Technische Universität Berlin

Modeling

Usability Engineering | Group 1 Topic 7

Muni Won, Qian Long Jamy Zhou Hu, Alexander Kens | 09.07.2024

TECHNISCHE UNIVERSITÄT BERLIN

Table of Contents

Technische Universität Berlin

01 Simple Linear Regression

Terms and Definitions

- **Statistical Modeling**: use of mathematical models and statistical assumptions to make predictions
- **Linear Regression**: data analysis technique that predicts the value of unknown data using known data values
- **Simple Linear Regression**: Uses only one independent variable
- **Linear regression is used for**
  1) prediction
  2) measurement of influence

Technische Universität Berlin

01 Simple Linear Regression

Ordinary Least Squares

$$Y = \beta_0 + \beta_1 * X + \varepsilon$$

Y : Dependent variable (DV) to be predicted
X : Independent variable (IV)
Unknown coefficients:
β_0: y-intercept, the expected value of Y when X = 0
β_1 : Slope of the line
ε: the error term

Slope: $$\beta_1 = \sum_1^i \frac{(x_i - \overline{x}) * (y_i - \overline{y})}{(x_i - \overline{x})^2}$$

mean of all y values

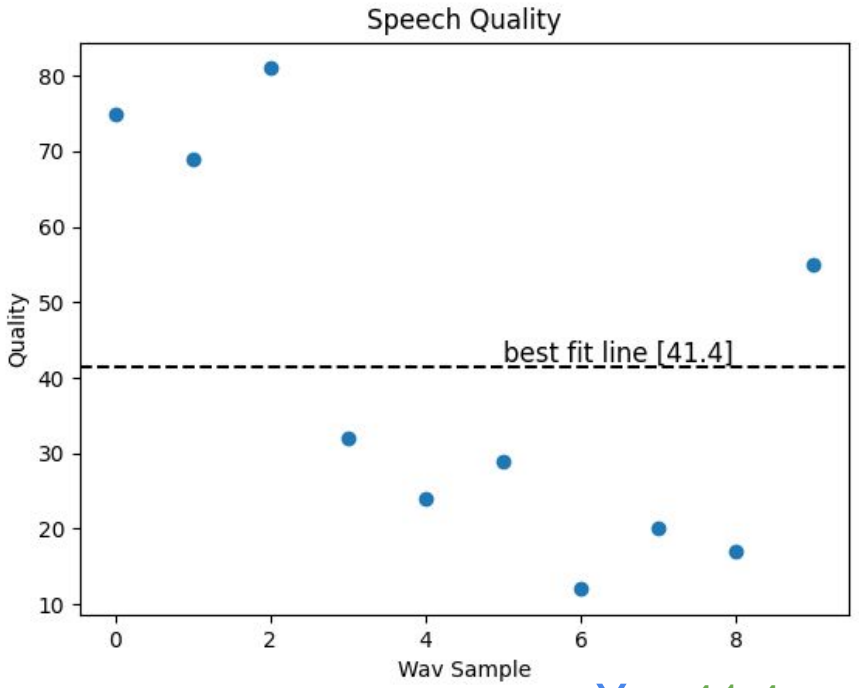
mean of all x values

Intercept: $$\beta_0 = \overline{y} - \beta_1 * \overline{x}$$

01 Simple Linear Regression

Speech Quality Example with one participant and only a Dependent Variable (DV)

Audio	Quality
wav1	75
wav2	69
wav3	81
wav4	32
wav5	24
wav6	29
wav7	12
wav8	20
wav9	17
wav10	55



Mean = 41.4

$Y = 41.4$

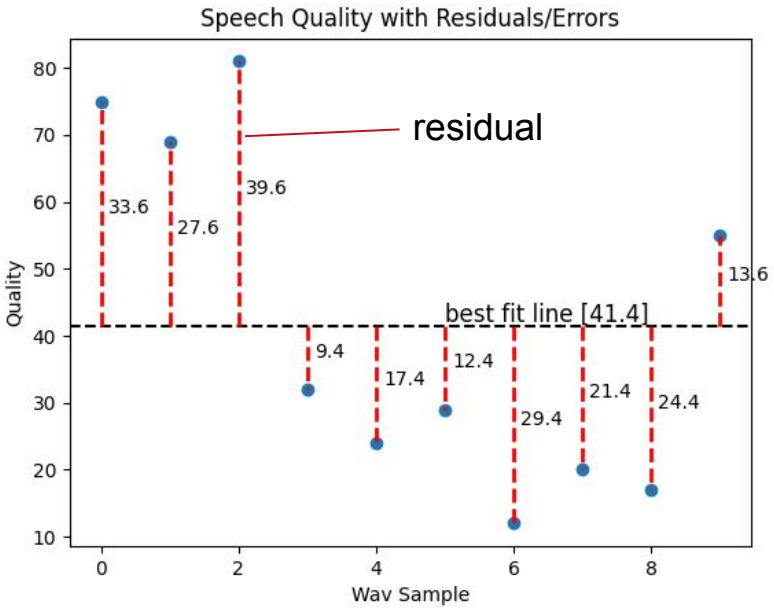
$Y = \beta_0 + \beta_1 * X + \varepsilon$

Technische Universität Berlin

01 Simple Linear Regression

Speech Quality Example with one participant and only a Dependent Variable (DV)

Audio	Quality	Residuals
wav1	75	-34
wav2	69	-28
wav3	81	-40
wav4	32	9
wav5	24	17
wav6	29	12
wav7	12	29
wav8	20	21
wav9	17	24
wav10	55	-14



Technische Universität Berlin

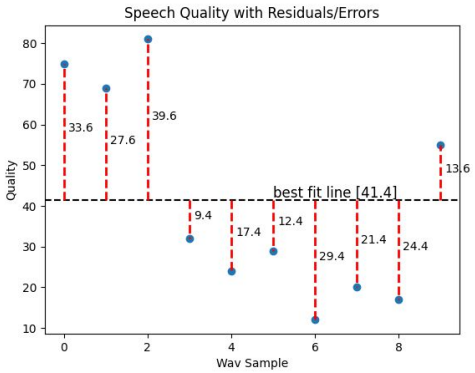
01 Simple Linear Regression

Speech Quality Example with one participant and only a Dependent Variable (DV)

Audio	Quality	Residuals	Residuals²
wav1	75	-34	1156
wav2	69	-28	784
wav3	81	-40	1600
wav4	32	9	81
wav5	24	17	289
wav6	29	12	144
wav7	12	29	841
wav8	20	21	441
wav9	17	24	576
wav10	55	-14	196

→ Σ = 6108

- The goal of linear regression is to minimize the sum of squares error (SSE)
- The regression line will "fit" the data better and minimize the residuals
- When there is only a dependent variable, then the best fit is the mean [41.4]

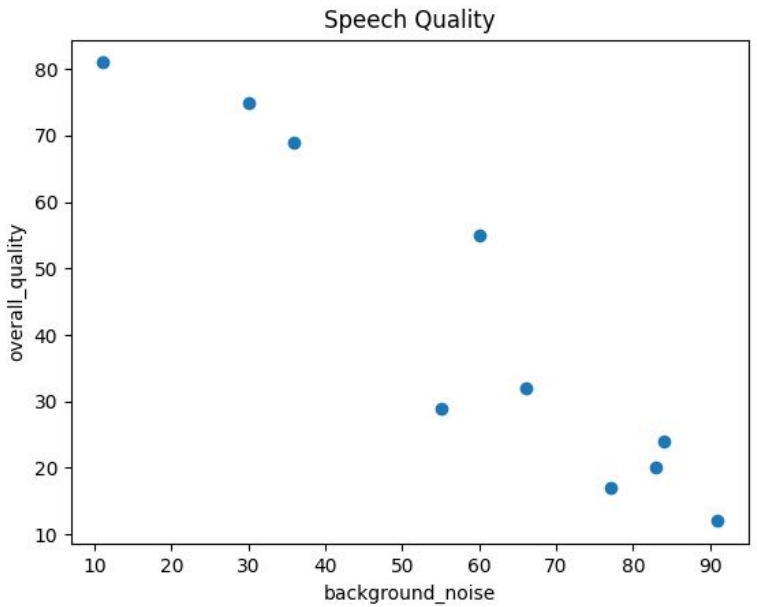


Technische Universität Berlin

01 Simple Linear Regression

Speech Quality Example with one participant and IV, DV

Audio	Quality	Noise
wav1	75	30
wav2	69	36
wav3	81	11
wav4	32	66
wav5	24	84
wav6	29	55
wav7	12	91
wav8	20	83
wav9	17	77
wav10	55	60

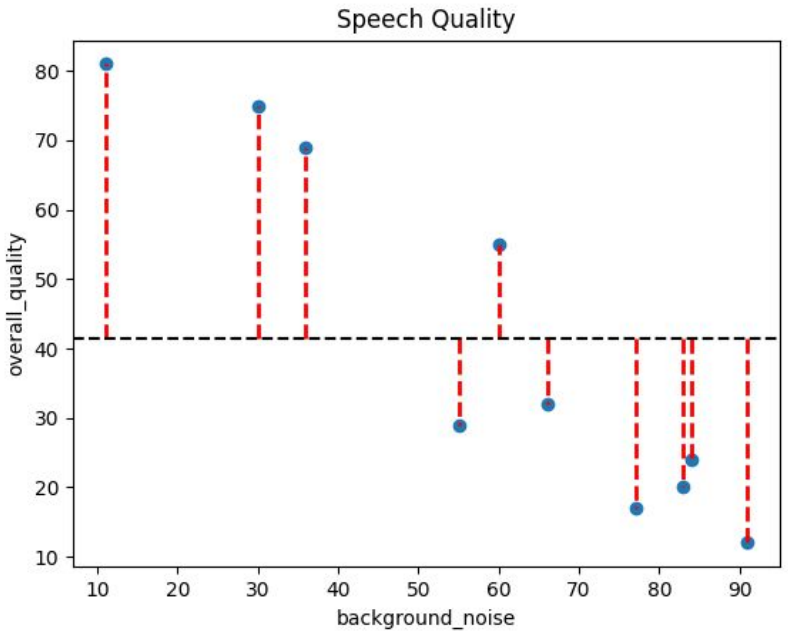


01 Simple Linear Regression

Speech Quality Example with one participant and IV, DV

Audio	Quality	Noise	Error	Error²
wav1	75	30	33.6	1128.96
wav2	69	36	27.6	761.76
wav3	81	11	39.6	1568.16
wav4	32	66	9.4	88.36
wav5	24	84	17.4	302.76
wav6	29	55	12.4	153.76
wav7	12	91	29.4	864.36
wav8	20	83	21.4	457.96
wav9	17	77	24.4	595.36
wav10	55	60	13.6	184.96

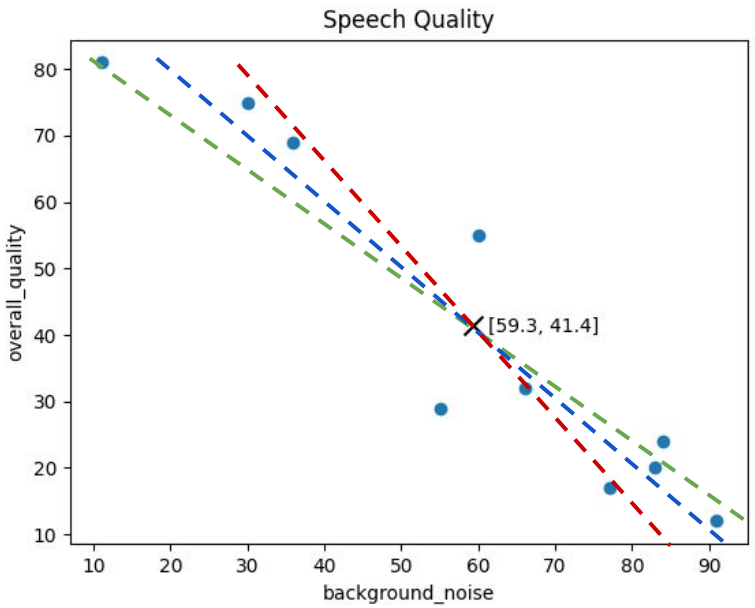
→ Σ = 6106.4



01 Simple Linear Regression

Speech Quality Example with one participant and IV, DV

Audio	Quality	Noise
wav1	75	30
wav2	69	36
wav3	81	11
wav4	32	66
wav5	24	84
wav6	29	55
wav7	12	91
wav8	20	83
wav9	17	77
wav10	55	60



X Mean = 59.3

Y Mean = 41.4

Technische Universität Berlin

01 Simple Linear Regression

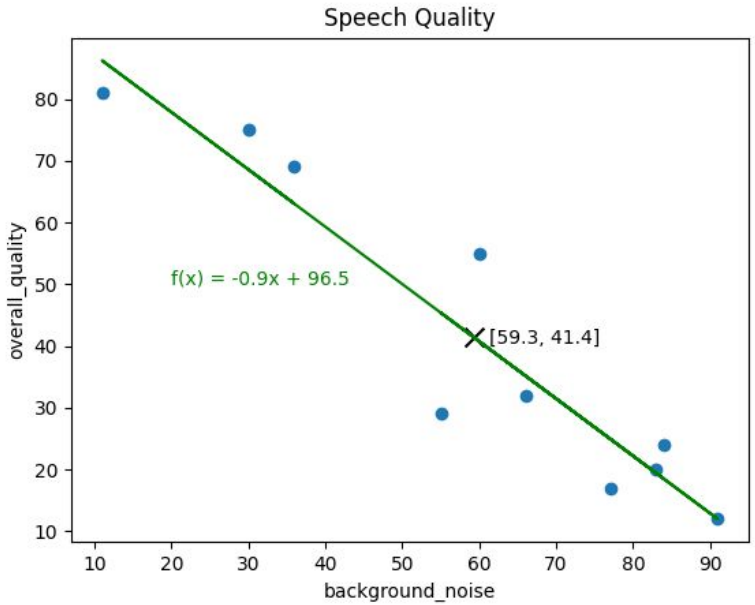
Ordinary Least Squares

Slope: $$\beta_1 = \sum_1^i \frac{(x_i - \overline{x}) * (y_i - \overline{y})}{(x_i - \overline{x})^2}$$

Intercept: $$\beta_0 = \overline{y} - \beta_1 * \overline{x}$$

$Y = -0.9 * X + 96.5$

$Y = \beta_0 + \beta_1 * X + \varepsilon$



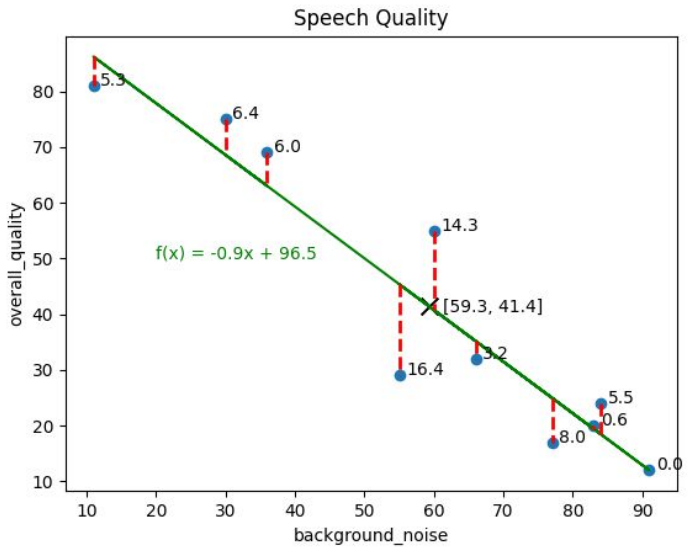
Technische Universität Berlin

01 Simple Linear Regression

Speech Quality Example with one participant and IV, DV

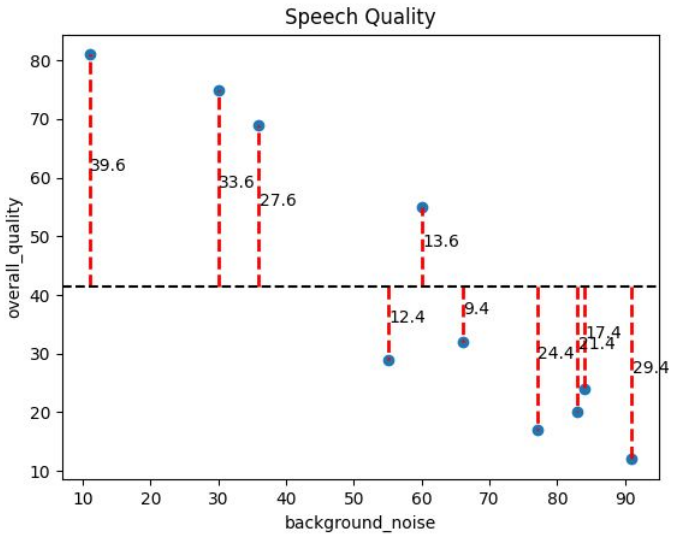
Audio	Quality	Noise	Error	Error²
wav1	75	30	-6.38	40.7
wav2	69	36	-5.96	35.52
wav3	81	11	5.27	27.77
wav4	32	66	3.18	10.11
wav5	24	84	-5.54	30.69
wav6	29	55	16.39	268.63
wav7	12	91	-0.05	0.0025
wav8	20	83	-0.62	0.38
wav9	17	77	7.96	63.36
wav10	55	60	-14.25	203.06

→ Σ = 680.22

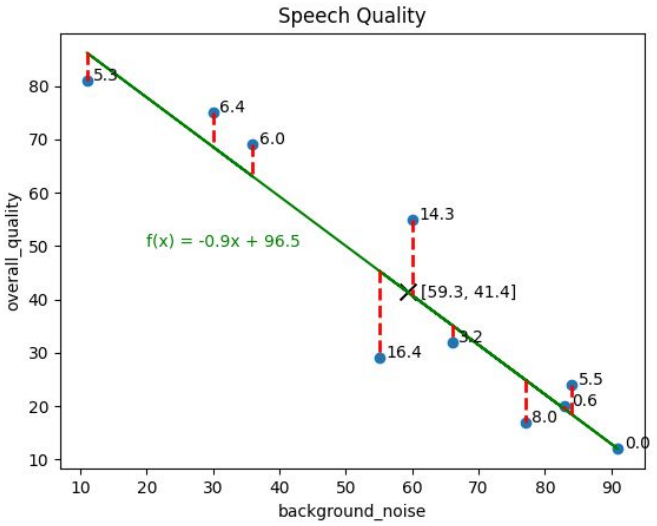


01 Simple Linear Regression

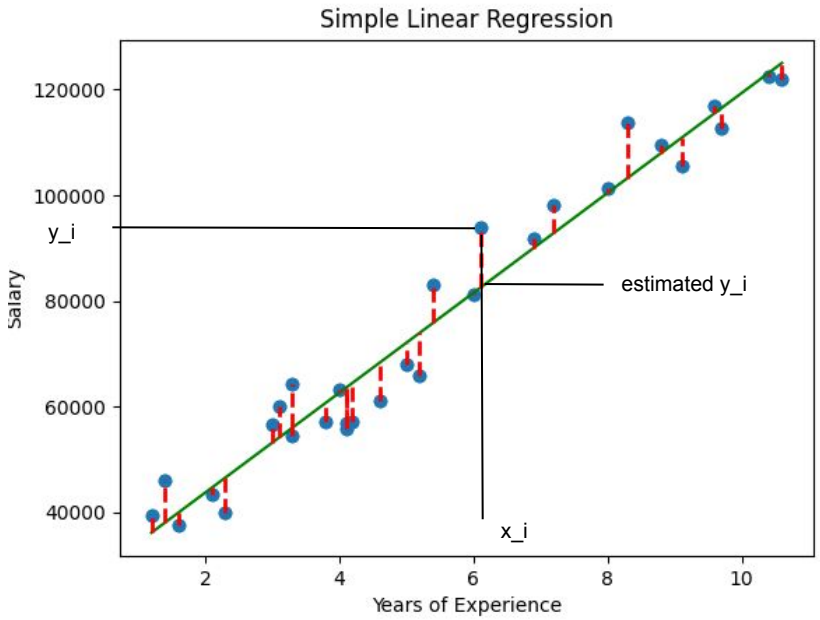
Comparison of Mean Line and OLS

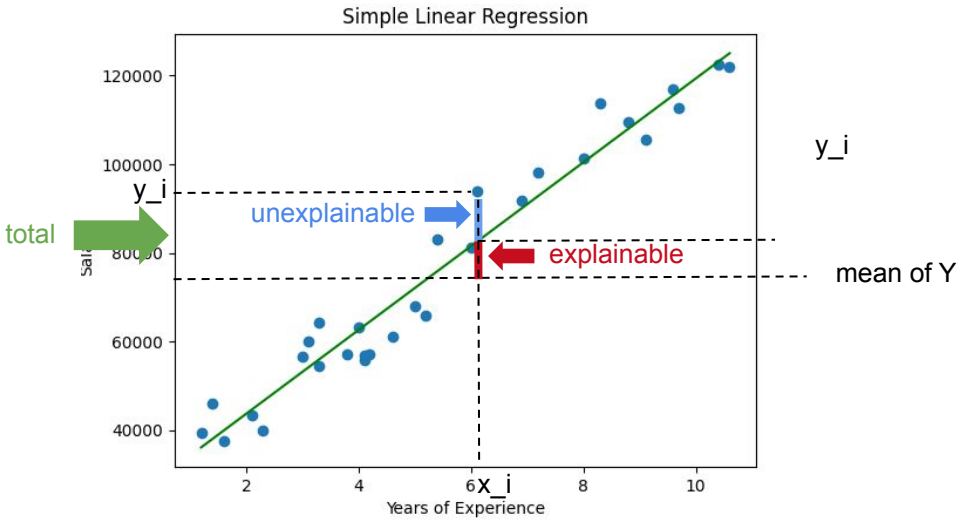


→ Σ = 6106.4



→ Σ = 680.22


Simple Linear Regression
120000
100000
80000
60000
40000
Salary
y_i
estimated y_i
x_i
2
4
6
8
10
Years of Experience


Simple Linear Regression
120000
100000
y_i
60000
40000
2
4
6
x_i
8
10
Years of Experience
y_i
unexplainable
explainable
mean of Y
total

01 Simple Linear Regression

Key assumptions and statistical significance

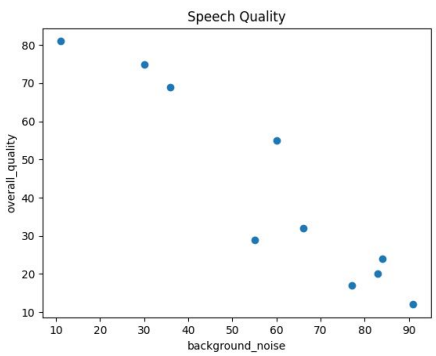
1. **Independence**: Observations are independent of each other.
2. **Linearity**: The relationship between X and the mean of Y is linear.
3. **Homoscedasticity (No Outliers)**: The variance of residual is the same for any value of X.
4. **Normality**: Distribution of the residuals should be approximately normal.

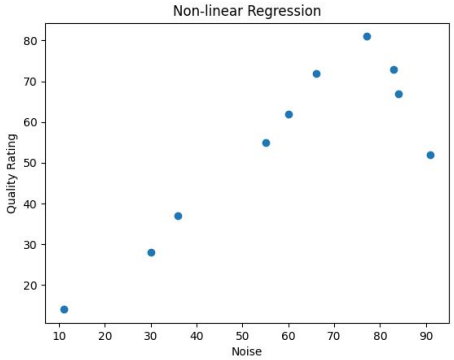
Technische Universität Berlin

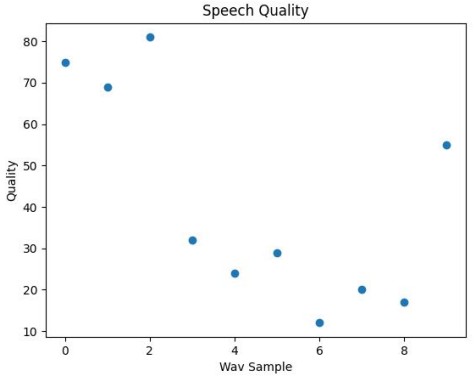
01 Simple Linear Regression

Key assumptions and statistical significance

2. Linearity: The relationship between X and the mean of Y is linear.





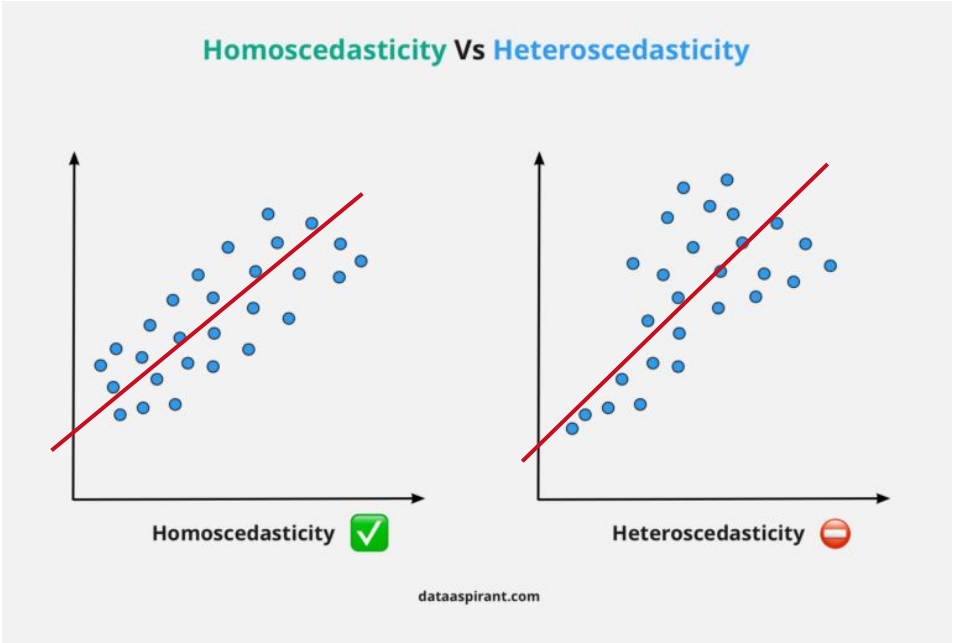


Technische Universität Berlin

01 Simple Linear Regression

Key assumptions and statistical significance

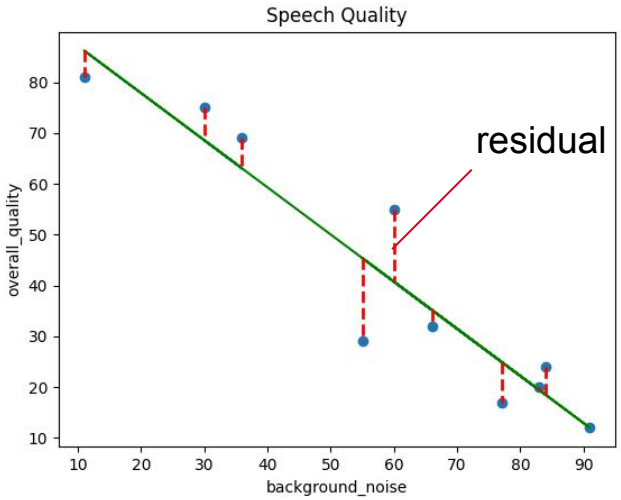
3. **Homoscedasticity**: The variance of residual is the same for any value of X.

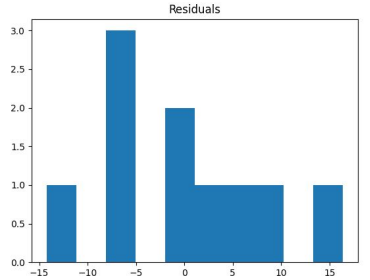


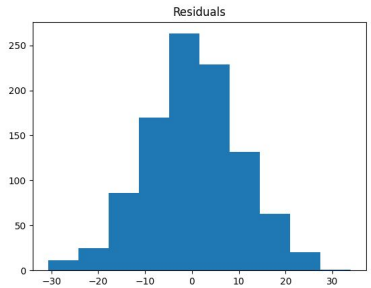
01 Simple Linear Regression

Key assumptions and statistical significance

4. Normality: Distribution of the residuals should be approximately normal.





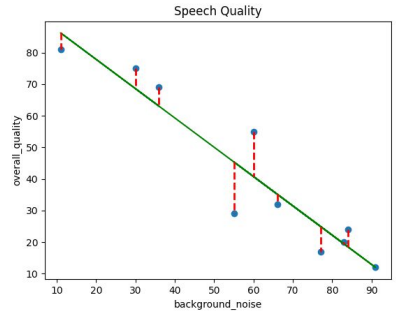


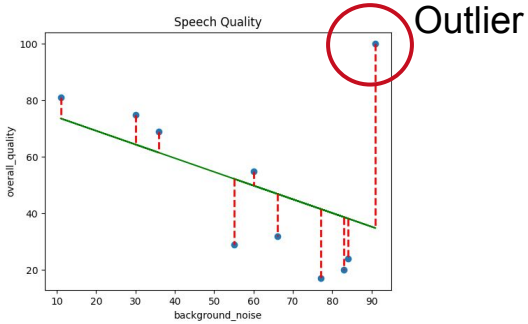
01 Simple Linear Regression

Key assumptions and statistical significance

Smaller Assumptions:

No Outliers





No Autocorrelation

- aspect of Independence of errors
- when residuals are not independent from each other, but exhibit a pattern across observations
- example: stock prices over time

Technische Universität Berlin

02 Non-Linear Regression

- Nonlinear regression is characterized by the fact that the prediction equation depends nonlinearly on one or more unknown parameters

- This is used in many real-world scenarios where the interactions between variables are complex

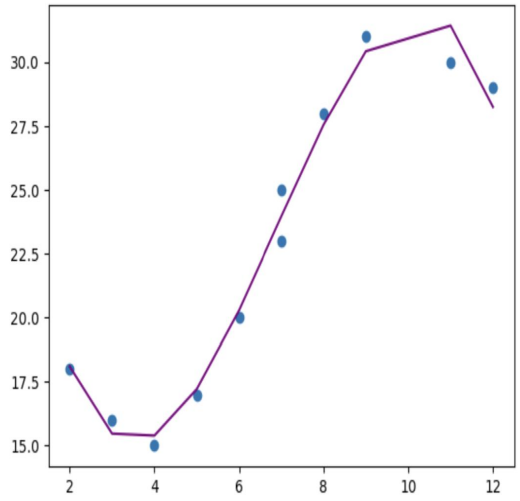
$$Y = f(X, c) + \varepsilon$$

Y : Dependent variable to be predicted
X : Independent variable
f : non-linear function
c : coefficients
ε: the error term

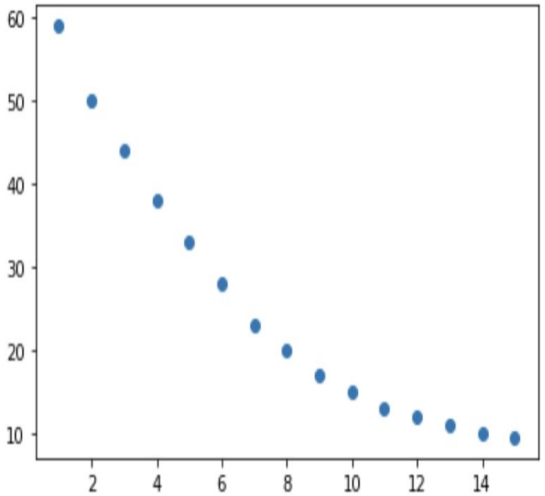
Technische Universität Berlin

02 Non-Linear Regression

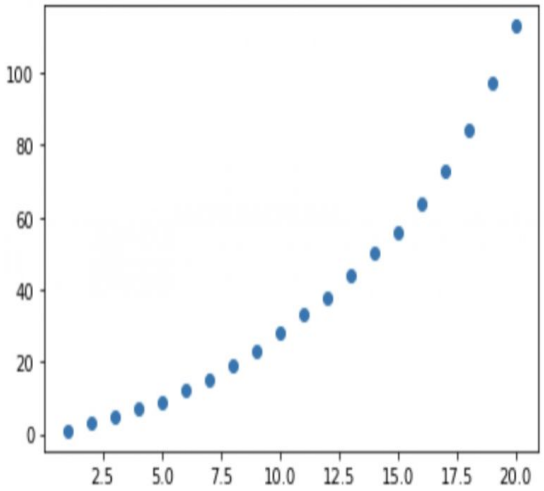
The most common types of Non-Linear Regression



Polynomial Regression



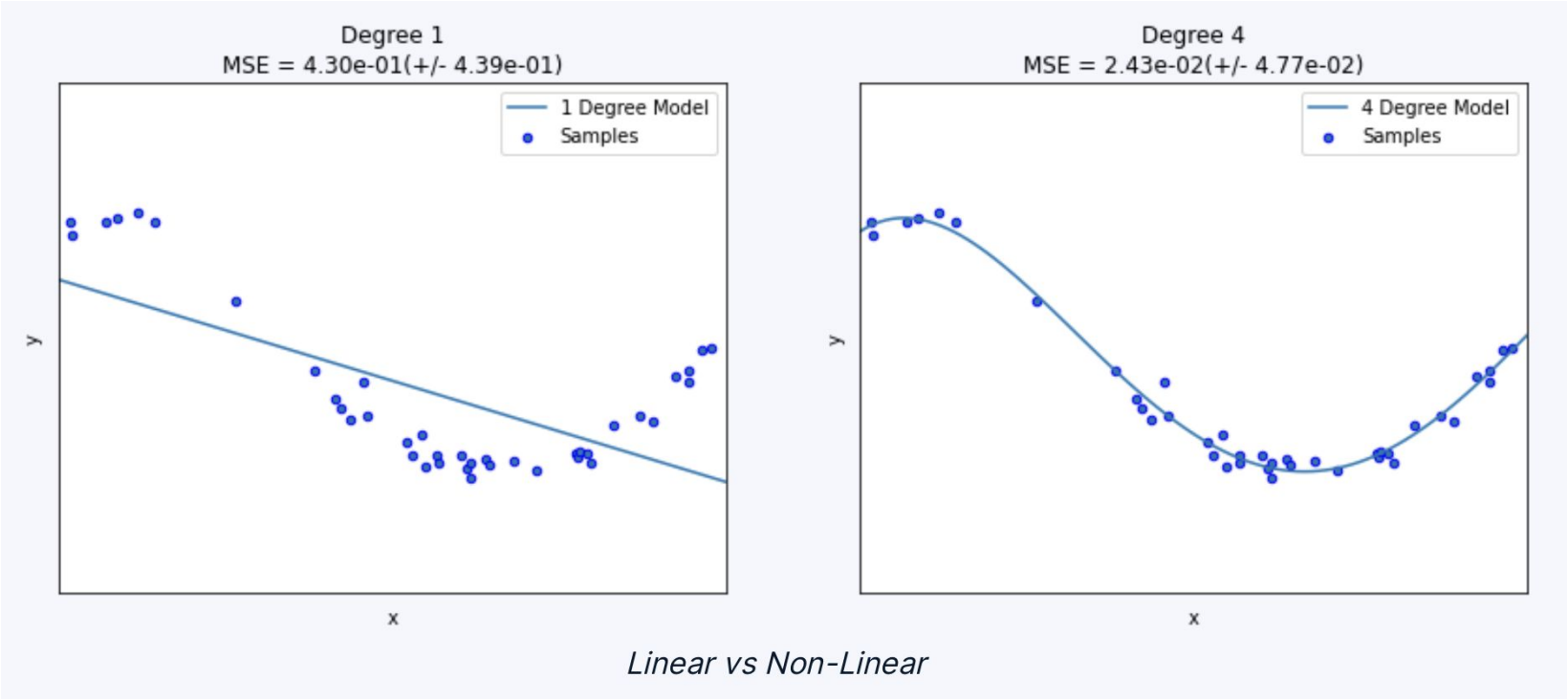
Logarithmic Regression



Exponential Regression

02 Non-Linear Regression

Linear Regression vs Non-Linear Regression



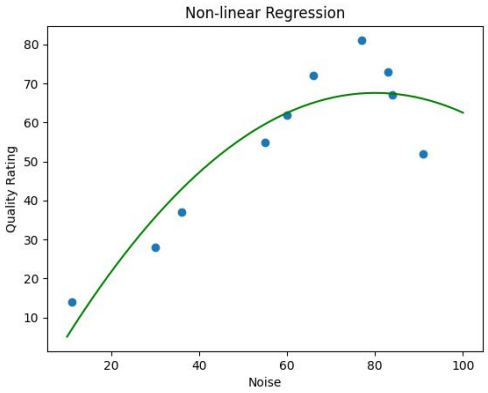
Linear vs Non-Linear

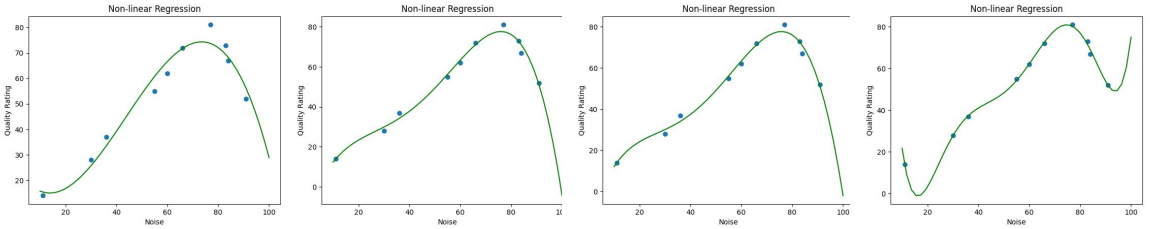
02 Non-Linear Regression

Curve fitting

- The process of finding a curve that best represents the relationship between a set of data points

Step 1. Generate Initial Curve → Step 2. Iterative Adjustment → Step 3. Stopping Criteria





...

Technische Universität Berlin

02 Non-Linear Regression

The Danger of Curve fitting : *Overfitting*

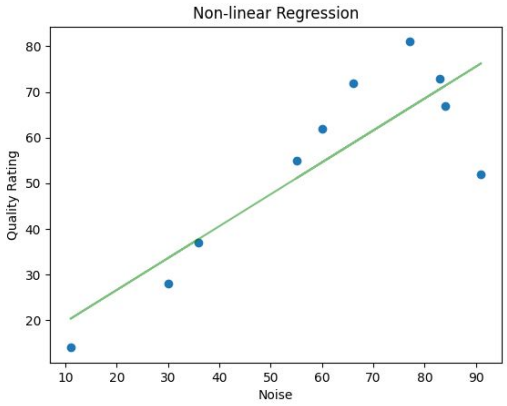
- Overfitting generally occurs when a model is excessively complex relative to the amount of data available
- Poor Generalization: The model performs well on training data but poorly on unseen data
- Increased Variance: The model is highly sensitive to fluctuations in the training data, leading to high variance and instability

Technische Universität Berlin

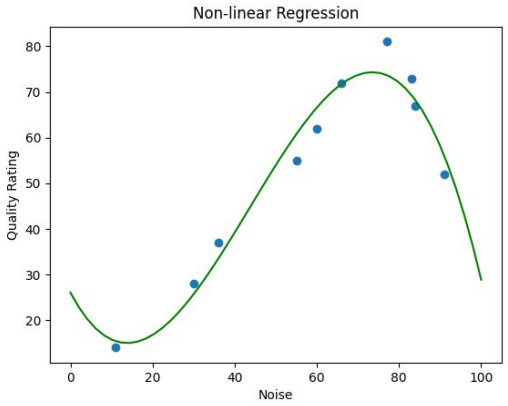
02 Non-Linear Regression

The Danger of Curve fitting : *Overfitting*

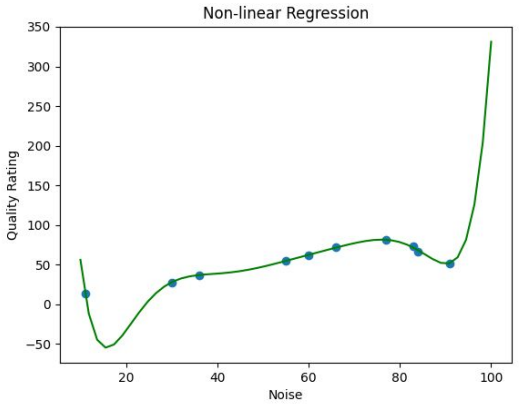
a simple model (underfitting)



an optimal model (good fit)



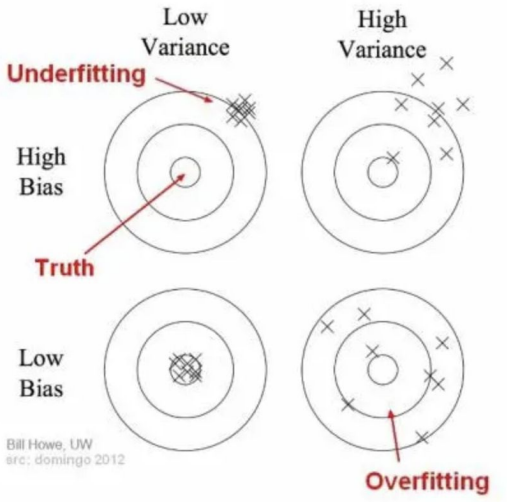
a complex model (overfitting)

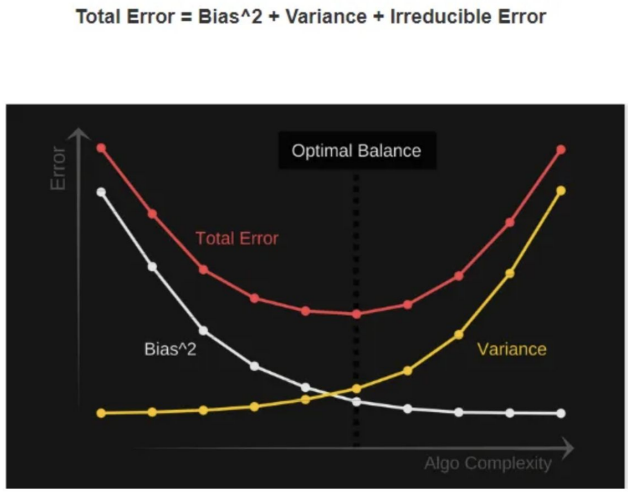


02 Non-Linear Regression

Bias-Variance Tradeoff

- ***Bias*** refers to the error introduced by approximating a real-world problem, which may be complex, with a simplified model (High Bias ⟶ underfitting)
- ***Variance*** refers to the error introduced by the model's sensitivity to small fluctuations in the training data (High Variance ⟶ overfitting)





Technische Universität Berlin

03 Multiple Linear Regression

- Generalization of simple linear regression
- SLR models relationship between the DV and one IV
- MLR models relationship between the DV and two or more IV
- By taking into account other IVs, it is possible to cancel the effects of these IVs on the prediction and isolate the IV of interest

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p + \varepsilon$$

Y : Dependent variable to be predicted

X_1 ,..., X_p : Independent variable

β_0 : y-intercept, the expected value of Y when X = 0

β_1 ,... , β_p: Slope of relationship between Y and X_p

ε: the error term

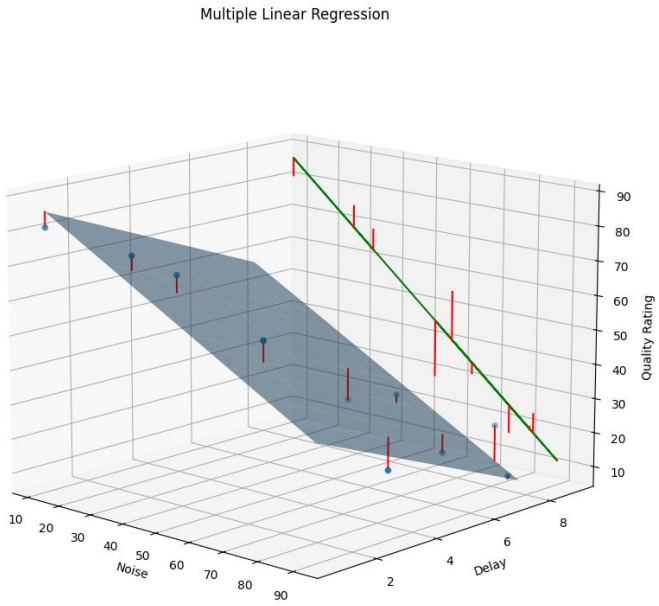
p : Number of dimensions/ independent variables

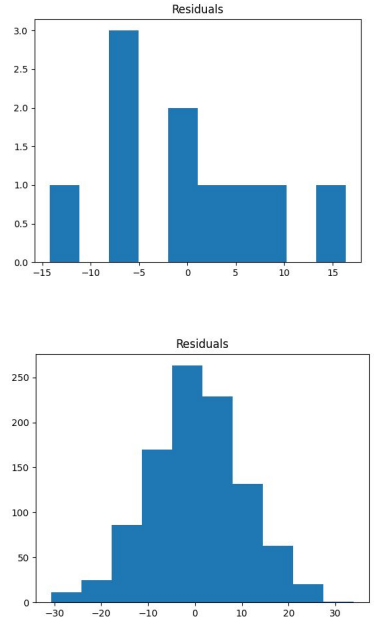
03 Multiple Linear Regression

Key assumptions

1. - 3. from Simple Linear Regression

4. Multivariate Normality: Instead of residuals from one IV and a DV, we have residuals for more than two IVs.





03 Multiple Linear Regression

Key assumptions

5. No Multicollinearity: No correlation between independent variables.

Multicollinearity: When two or more independent variables are strongly correlated with each other.

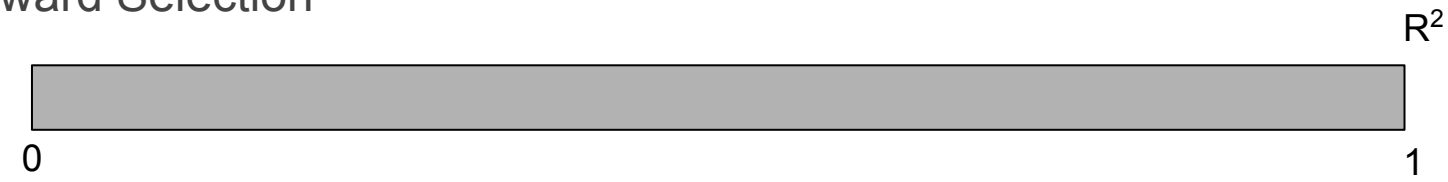
Problem: Effect of individual variables can not be clearly separated.

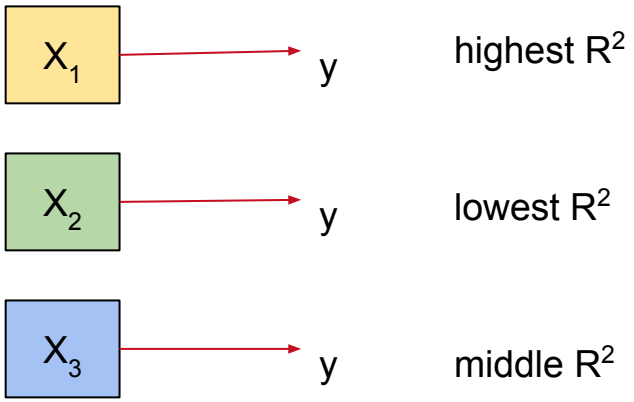
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Prediction →ok **Measurement of influence →bad**

04 Developing Multiple Linear Regression

Forward Selection

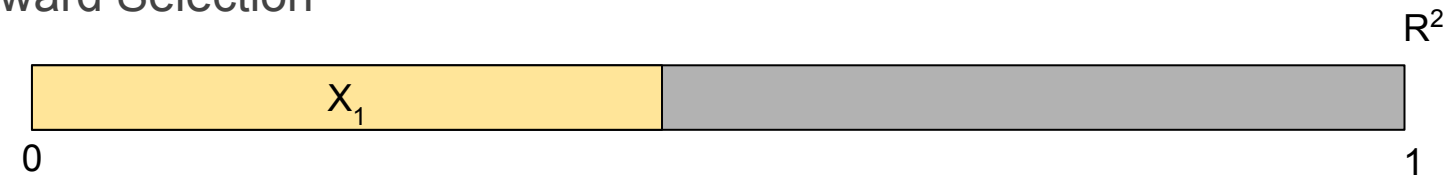


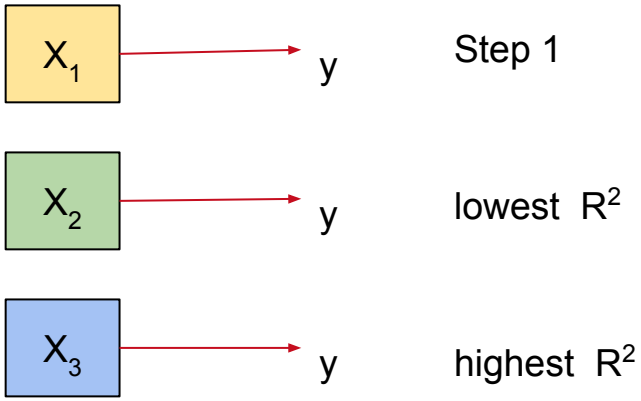


Step 1:
Choose the variable with the highest significance
Only keep the variable if it is above the significance threshold

04 Developing Multiple Linear Regression

Forward Selection

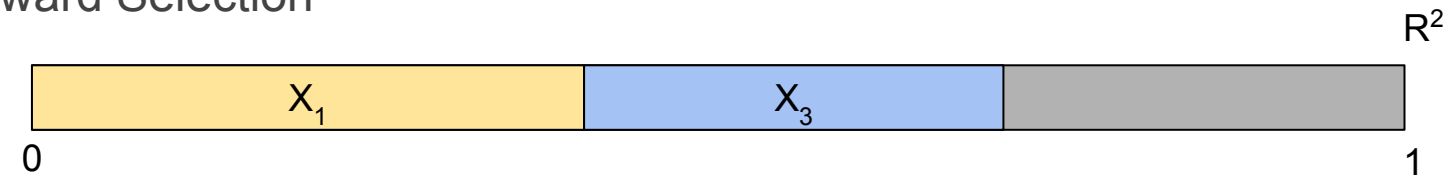


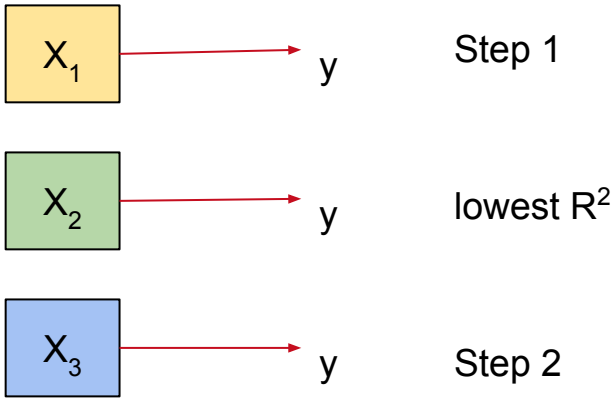


Step 2:
When variable added it cannot be removed
Repeat step 1

04 Developing Multiple Linear Regression

Forward Selection

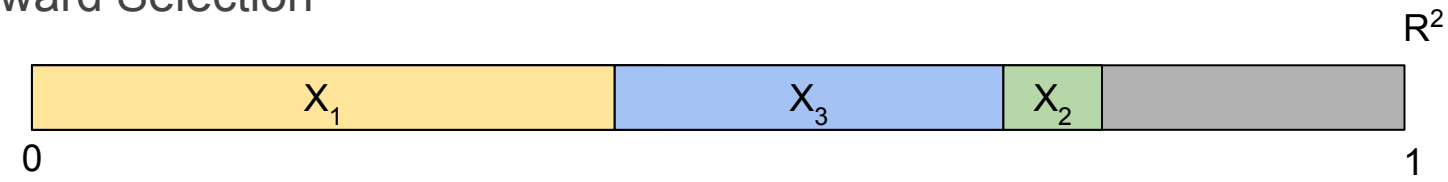




Step 2:
R^2 will never decrease when variable added, but ratios might change

04 Developing Multiple Linear Regression

Forward Selection



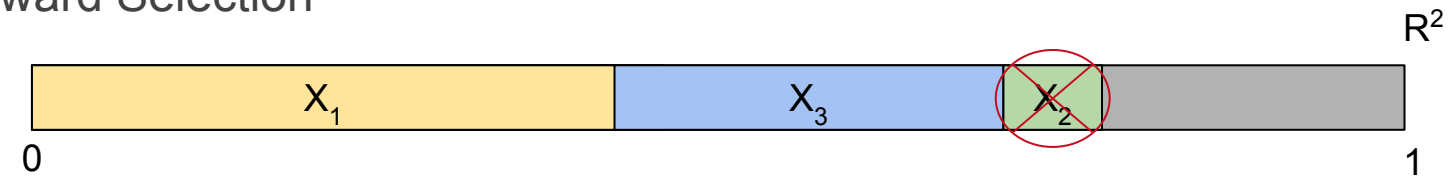
X_1 → y kept from step 1

X_2 → y Added

X_3 → y kept from step 2

04 Developing Multiple Linear Regression

Forward Selection

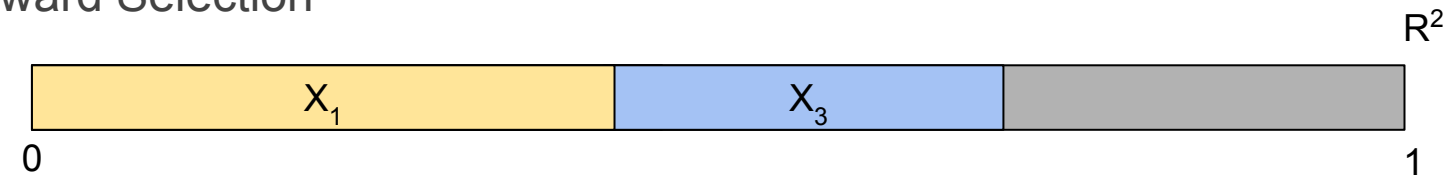


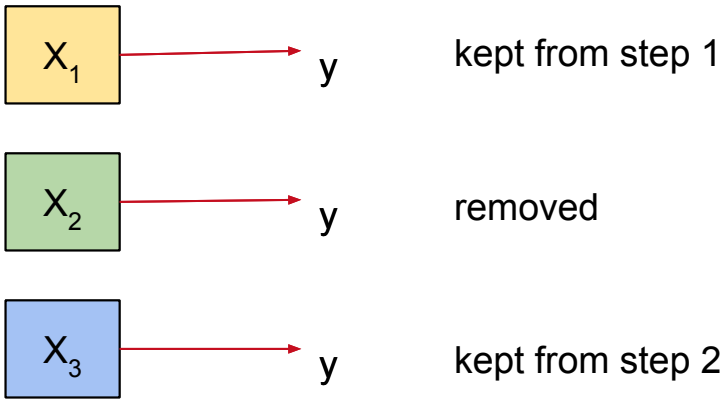
X_1 →	y	kept from step 1
X_2 →	y	removed
X_3 →	y	kept from step 2

Termination step:
If variable is not above the significance threshold remove it and terminate the algorithm

04 Developing Multiple Linear Regression

Forward Selection



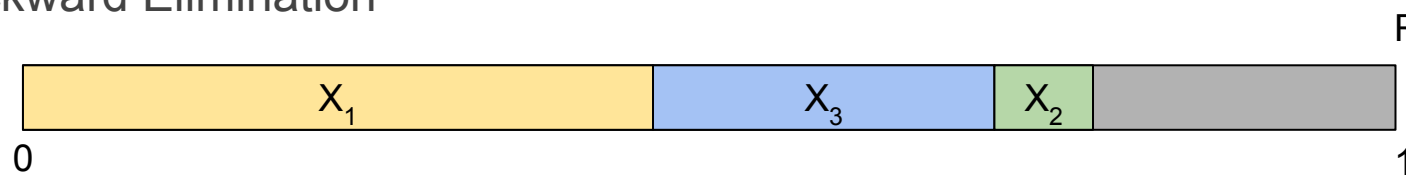


Result:

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$$

04 Developing Multiple Linear Regression

Backward Elimination



y = X_1 + X_2 + X_3 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

---

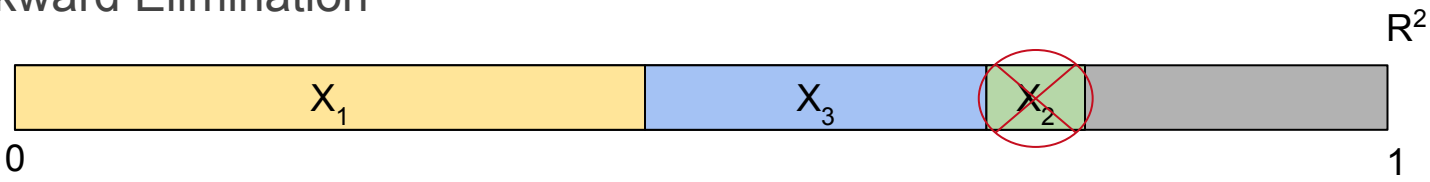
y = X_1 + X_2 + X_3 $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

y = X_1 + X_2 + X_3 $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$

y = X_1 + X_2 + X_3 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

04 Developing Multiple Linear Regression

Backward Elimination



y = X_1 + X_2 + X_3 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

---

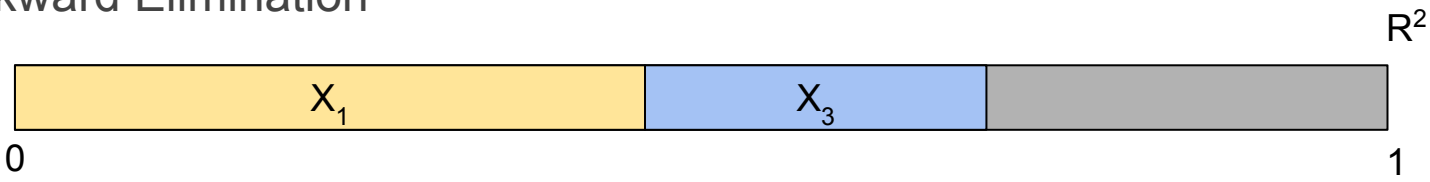
y = X_1 + X_2 + X_3 $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

y = X_1 + X_2 + X_3 $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$

y = X_1 + X_2 + X_3 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

04 Developing Multiple Linear Regression

Backward Elimination



y = X_1 + X_3 $\quad Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$

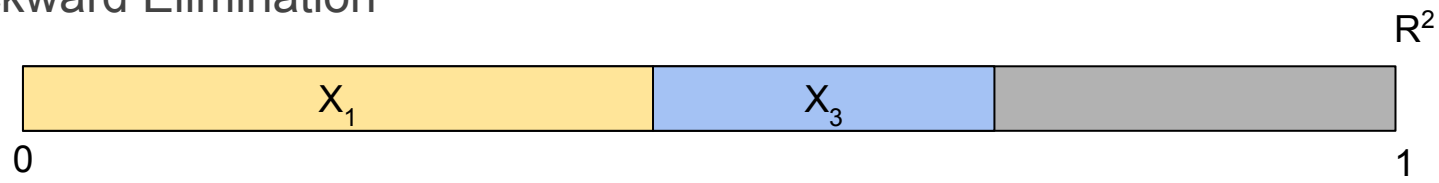
---

y = X_1 + X_3 $\quad Y = \beta_0 + \beta_3 X_3 + \varepsilon$

y = X_1 + X_3 $\quad Y = \beta_0 + \beta_1 X_1 + \varepsilon$

04 Developing Multiple Linear Regression

Backward Elimination



y = X_1 + X_3

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$$

1. Add all variables into the function
2. Remove the variable with the least significance from the function
3. Repeat Step 2 until all variables with low significance are removed

Technische Universität Berlin

04 Developing Multiple Linear Regression

Stepwise Selection

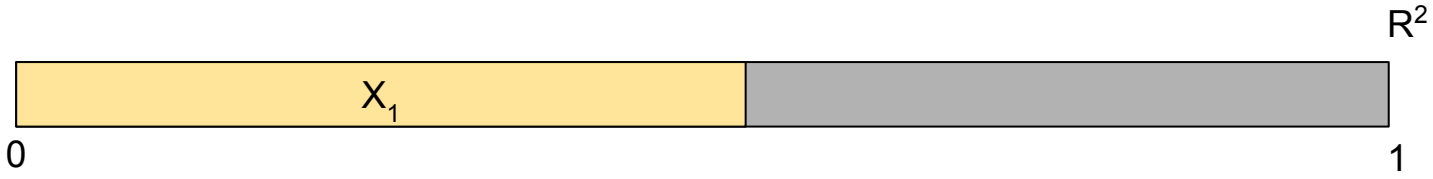
- Combination of Forward Selection and Backward Elimination
- At every step variables can be added or removed to get a good model, without exclusion

Technische Universität Berlin

04 Developing Multiple Linear Regression

Stepwise Selection

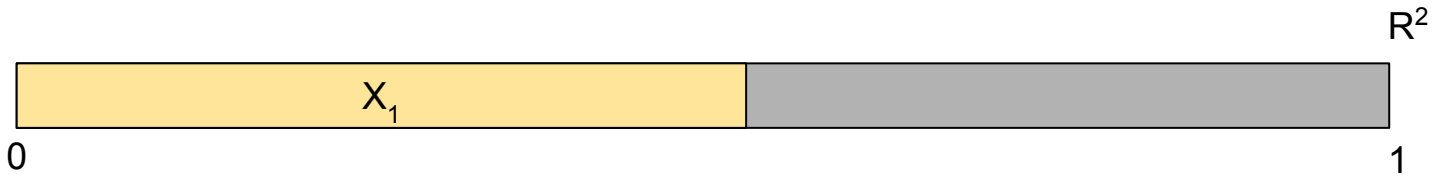
- Combination of Forward Selection and Backward Elimination
- At every step variables can be added or removed to get a good model, without exclusion



Step 1:
Choose most significant variable (Similarly in Forward Selection)

04 Developing Multiple Linear Regression

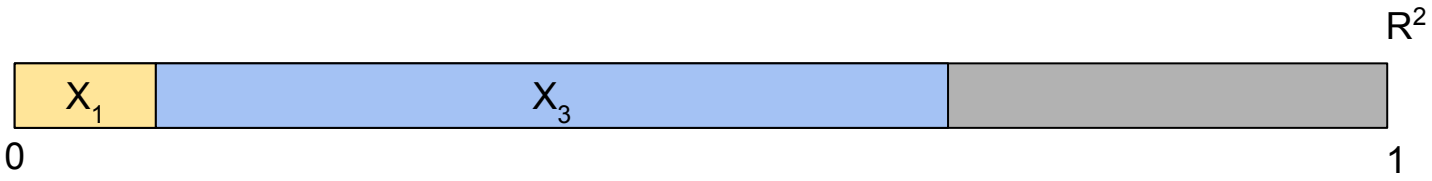
Stepwise Selection



Step 2:
Check if any variable can be removed, due to changed ratio of significance (Similarly in Backward Elimination)

04 Developing Multiple Linear Regression

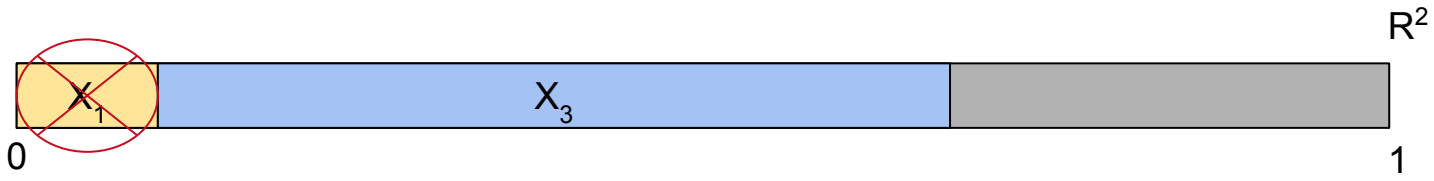
Stepwise Selection



Step 3:
Repeat step 1 and 2

04 Developing Multiple Linear Regression

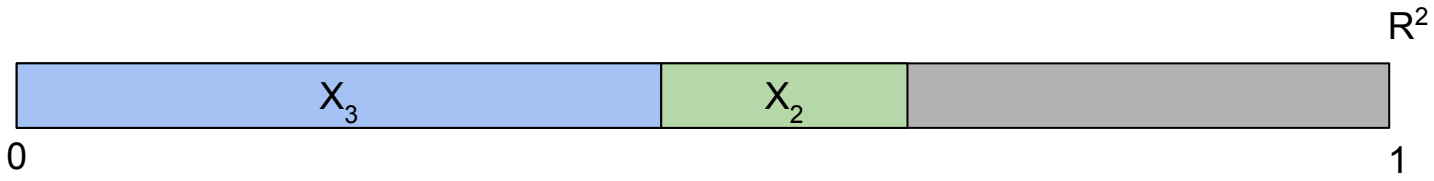
Stepwise Selection



- Due to the change of significance of X_1 when X_3 has been added, X_1 it will be removed

04 Developing Multiple Linear Regression

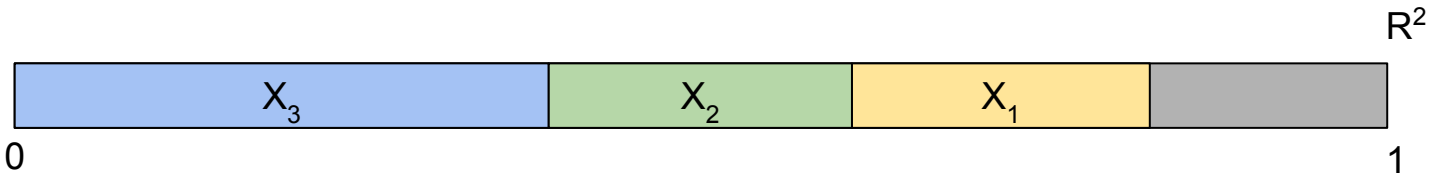
Stepwise Selection



- Adding of X_2, no variable is insignificant = thus no removal

04 Developing Multiple Linear Regression

Stepwise Selection



- Due to the effect increased significance X_1 , it will not be removed like before
- The combination of all 3 variables changed the ratio, so that every variable play significant roles in the model

04 Developing Multiple Linear Regression

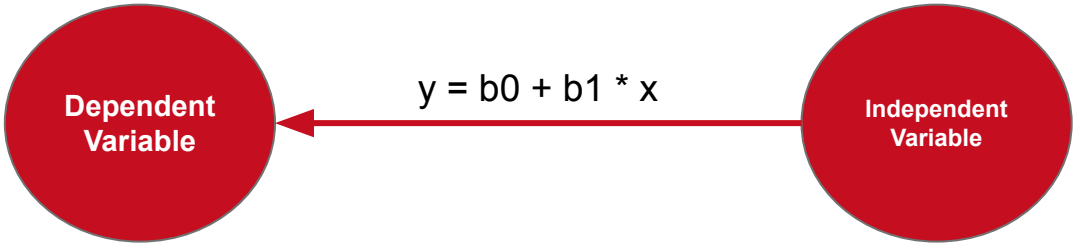
Interaction Effect

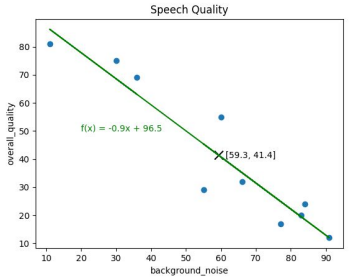
- occurs when effect of one IV depends on the value of another IV
- Effect on the DV is multiplicative not additive

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + \varepsilon$$

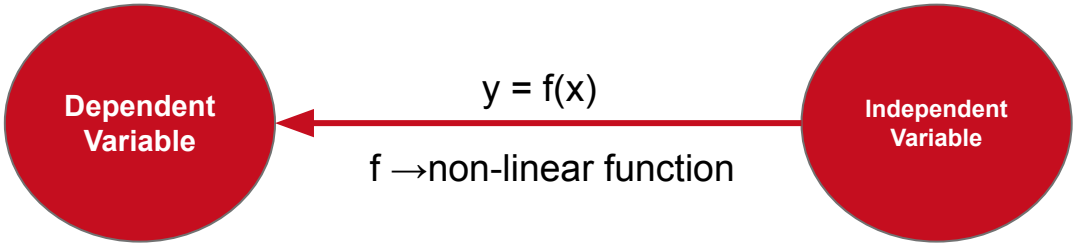
Recap Modeling

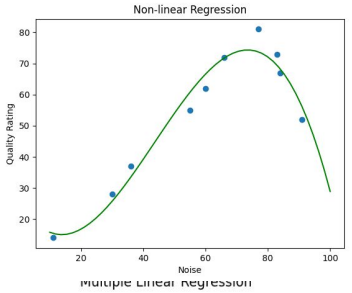
Simple Linear Regression



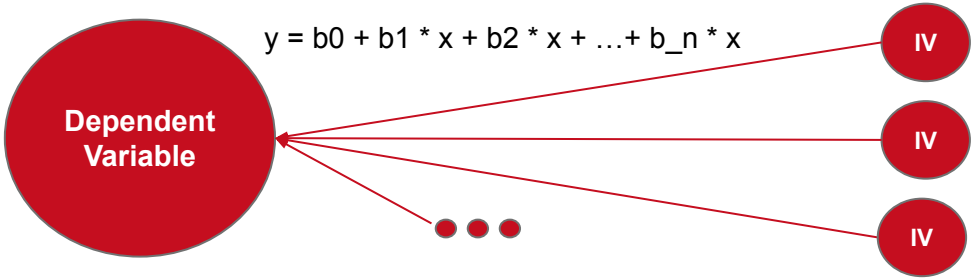


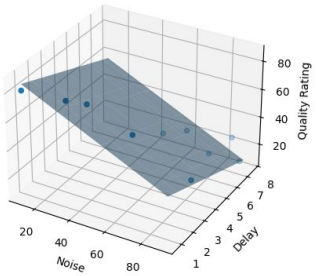
Non-linear Regression





Multiple Linear Regression



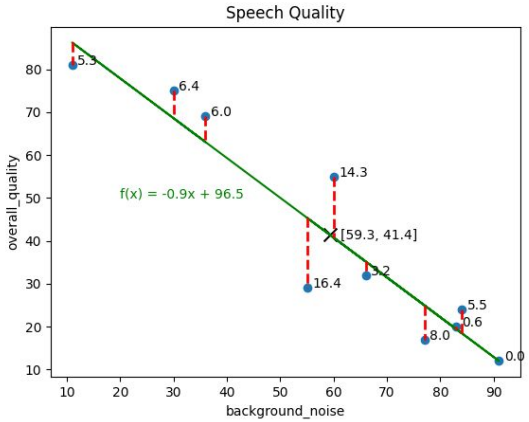


Technische Universität Berlin

Recap

Ordinary Least Squares: optimization problem which seeks to minimize the sum of squares error (SSE)

Σ = 680.22 ←



Bias-Variance Tradeoff: tradeoff in complexity between both categories so total error is minimized
Bias: error between the model's predicted value and the actual value
Variance: amount by which the estimates of the target function change

Total Error = Bias^2 + Variance + Irreducible Error

underfitting | optimum | overfitting

Technische Universität Berlin

Recap

Forward Selection: starts with no variables and adds variables incrementally

Backward Elimination: begins with all variables and removes the least statistically significant ones

Stepwise Selection: iteratively examines the statistical significance of each variable, combining forward selection and backward elimination

Interaction Effect: occurs when the impact of one variable on the outcome depends on the level of another variable, indicating that their combined effect is not simply additive

Technische Universität Berlin

References

https://datatab.net/tutorial/linear-regression
https://aws.amazon.com/what-is/linear-regression/?nc1=h_ls
https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/
R5_Correlation-Regression4.html
https://youtu.be/Rb8MnMEJTI4?t=101
https://pieriantraining.com/nonlinear-regression-in-machine-learning-python/
http://pzs.dstu.dp.ua/DataMining/mls/bibl/Nonlinear%20regression.pdf
https://www.heavy.ai/technical-glossary/statistical-modeling
https://www.statology.org/sklearn-polynomial-regression/
https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-linear-regression/
https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229