, the following model was obtained:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 \ , \ \hat{\beta}_0 = 10 \text{ and } \hat{\beta}_1 = 0$$

What is the R^2 of this model:

A. 0

B. 1

C. $-\infty$

D. $+\infty$

3. What can linear regression be used for?

1 - Make predictions for future, unseen, data

2 - Quantify explained variance

3 - Model linear relationship between variables

Select all that apply.

4. What are the implications of Heteroscedascity on linear regression. Check all that are true:

1 - The fit can be suboptimal because the least squares errors give too much importance to the points with high noise

2 - The statistical tests are flawed

3 - The coefficient estimates are biased

Tutorial

The following exercises will be solved during the tutorial sessions.

Section 00 - Getting Ready

1. Make sure you have already installed and loaded the following libraries:

```
library(ggplot2)
library(data.table)
library(magrittr)
library(tidyr)
library(dplyr)
library(patchwork) # optional, makes plots nicer
library(cowplot)
```

Section 01 - Linear regression for Predicting Heights

To start, read the provided heights dataset using the following line of code (it's your own heights data):

```
heights <- fread("extdata/height.csv") %>% na.omit() %>%
  .[, sex:=as.factor(toupper(sex))]
heights
```

1. Predict each student's height, given their sex and their parents heights.

2. Check the plot of the residual vs the predicted values and the Q-Q plot of the residuals. Do these plots provide evidence against the assumptions of linear regression?

3. Let's focus on one sex. Fit a linear model for each male's height given the father's height. Then, fit another linear model for the father's height given each male's height.

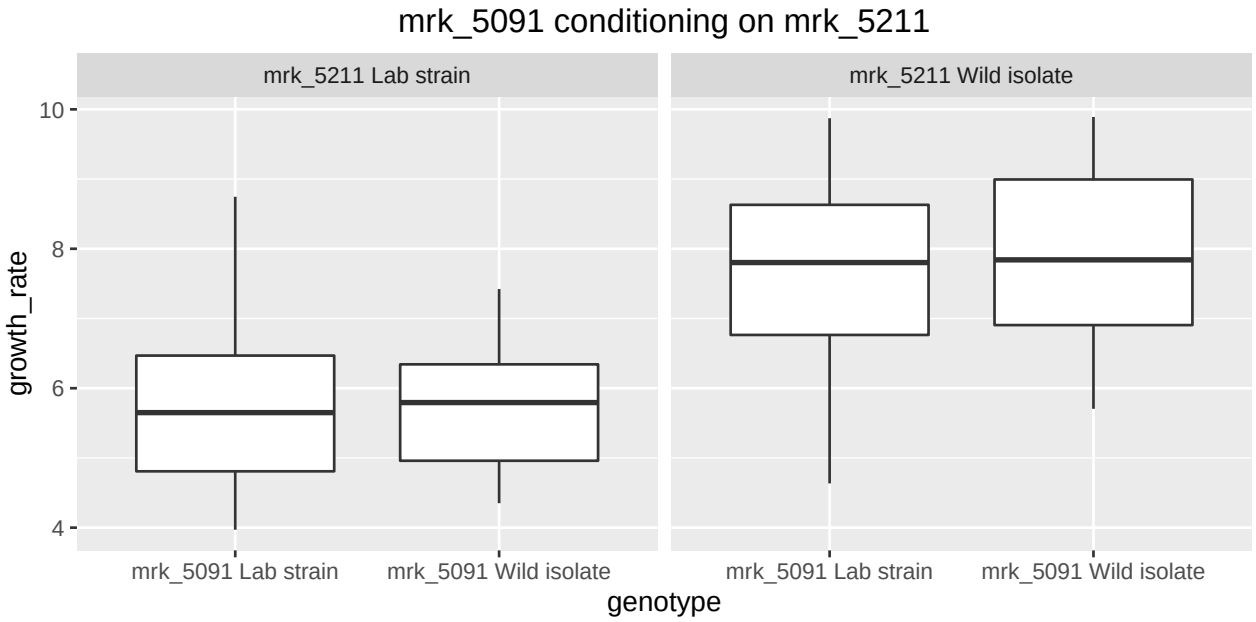
4. Predict each male student's height given the father's height (`predict()`) and predict each father's height given each male student's height. Store both predictions into new columns of a data table. Then, plot the original data and additionally both regression lines. Hint: use `geom_line`.

5. Additionally, run a PCA on the same subset of the data and plot the first principal component line into the same figure. Hint: you can get the unit vector of the first principal component by accessing the `loadings` attribute of the object obtained from `princomp`. Given this unit vector compute the slope and intercept and use `geom_abline` to plot the resulting line. Remember that the slope, m, of a line can be calculated from the coordinates (u_x, u_y) of its unit vector, $m = \frac{u_y}{u_x}$ and recall that the principal component line contains the center of the data. Use `geom_abline` to plot the line from the computed intercept and slope.

6. Interpret the plot from above. How can we explain the different slopes of the two linear models and the pca?

Section 02 - Adjusting for confounding variables - Yeast QTL

Recall the yeast QTL dataset from the previous exercises and lecture. In particular, we consider once again the question: Does marker 5091 still associate with growth in maltose when conditioned on marker 5211? Here is the plot of the data:




```
growth <- fread(file.path(eqtl_dir, "growth.txt"))
growth <- growth %>% melt(id.vars="strain", variable.name='media', value.name='growth_rate')
growth <- growth[media=="YPMalt"]

genotype <- fread(file.path(eqtl_dir, "genotype.txt"))
genotype <- genotype[, .(strain, mrk_5211, mrk_5091)]

head(genotype)
```

```
##     strain     mrk_5211     mrk_5091
## 1: seg_01B   Lab strain   Lab strain
## 2: seg_01C   Lab strain   Lab strain
## 3: seg_01D Wild isolate Wild isolate
## 4: seg_02B   Lab strain Wild isolate
## 5: seg_02C   Lab strain   Lab strain
## 6: seg_02D Wild isolate   Lab strain
```

```
head(growth)
```

```
##     strain  media growth_rate
## 1: seg_01B YPMalt    6.720447
## 2: seg_01C YPMalt    7.429273
## 3: seg_01D YPMalt    6.905589
## 4: seg_02B YPMalt    4.924324
## 5: seg_02C YPMalt    4.413402
## 6: seg_02D YPMalt    7.926200
```

1. Run a linear model predicting the growth given the genotypes of both markers and interpret the result. Call this model `full`.

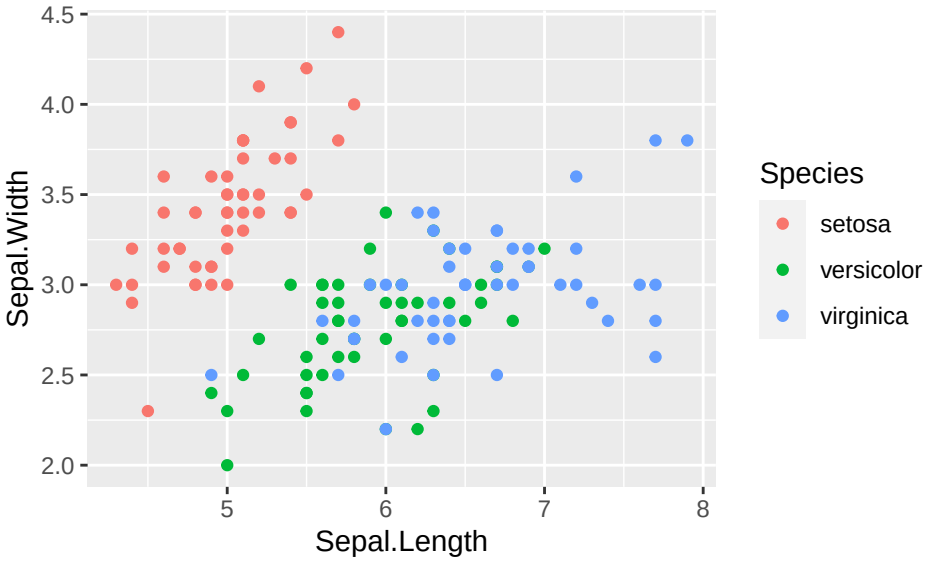
2. Create a reduced model that only depends on the genotype of `mrk_5211`. Then run ANOVA to compare the full and the reduced model. Suppose that all the assumptions of linear regression hold. What do you conclude?

Homework

Please solve the exercises below at home. The solutions will be discussed in the central exercise.

Section 03 - Adjusting for confounding variables - Iris dataset

Recall the plot showing the relationship between sepal width and sepal length in the Iris dataset:



1. Fit three linear models predicting Sepal.Width from Sepal.Length: the base model that simply predicts sepal width from sepal length, one where you use the species as a covariate in linear regression (i.e., different intercept for different species) and one where you use separate slopes and intercepts for different species by using the `*` operator in `lm`: `lm(y ~ Sepal.Length * Species)`.

2. What are the slopes and intercepts of each one of the species for the model with separate slopes and intercepts?

3. Overlay the resulting fits on the plot above (Hint: use `predict` to generate predictions)

4. Use `anova` to test if the second model is a better model than the base and also if the third model is better than the second. Suppose all the assumptions of linear regression hold.