Multiple Linear Regression

AI Concepts and Algorithms (SS2025)

FH University of Applied Sciences TECHNIKUM WIEN

Lecturer: Rosana de Oliveira Gomes
Author: M. Blaickner, B. Knapp, S. Rezagholi, R.O. Gomes

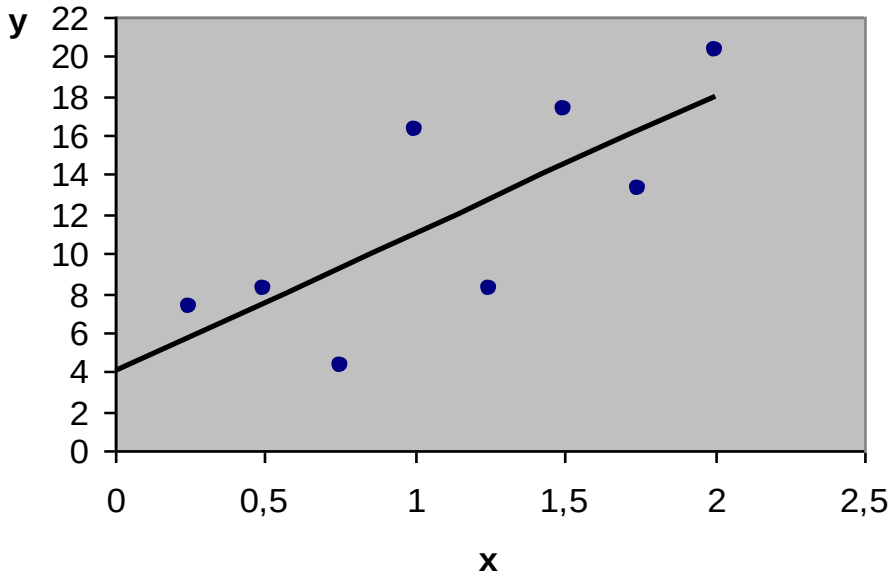
Repetition: Simple Linear Regression

y, dependent variable (observation, response)
x, independent variable (predictor)

Intercept: β_0
Slope: β_1
$y = \beta_0 + \beta_1 x$

Residual of the i^{th} observation: e_i
Residual sum of squares: RSS

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

FH University of Applied Sciences TECHNIKUM WIEN

Multiple Linear Regression (MLR)

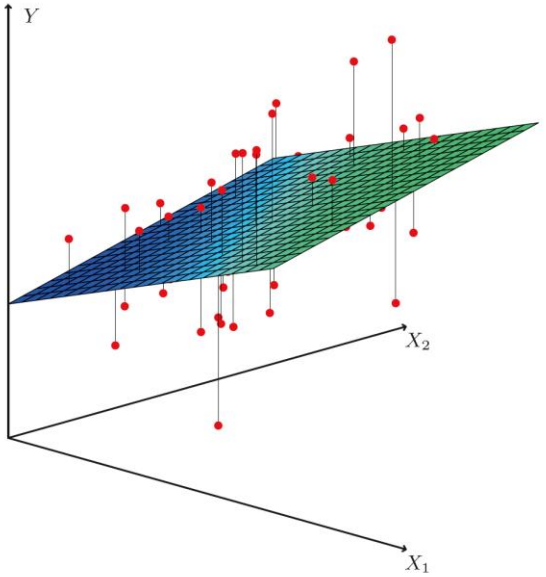
- In multiple linear regression a quantitative response **y** with **p different predictors** x_1, x_2, ..., x_p is written in the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon = \epsilon + \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

- where ε is a stochastic error term.

- The parameters β_0, β_1, ..., β_p are estimated using the least squares approach as in simple linear regression.

- In a 3-dimensional setting (2 predictors and 1 response), the least squares regression line becomes a plane that minimizes the sum of the squared vertical distances between each observation (shown in red) and the plane.

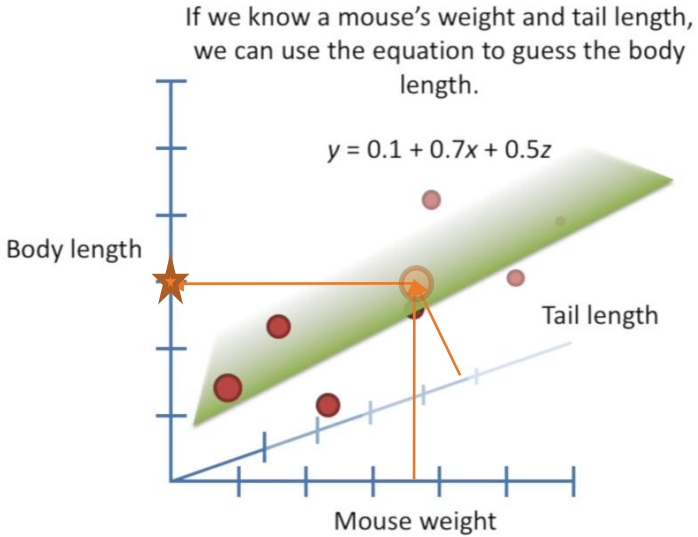


Taken from [1]

Example

2 predictors and 1 response

from https://www.youtube.com/watch?v=nk2CQITm_eo&list=PLblh5JKOoLUIzaEkCLIUxQFjPIlapw8nU&index=2



Example: Multiple Linear Regression

X
Features

	Age	Education level	Years experience	Manager of	Sick days
Kate Mayer	25	2	1	0	3
Angelo Black	37	5	15	5	0
John Smith	32	0	2	0	21
…	…	…	…	…	…

y
Target variable

Income/year
39 356
77 834
25 899
…

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

incomePerYear = β_0 + β_1 * age + β_2 * educationLevel + β_3 * yearsExperience + β_4 * managerOf + β_5 * sickDays

We are trying to find the ideal values for the parameters β_0, …, β_p such that the RSS is minimized.

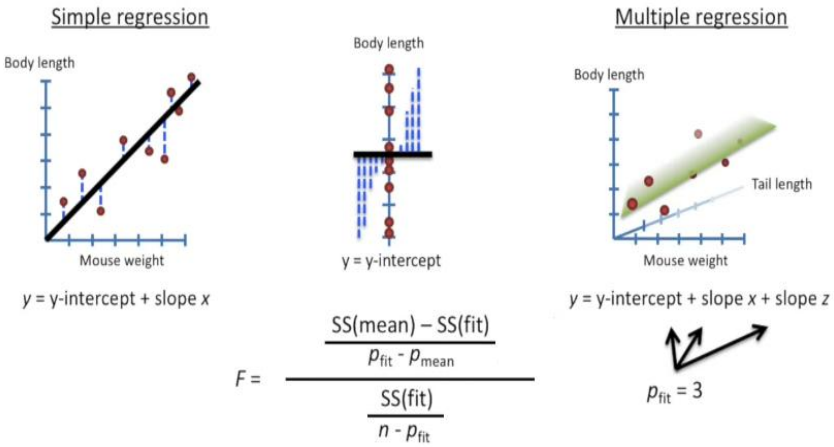
FH University of Applied Sciences TECHNIKUM WIEN

Does a linear regression with multiple dependent variables exist as well?

- Yes, it is called **multivariate multiple regression**.
- E.g. math scores and reading scores as determined by socioeconomic factors.
- Multivariate multiple regression regresses each dependent variable separately on the predictors. But hypothesis testing is more complicated, especially regarding p-values (e.g. Holm-Bonferroni method).
- No further details in this lecture.

Statistical Metrics for Multiple Linear Regression

- The same as in one-dimensional case (see slides on simple linear regression in ML1).



[From https://www.youtube.com/watch?v=zITIFTsivN8]

p-value: tests contribution of individual variables

F-score: tests model significance
(e.g. Regression assumption)

R^2: tests if model can describe the data (fit)

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \qquad \mathrm{TSS} = \sum (y_i - \bar{y})^2$$

Issues in Multiple Linear Regression

- Is at **least one** of the predictors x_1, x_2, ..., x_p useful in predicting the response?
- Do **all** the predictors help to explain Y, or only a **subset**?

E.g.: If we are trying to predict the *weight of a mouse,* which of the following predictors might be useful? *Mouse length, blood volume, paw size, tail length, color of fur, color of eyes, astrological sign*?

- Is there an **interaction** effect between explanatory variables?

Are variables correlated?

Pearson correlation coefficient:
Measurement of the linear relationship among two continuous variables

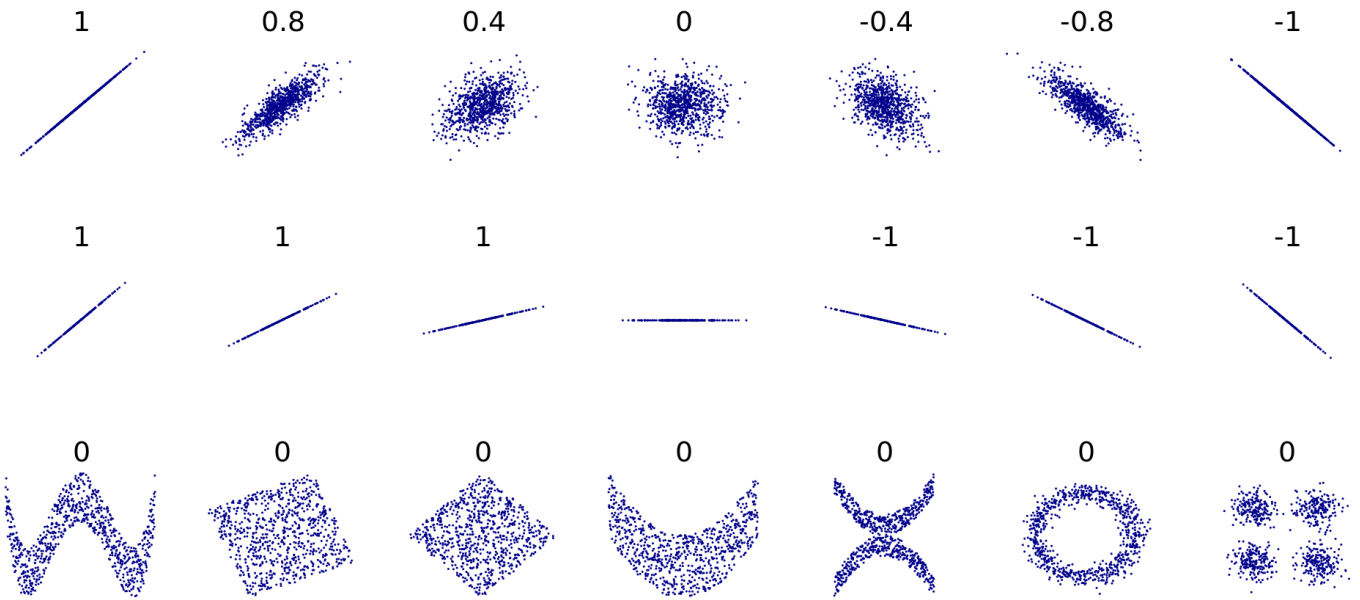
r = 1, perfect positive correlation
r = 0, no correlation
r = -1, perfect negative correlation

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

FH University of Applied Sciences TECHNIKUM WIEN

Examples: Pearson correlation coefficient



FH University of Applied Sciences TECHNIKUM WIEN

- Only sensitive to **linear** interaction!
- Quantifies degree of linear relationship, not the slope!

Remember?

OLS Regression Results

Dep. Variable:	y	R-squared:	0.900
Model:	OLS	Adj. R-squared:	0.892
Method:	Least Squares	F-statistic:	113.7
Date:	Fri, 26 Nov 2021	Prob (F-statistic):	7.57e-75
Time:	13:48:26	Log-Likelihood:	-1.6977
No. Observations:	178	AIC:	31.40
Df Residuals:	164	BIC:	75.94
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	3.4733	0.498	6.980	0.000	2.491	4.456
alcohol	-0.1170	0.037	-3.166	0.002	-0.190	-0.044
malic_acid	0.0302	0.022	1.369	0.173	-0.013	0.074
ash	-0.1486	0.103	-1.441	0.151	-0.352	0.055
alcalinity_of_ash	0.0399	0.009	4.650	0.000	0.023	0.057
magnesium	-0.0005	0.002	-0.307	0.759	-0.004	0.003
total_phenols	0.1443	0.064	2.268	0.025	0.019	0.270
flavanoids	-0.3724	0.051	-7.334	0.000	-0.473	-0.272
nonflavanoid_phenols	-0.3035	0.206	-1.473	0.143	-0.710	0.103
proanthocyanins	0.0394	0.047	0.838	0.403	-0.053	0.132
color_intensity	0.0756	0.014	5.268	0.000	0.047	0.104
hue	-0.1492	0.134	-1.116	0.266	-0.413	0.115
od280/od315_of_diluted_wines	-0.2701	0.052	-5.152	0.000	-0.374	-0.167
proline	-0.0007	0.000	-6.868	0.000	-0.001	-0.000

```
from sklearn import datasets
import statsmodels.api as sm

dataset = datasets.load_wine()
data = sm.add_constant(dataset.data)
targets = dataset.target

ols = sm.OLS(targets, data).fit()

features = dataset.feature_names
features.insert(0, 'intercept')

print(ols.summary(xname=features))
```

FH University of Applied Sciences TECHNIKUM WIEN

… now we are going to look a little bit into how p-values are calculated!

Recap: p-value

"The p-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct."

- In other words: The p-value is the probability to observe the given data by chance in the absence of any real association.
- A small p-value indicates that there is an association, which validates rejecting the null hypothesis. Typical p-value cutoffs for rejecting the null hypothesis are 0.01, 0.05, and 0.1.

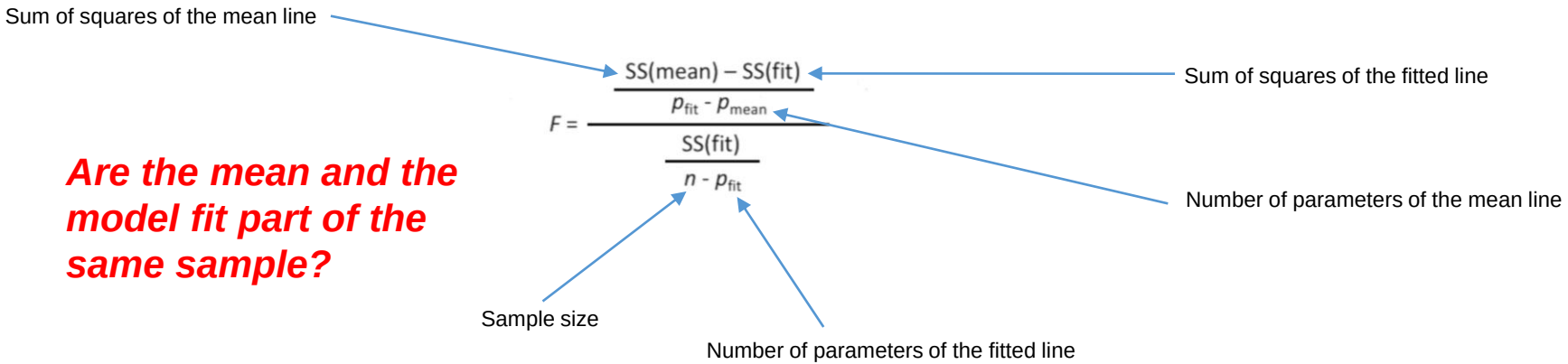
Example: Predicting house prices using variables *number of rooms , square meters, street name.*

Predictor	p-value
Number of rooms	0.02
Square meters	0.001
Street name	0.9

FH University of Applied Sciences TECHNIKUM WIEN

F-statistic

***F-test**: determines if the variances of two samples are significantly different*

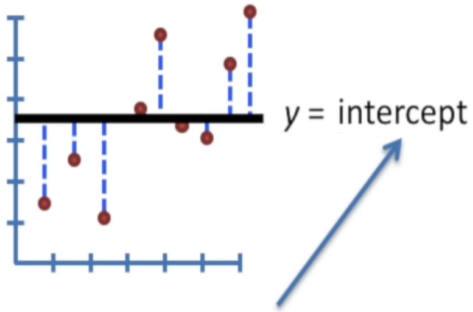


Are the mean and the model fit part of the same sample?

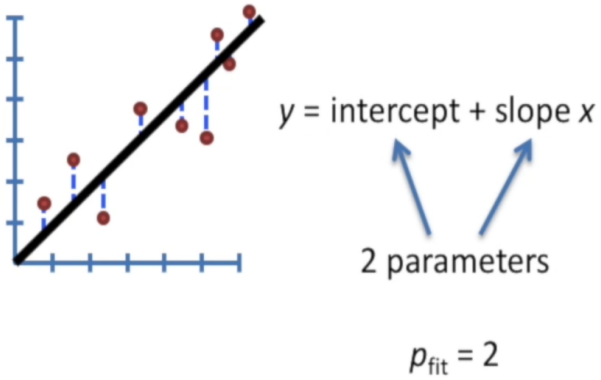
A little “punishment“ if we use more parameters

Example: if we predict someone‘s income in € based on toe size, number of neighbours during childhood, height of their grandmother, number of mice living in his backyard etc).

F-statistic and p-value



$p_{mean} = 1$



$p_{fit} = 2$

If we had used 2 features then p_{fit}=3.
Not 2.
Do not forget the intercept.

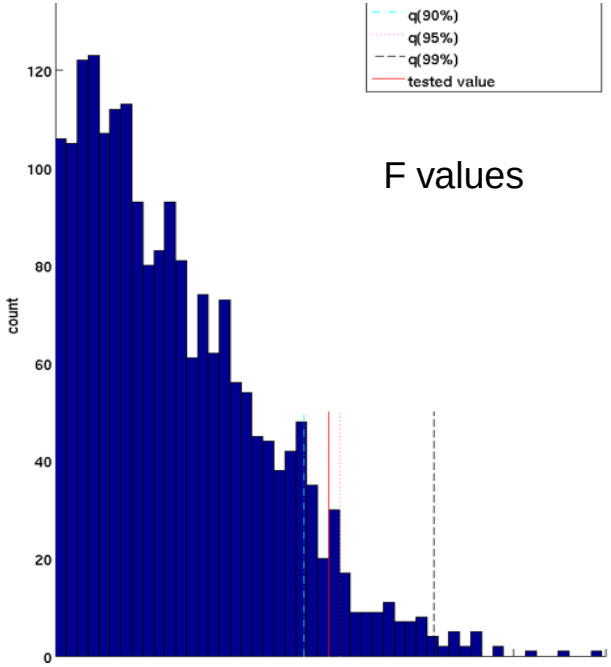
$$F = \frac{SS(mean) - SS(fit)}{\frac{SS(fit)}{n - 2}}$$

F-score and p-value

- How do we get a p-value from the F-statistic?
- Complex calculation: we look it up in a table
- But how was the table created?

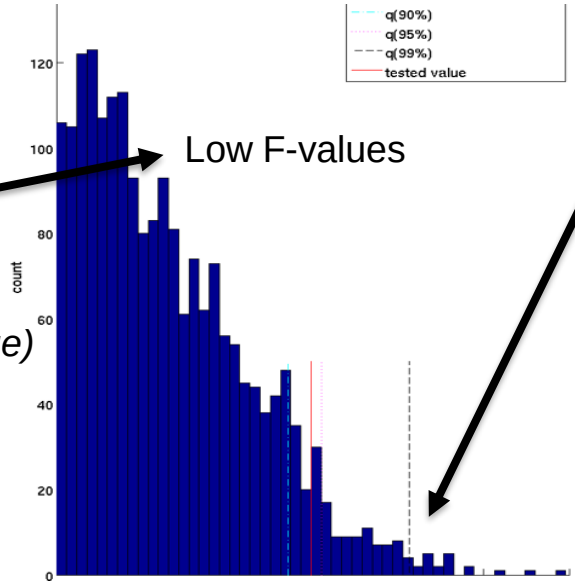
Answer:

- The distribution of the F-statistic under the null hypothesis has been derived theoretically.



F-score and p-value

Small p-values (e.g. p<0.05) indicate that the optimized parameters (e.g. slope and intercept) are significantly better than the model that always predicts the mean



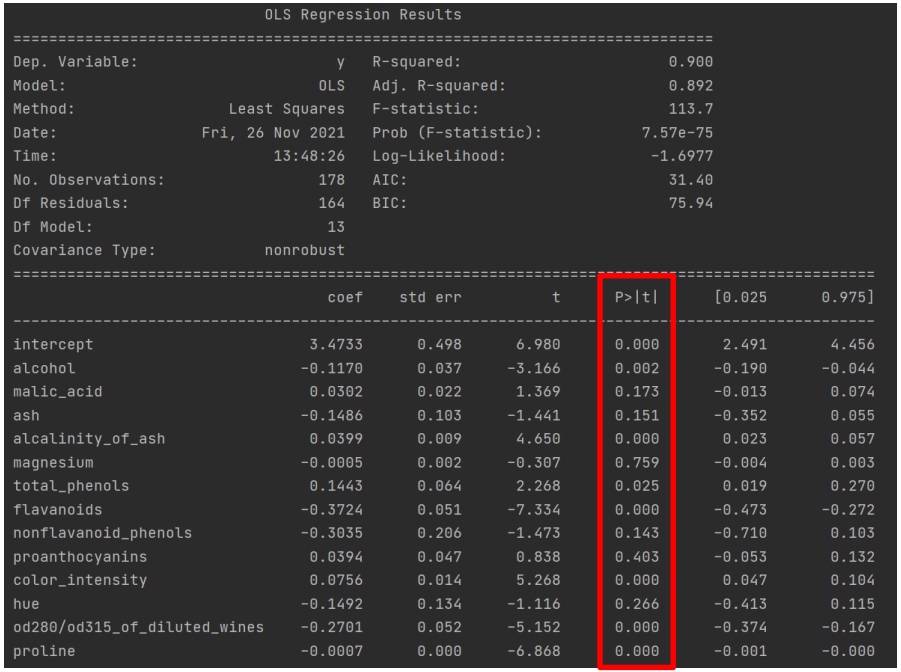
Low F-values

the model does not explain much variance (null hypothesis likely true)

Small p-values
High F-values

the explained variance is much greater than the unexplained variance (unlikely under the null hypothesis)

F values



OLS Regression Results

Dep. Variable:	y	R-squared:	0.900
Model:	OLS	Adj. R-squared:	0.892
Method:	Least Squares	F-statistic:	113.7
Date:	Fri, 26 Nov 2021	Prob (F-statistic):	7.57e-75
Time:	13:48:26	Log-Likelihood:	-1.6977
No. Observations:	178	AIC:	31.40
Df Residuals:	164	BIC:	75.94
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	3.4733	0.498	6.980	0.000	2.491	4.456
alcohol	-0.1170	0.037	-3.166	0.002	-0.190	-0.044
malic_acid	0.0302	0.022	1.369	0.173	-0.013	0.074
ash	-0.1486	0.103	-1.441	0.151	-0.352	0.055
alcalinity_of_ash	0.0399	0.009	4.650	0.000	0.023	0.057
magnesium	-0.0005	0.002	-0.307	0.759	-0.004	0.003
total_phenols	0.1443	0.064	2.268	0.025	0.019	0.270
flavanoids	-0.3724	0.051	-7.334	0.000	-0.473	-0.272
nonflavanoid_phenols	-0.3035	0.206	-1.473	0.143	-0.710	0.103
proanthocyanins	0.0394	0.047	0.838	0.403	-0.053	0.132
color_intensity	0.0756	0.014	5.268	0.000	0.047	0.104
hue	-0.1492	0.134	-1.116	0.266	-0.413	0.115
od280/od315_of_diluted_wines	-0.2701	0.052	-5.152	0.000	-0.374	-0.167
proline	-0.0007	0.000	-6.868	0.000	-0.001	-0.000

Which features are “significant“?

Testing the Null Hypothesis

- **Null hypothesis H_0**: There is **no relationship** between predictor and response.

$$H_0: \beta_1 = \beta_1 = \ldots = \beta_p = 0.$$

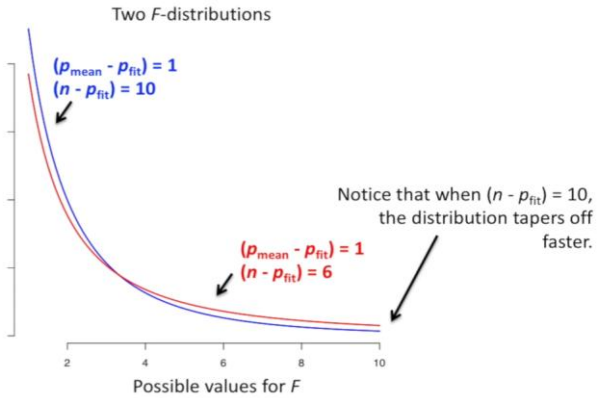
- In case of the alternative H_1 at least one β_j is unequal to zero.
- This can be tested with the **F-statistic:**

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}, \qquad \mathrm{TSS} = \sum(y_i - \bar{y})^2$$

where

- n denotes the number of observations

- p the number of predictors



F-statistic

- In case of H_0 the F-statistic has a value close to 1.
- In case of H_a the F-statistic has a value > 1.

How large must the F-statistic be to reject H_0?

- Any statistical software package can be used to compute the **p-value associated with the F-statistic**
- Based on the p-value, we can determine whether or not to reject H_0.

FH University of Applied Sciences TECHNIKUM WIEN

See the slides on simple linear regression (ML1)!

F-test vs. Individual p-values

- Consider an **MLR** model with several coefficients, **each** of which has been tested for statistical significance (t-test), leading to an associated **p-value**.
- The global test for statistical significance of the model (**global F-test**) also leads to a p-value.
- Attention: **The p-values for individual coefficients can not be interpreted globally, neither can the p-value for the global F-test be thought of as applying to all coefficients!**

In case of $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ there is only a 5% chance that the F-statistic results in a p-value below 0.05, regardless of the value of p and n. (n >> p)

Having a single significant parameter in a model with many parameters does not imply the same level of significance for the model as a whole.

- The **F-statistic** avoids this problem because it **adjusts** for the number of **parameters**

Variable Selection

- Suppose that the H_0 in an MLR model is rejected on the basis of the p-value associated with the F-statistic. A question remains:
  - Which predictors are related to the response? All of them? Just a subset?
- For p variables there are 2^p-1 nonempty subsets.

  For p=10 there are 1023 models to consider!

➢ Variable Selection:
  - Forward selection
  - Backward selection
  - Mixed selection

FH University of Applied Sciences TECHNIKUM WIEN

Forward Selection

1. Start with the null model which contains an intercept but no predictors.
2. $y = \beta_0$
3. Fit p simple linear regressions and **add** to the null model the variable that results in the lowest RSS. (Unless the last model had lower RSS. In that case the last model is the final model.)
4. $y = \beta_0 + \beta_{i1} x_{i1}$
5. Add to the new model the variable that reduces RSS the most (if possible). This yields a two-variable model.
6. $y = \beta_0 + \beta_{i1} x_{i1} + \beta_{i2} x_{i2}$
7. Continue until a stopping rule is satisfied.

Backward Selection

1. Start the MLR with all variables and **remove** the one with the largest p-value (that is the least statistically significant).
2. Fit the new (p-1)-variable model and remove the variable with the largest p-value.
3. Continue until a stopping rule is satisfied. (E.g. all remaining variables have a p-value below some threshold.)

Mixed Selection

1. Combination of forward and backward selection.
2. Start with forward selection and add the variable that provides the best fit.
3. Continue to add variables one-by-one.
4. If at any point the p-value for one of the variables in the model rises above a certain threshold, remove that variable from the model.
5. Continue these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

Note: The inclusion-threshold should be stricter than the exclusion-threshold.
Explain!

Goodness of Fit

- The most common numerical measure of model fit is R^2 (Range: (0,1); the closer to 1 the better) and the **residual standard error (RSE)** (the lower the better).
- **Caution:** Adding a new variable will **always (slightly) increase R^2**. A small increase implies that the new variable is not very significant (also check the p-value).
- One often uses **adjusted R^2** to compare models of unequal dimensionality.

$$\bar{R}^2 = 1 - \frac{\frac{\mathrm{RSS}}{n-p}}{\frac{\mathrm{TSS}}{n-1}}$$

- Note: If p=1, then adjusted R^2 equals the usual R^2.

FH University of Applied Sciences TECHNIKUM WIEN

Interaction

- Variables may not be (completely) independent of each other.
  - ➢ Example: The number of heart failures in a population could be a function of age and hypertension (high blood pressure). But with higher age the probability of hypertension rises. This is referred to as an **interaction** effect.

- Consider an MLR with two variables: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

- Introducing an **interaction term**:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

FH University of Applied Sciences TECHNIKUM WIEN

Hierarchical Principle

- What if β_3 (interaction term) has a very low p-value (seems important), but β_1 and β_2 (main effects) have not?
- The **hierarchical principle** states that if we include an **interaction** in a model, we should **also include the main effects**, even if the p-values associated with their coefficients are not significant.

Why Does This Principle Matter?

- Interaction terms depend on the individual variables removing a variable while keeping another makes interpretation difficult and the model unstable.
- If a variable in the mixed term is significant but the other one is not, removing a variables changes the meaning of the interaction term
- Omitting main effects while keeping interactions can lead to biased coefficient estimates and incorrect conclusions.

Take-aways

Statistical metrics

- **p-value**: Tests if an individual variable is useful
- **F-score**: Tests if the overall model is better than guessing
- **R-squared**: Measures how well the model explains the data

Too little data *leads to unreliable statistical tests*

Higher dimentionality

- **Too many variables**: Risk of false significance and collinearity
- **Correlation**: Can cause instability in coefficients if not handled properly

Feature selection process

Assignment: Multiple Regression

a) Explain multiple regression as pseudo-code or by visualizations

Use self-made images or even hand drawings (of which you take a photo).

Use self written explanations.

Do not copy from the lecture slides or the internet (neither text nor images).

b) Use a library implementation of multiple regression (e.g. Scikit)

Compute F-test on all variables.

Interpret the correlation matrix of features.

Implemented forward selection and backward selection by yourself and compare R^2, adjusted R^2, p-values, ...

Run your implementation on a dataset of your choice.

Most of you have jobs in IT where you might have access to interesting datasets. Otherwise you can always use something from https://scikit-learn.org/stable/datasets.html or https://www.data.gv.at. You can also create artificial datasets with sklearn.datasets.make_classification() and sklearn.datasets.make_regression().

FH University of Applied Sciences TECHNIKUM WIEN

References

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013.