Multiple linear regression
Multiple linear regression

Extension of the simple linear regression

Two or more covariates

No longer easily represented graphically

Model selection and over fitting
Multiple regression model

\[ y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{i_k} + \epsilon_i, \quad i = 1, \ldots, n \]

\[ \epsilon_i \sim N(0, \sigma^2) \]

\[
\begin{bmatrix}
  y_1 \\
  \vdots \\
  y_n
\end{bmatrix}
= 
\begin{bmatrix}
  x_{11} & \cdots & x_{1k} \\
  \vdots & \ddots & \vdots \\
  x_{n1} & \cdots & x_{nk}
\end{bmatrix}
\begin{bmatrix}
  \beta_1 \\
  \vdots \\
  \beta_k
\end{bmatrix}
+ 
\begin{bmatrix}
  \epsilon_1 \\
  \vdots \\
  \epsilon_n
\end{bmatrix}
\]

\[ y = X\beta + \epsilon \]
Observations (rows)

Variables (columns):
- Response $y_1, \ldots, y_n$
- Covariates (or predictors) $x_{11}, \ldots, x_{nk}$

Parameters $\beta_1, \ldots, \beta_k$

Errors $\epsilon_1, \ldots, \epsilon_n$
Assumptions

Independent and normally distributed errors

Variability of errors is constant
(regardless of covariate observations)

Linear relationship between covariates and response
(non-linear relationships later in the course)
Model selection for linear models

Using more covariates will always increase fit to data, but not to “unseen” data!

Overfit vs. underfit (bias vs. variance)

When in doubt, prefer a simpler model!

Parameter t-test can help, but not a sufficient solution

Information criteria, step-wise selection
Evaluating model size

Adjusted R-squared

- Proportion of variation explained
- Adjusts for number of variables

F-test for the full model

- Null: covariates are irrelevant
- Alternative: at least one covariate is relevant
- Adding irrelevant covariates tends to increase the p-value
Evaluating individual covariates

T-test for individual parameters

Not satisfactory:

• Troublesome e.g. when several covariates contain information about the same thing

• Adding or removing a covariate affects p-values of other covariates

However, we should avoid models with non-significant parameters
Model evaluation with ANOVA

ANOVA on the model, in R: `anova(myModel)`

- Adds covariates one by one and tests, in each step, if the decrease in RSS is significant

ANOVA on two models

In R: `anova(smallerModel, biggerModel)`

- Tests the alternative hypothesis: the bigger model has a significantly lower RSS
Example: Backward step-wise selection

1. Start with the full model (all covariates)

2. Remove the least significant covariate

3. Did the model improve? (e.g. adjusted $R^2$, F-statistic, ANOVA)
   - Yes: Continue and remove another covariate
   - No: Stop, include remaining covariates and the last removed one in the final model

Alternatively: forward step-wise selection
Some challenges

There is no one method that is always best

As usual: a combination of what the data tells us and what we know about the domain

“Best” model depends on the purpose
Next time

Non-linear models and more about linear modeling

Friday morning