Excercise 3 Increasing Well-Being with Data Analytics

TECHNISCHE UNIVERSITÄT DARMSTADT

Summer Term 2023

ise. Information Systems and Electronic Services

Agenda

TECHNISCHE UNIVERSITÄT DARMSTADT

Data Analytics and Regression Analysis

Ordinary Least Squares (OLS)

Hypothesis Testing: The t-test

What is Data Analytics?

- “a careful and complete analysis of data using a model, usually performed by a computer; information resulting from this analysis”.[1]
- Within data analytics (inferential), statistics are utilized to draw conclusions from a population sample in order to identify potential causal relationships between the independent and the dependent variables.[2]
- Data Science and Data Analytics can be considered applied branches of statistics.[3]

1 Oxford Dictionary
2 Judd, Charles and, McCleland, Gary (1989). Data Analysis. Harcourt Brace Jovanovich. ISBN 0-15-516765-0.
3 Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.

Regression Analysis

Comprehensive methodology to analyze quantitative data (usually including at least one numerical variable):

“set of statistical processes for estimating the **relationships** between a **dependent variable** and one or more **independent variables.**”

The most widely applied methodology to perform (inferential) statistical analyses

Typical data sources:

Survey, (Field) Experiment, Observational Data

Source: https://en.wikipedia.org/wiki/Regression_analysis

Agenda

TECHNISCHE UNIVERSITÄT DARMSTADT

Data Analytics and Regression Analysis

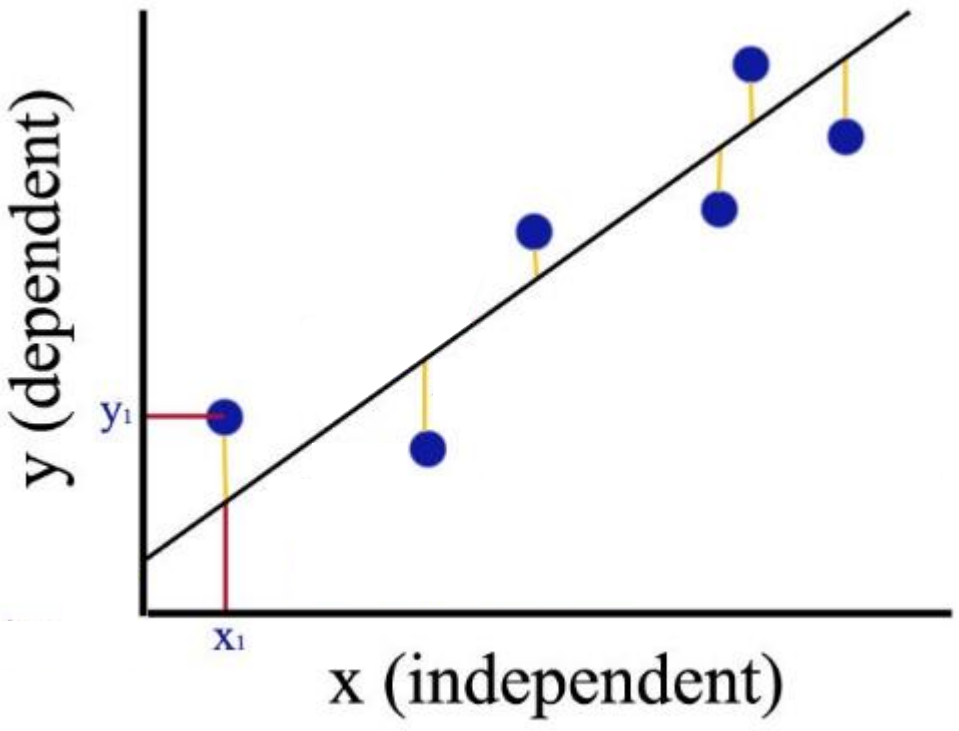
Ordinary Least Squares (OLS)

Hypothesis Testing: The t-test

Fundamental Regression Model: ordinary least squares (OLS)

TECHNISCHE UNIVERSITÄT DARMSTADT

OLS: a method for estimating the unknown parameters in a linear regression model



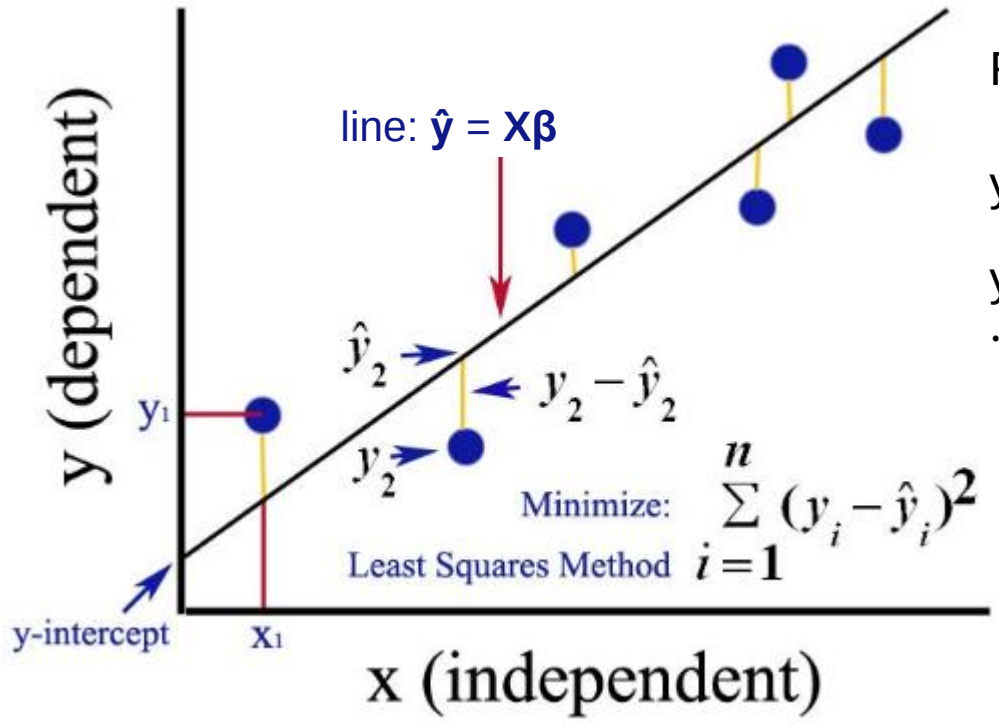
Sources:
https://medium.com/analytics-vidhya/ordinary-least-square-ols-method-for-linear-regression-ef8ca10aadfc;
https://en.wikipedia.org/wiki/Ordinary_least_squares

ise Information Systems and Electronic Services

Fundamental Regression Model: ordinary least squares (OLS)

TECHNISCHE UNIVERSITÄT DARMSTADT

OLS: a method for estimating the unknown parameters in a linear regression model



Regression Equation:

$$y = b_0 + \beta_1 * x + \varepsilon$$

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_k * x_k + \varepsilon$$

Sources:
https://medium.com/analytics-vidhya/ordinary-least-square-ols-method-for-linear-regression-ef8ca10aadfc;
https://en.wikipedia.org/wiki/Ordinary_least_squares

ise Information Systems and Electronic Services

Agenda

TECHNISCHE UNIVERSITÄT DARMSTADT

- Data Analytics and Regression Analysis
- Ordinary Least Squares (OLS)
- **Hypothesis Testing: The t-test**

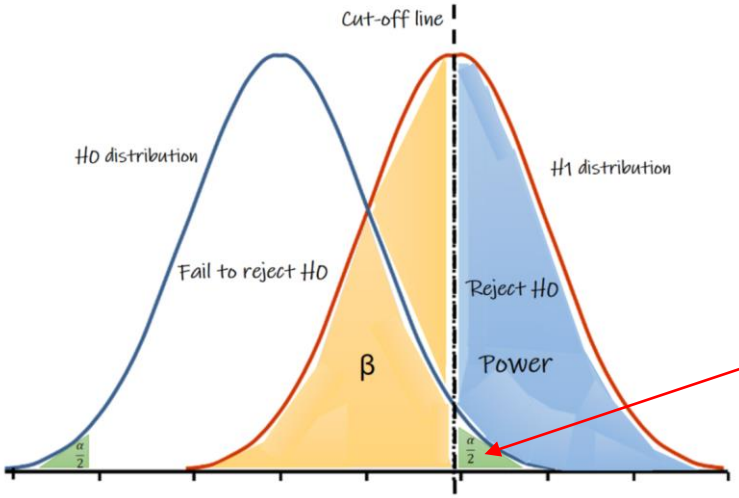
Hypothesis Testing: The t-test

To test if we have a significant independent variable, we can use hypothesis testing, such as a t-test.

To understand hypothesis testing, we have to think of two variables:

H0: Null hypothesis — the default **no-difference hypothesis** we are trying to reject.

H1: Alternative hypothesis — the difference hypothesis.



α (type 1 error)

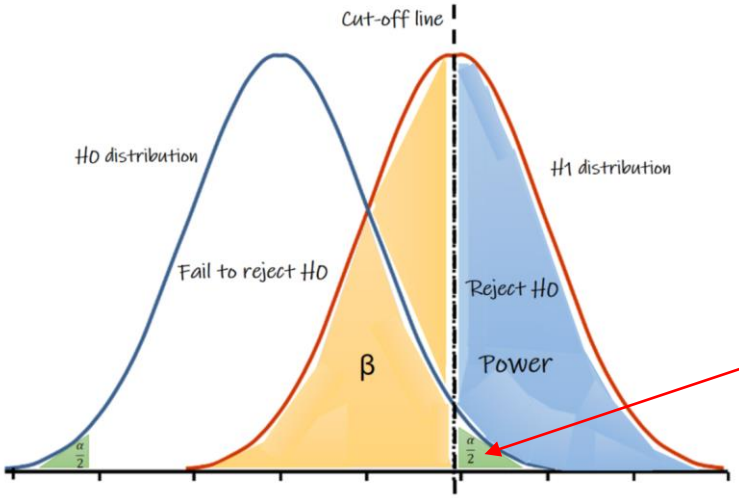
Source: https://medium.com/evidentebm/the-value-of-the-p-value-9460797a92d6

Hypothesis Testing: The t-test

The β estimators are derived from randomly distributed data; therefore, parameter values of OLS regressions can be tested.

To test if an independent variable of regression has a significant relationship with our dependent variable, we set our hypothesis as follows:

H0: Null hypothesis $\beta = 0$ | H1: Alternative hypothesis $\beta \neq 0$



α (type 1 error)

Source: https://medium.com/evidentebm/the-value-of-the-p-value-9460797a92d6

Regression: P-Value

As a result of the t-test, we receive a p-value.

The p-value and a self-defined cut-off of α (usually 0.05) indicate if the value of β is statistically significant.

If p-value<α (e.g., p-value=0.02), we reject the null hypothesis and say there is a statistically relevant difference, but if p-value≥α (e.g., p-value=0.50), we cannot reject the null hypothesis and there is no statistically relevant difference.

The p-value indicates the probability of the result occurring by chance.

Source:
https://medium.com/evidentebm/the-value-of-the-p-value-9460797a92d6

OLS-Regression output in R

```
Call:
lm(formula = WORK_LIFE_BALANCE_SCORE ~ FRUITS_VEGGIES + BMI_RANGE +
    TODO_COMPLETED + DAILY_STEPS + SUFFICIENT_INCOME + TIME_FOR_PASSION +
    WEEKLY_MEDITATION + AGE + GENDER + YEAR, data = KA_sub)

Residuals:
    Min      1Q  Median      3Q     Max
-68.492 -15.703   1.191  15.920  42.681

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       540.0369     7.5066  71.941  < 2e-16 ***
FRUITS_VEGGIES      6.1293     0.8817   6.952 1.57e-11 ***
BMI_RANGE         -12.4581     2.4538  -5.077 6.00e-07 ***
TODO_COMPLETED      4.5360     0.4677   9.699  < 2e-16 ***
DAILY_STEPS         3.0800     0.4316   7.136 4.87e-12 ***
SUFFICIENT_INCOME  30.5668     2.6603  11.490  < 2e-16 ***
TIME_FOR_PASSION    6.2116     0.4522  13.735  < 2e-16 ***
WEEKLY_MEDITATION   2.3515     0.4020   5.849 1.06e-08 ***
AGE36 to 50        -2.0451     2.7990  -0.731   0.4654
AGE51 or more      -2.3778     3.4146  -0.696   0.4866
AGELess than 20     8.9271     3.6152   2.469   0.0140 *
GENDERMale         -4.2973     2.3816  -1.804   0.0720 .
YEAR2016           -7.9251     3.9975  -1.983   0.0481 *
YEAR2017           -3.0408     4.0172  -0.757   0.4495
YEAR2018           -4.7586     4.5987  -1.035   0.3014
YEAR2019           -7.5908     4.7900  -1.585   0.1139
YEAR2020           -8.2080     4.0206  -2.041   0.0419 *
YEAR2021            4.9846     8.5954   0.580   0.5623
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.23 on 382 degrees of freedom
Multiple R-squared:  0.7783,    Adjusted R-squared:  0.7684
F-statistic: 78.89 on 17 and 382 DF,  p-value: < 2.2e-16
```

regression model

parameter estimates

p-values

Thank You!

TECHNISCHE UNIVERSITÄT DARMSTADT

ise Information Systems and Electronic Services