



## Quantitative Decision Making: Linear Regression

Faculty of Economics & Management, Institute of Technology and Management

---

25. July 2018

# Agenda

1. Introduction to Predictive Analytics
2. Introduction to Linear Regression with Spreadsheets
  1. Linear regression
  2. Evaluation of linear regression
  3. Non-linear influences



95

... percent is the accuracy with which researchers at Cambridge University have predicted the skin color of a Facebook user solely on the basis of his "like" information. They applied machine learning models to data from 58,466 subjects.

Gender:

93%

marital status:

67%

Hetero-/homo-  
sexuality ...

for men:

88%

for women:

75%

Democrat or  
Republican:

85%

Christian or Muslim:

82%

Smoker:

73%

alcoholic:

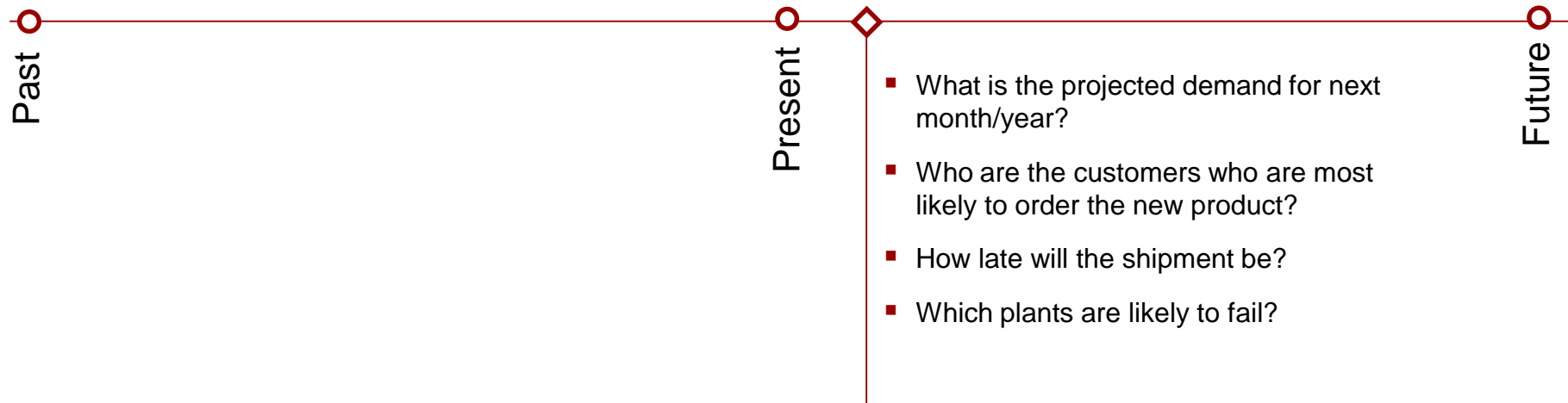
70%

Drug addiction:

65%



- ”
- *Predictive analytics predicts future trends, behaviors and events to support decisions*
  - *Predictive analytics creates models that estimate the value of a target variable of an unseen example*
  - *Predictive analytics predict what will happen in the future*



# Case: Internet Of Trains from Siemens



**SIEMENS**

- Siemens is an international technology group organized in various areas (Automation and Control, Lighting, Medical, Power and Transportation) with ~364,000 employees and sales of €79.644 billion.
  - In the transportation sector, especially for rail transport, Siemens has begun to monitor vehicles and evaluate train sensor data for various operators ("Internet of Trains").
  - The aim is to identify the main causes of vehicle problems and to provide predictive maintenance for the vehicles. Operators benefit from the availability and reliability of the vehicles and the transport services offered.
  - In the pilot projects, time series of sensor data were combined with faults to identify and predict patterns of faults and failures and thus make cost-effective maintenance decisions (repairs can be carried out location-optimally and cost less time and effort).
- Siemens develops new business model
  - In one example, a Spanish rail operator is only 5 minutes late on 2,300 journeys.

# Case: Automated weighing at Zalando



- Zalando is a European eCommerce retailer specialising in clothing with 12,000 employees and a turnover of €987 million.
  - The weight of the products is a decisive cost factor for the products shipped (the expected weight is precalculated before shipment) and must be determined by labor-intensive weighing of the individual products. It should therefore be ensured that the expected weight is as accurate as possible and that the work involved in weighing is reduced at the same time.
  - Since the final packages are weighed individually before dispatch, there is an automated process. However, the packages contain additional packaging material and during the inspection. It was noticed that the package weight would have been negative if the individual products had been subtracted.
  - The weights of the products were therefore estimated using a Bavarian model, which revealed some measurement errors in the test.
- The determination of the weight of the products is now more reliable
  - The manual weighing step is no longer necessary

# Define Goal

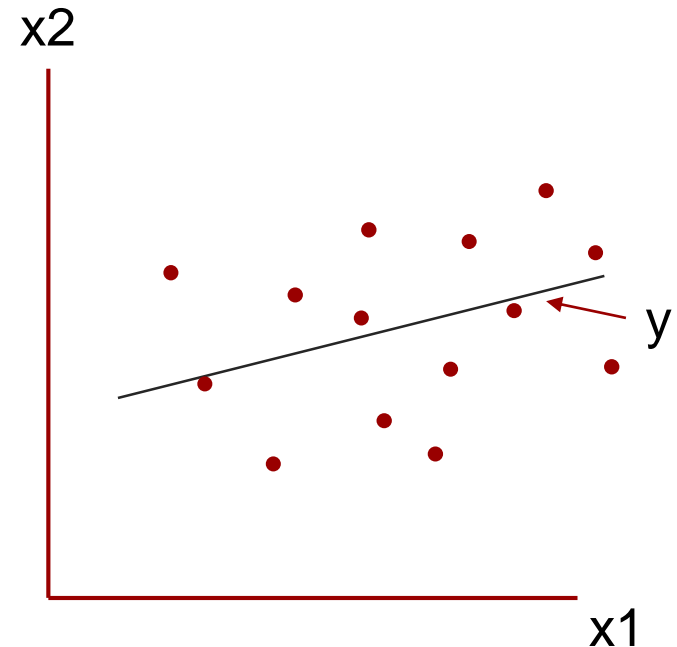
- The variable to be determined - **the target variable (y)** - plays a decisive role in prediction models
  - It should therefore be clearly defined
  - Check before an analytical step:
    - Robustness (not susceptible to interference)
    - Stability (constant over time) of the target variable
  
- Example: Customer churn
  - Active customer churn: Customer terminates the business relationship
    - It is difficult to determine without a contractual agreement: Customers stop shopping in stores, wholesalers no longer accept goods, logistics service provider is no longer commissioned
    - BUT: from **when** is this noticeable 1 month without purchase/order, 3 months, 6 months?
  - Passive customer churn: customer buys/contracts less
    - Even with contractual agreements, order quantities/purchase quantities are variable to a certain extent and a reduction in volume may occur intentionally or accidentally
    - THEREFORE: from **which volume** reduction over **which period** can this be determined??

# Different Prediction Methods I/VIII

## Linear Regression

- Modeling of continuous target variables ( $y$ ) based on one or more explanatory variables ( $x$ )\*

- Identification of linear factors ( $y=f(x_i)$ ) (multiplied by the explanatory variables,  $\beta$  parameters) and an additive constant by minimizing the square error



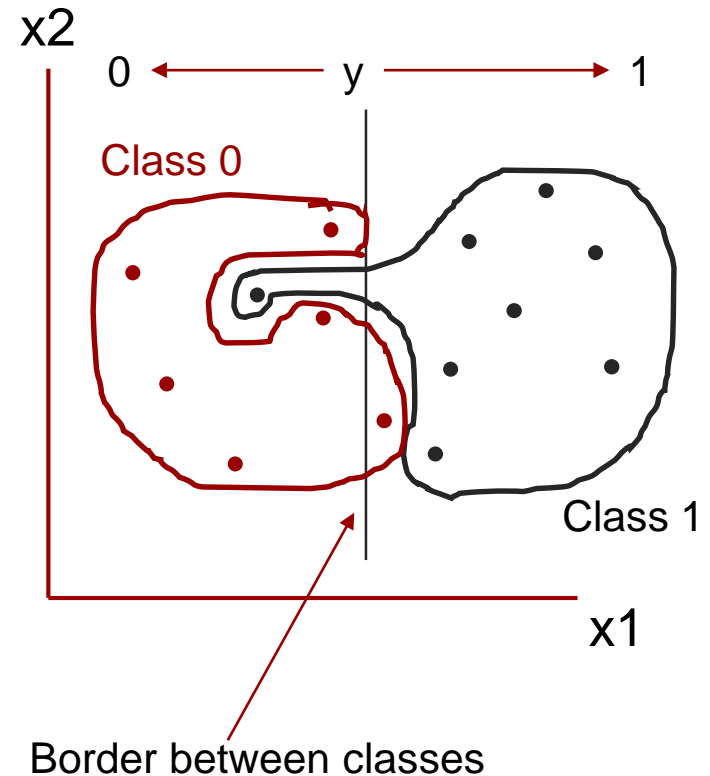
Note: The algorithms require test data for which the target variable is known.



# Different Prediction Methods II/VIII

## Logistic Regression

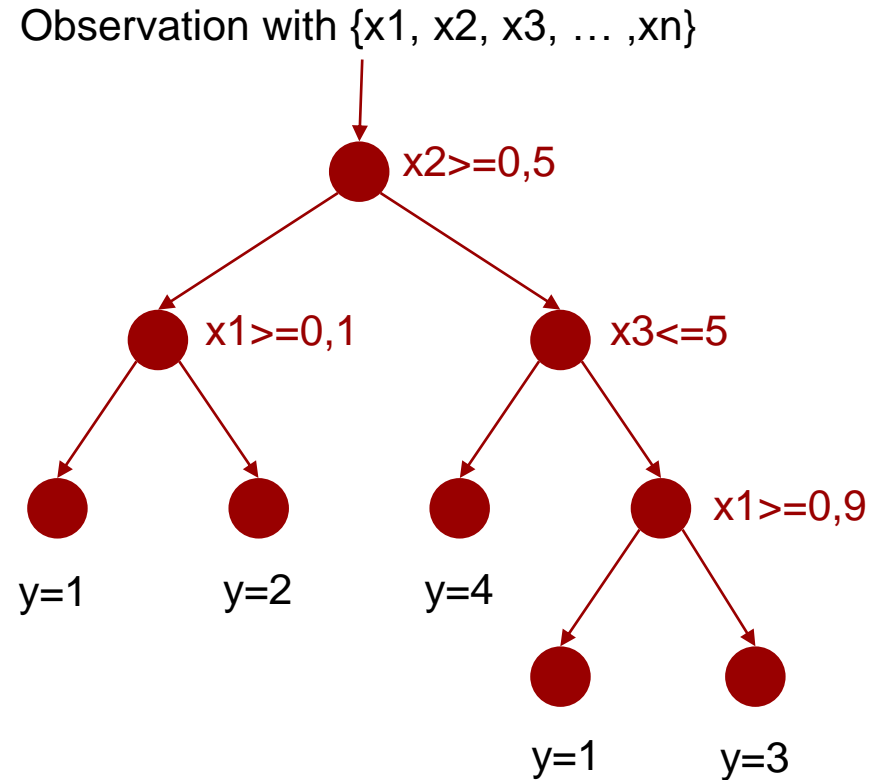
- Modeling of a binary classified target variable
  - With one or more explanatory variables
- Application of linear regression, but with limitation to values between 0 and 1
  - The target variable represents the (log) probability that *class 1* will occur
  - The probability which *best* represents the boundary between classes 1 and 0 must be determined.
    - (Depending on the purpose of the model, does not have to be 0.5! For example, if a wrong estimation of class 1 has severe consequences.)



# Different Prediction Methods III/VIII

## Decision Tree

- Recursive division algorithm represented in a tree structure
  - Each node of the tree → condition to be tested (for example, fixed value of a variable) that determines the path to the next condition.
  - The path ends in a "sheet" → predicted value of the target variable (e.g. for classification with several classes)
- Algorithm contains splitting decisions of conditions, stop decisions and allocation decisions
  - Allocation is often made by the majority (majority class), on the sheet
  - Remaining decisions → Minimization of "impurity" (if possible only one remaining class per sheet)

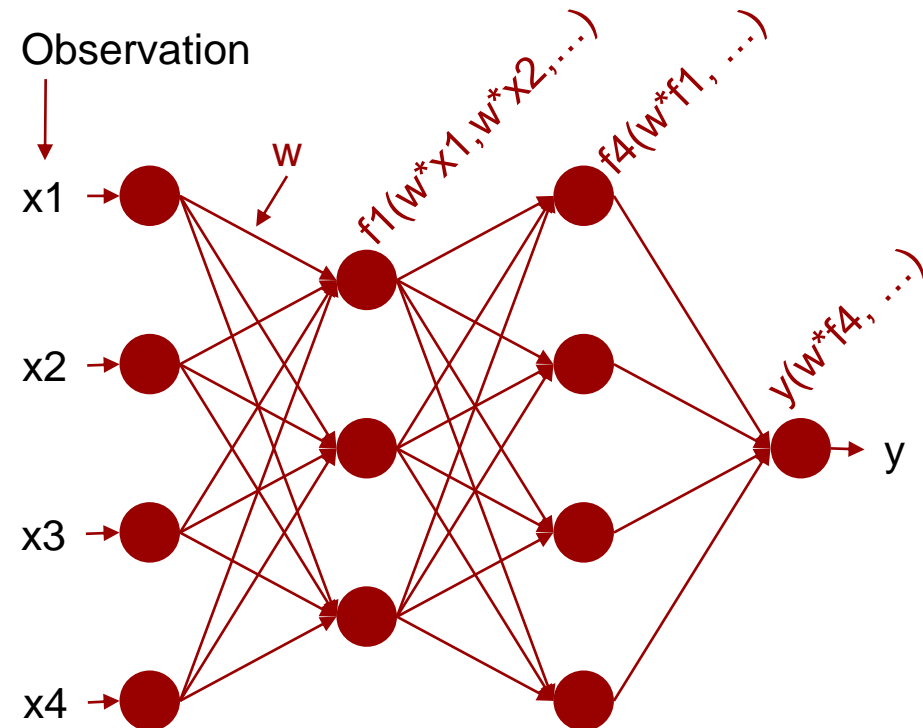


# Different Prediction Methods IV/VIII

## Neural Networks

- Mathematical representations inspired by the functioning of the human brain
- Node  $\rightarrow$  neurons (side by side and in series) determine the value of the target variable
- Is "trained" with the test data

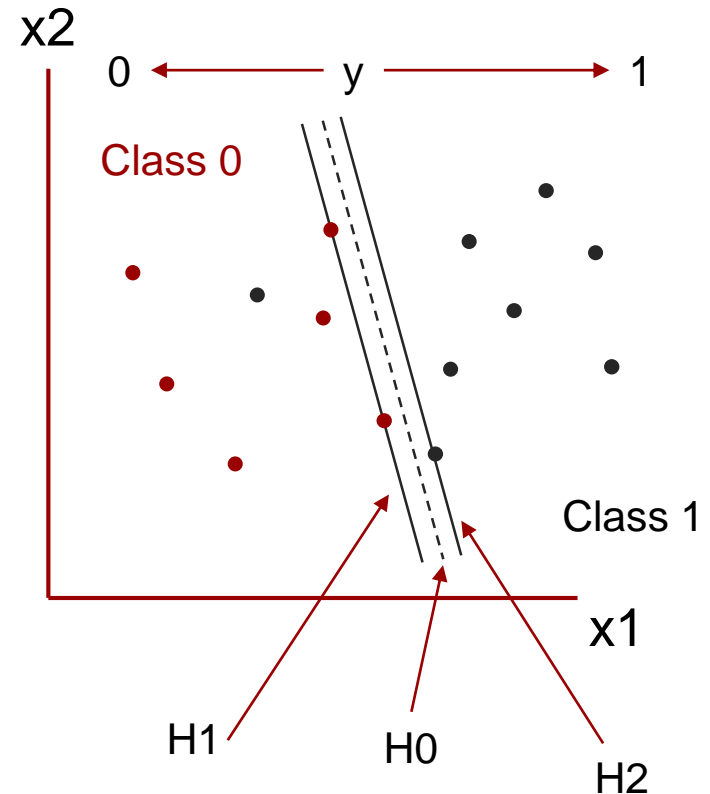
- Neurons  $\rightarrow$  individual statistical models (logistic regression or linear regression)
- Weightings ( $w$ ) are adjusted iteratively
- Training  $\rightarrow$  Predicted values are compared with actual values of the target variables; weighting is adjusted accordingly (the model "learns").



# Different Prediction Methods V/VIII

## Support Vector Machine

- Three hyperplanes (one  $n-1$  dimensional plane in  $n$ -dimensional space) for differentiation  $\rightarrow$  two hyperplanes (H1, H2) at the edges of the classes, one boundary hyperplane in the center
  - Data points on H1 and H2  $\rightarrow$  Support vectors
  - Aim  $\rightarrow$  Maximizing the distance between H1 and H2
- Square optimization  $\rightarrow$  optimal hyperplanes or support vectors that describe them
  - The quadratic optimization can be solved with Lagrange optimization
  - A perfect division is pursued  $\rightarrow$  for realistic calculation an error term is used (which allows misclassification)



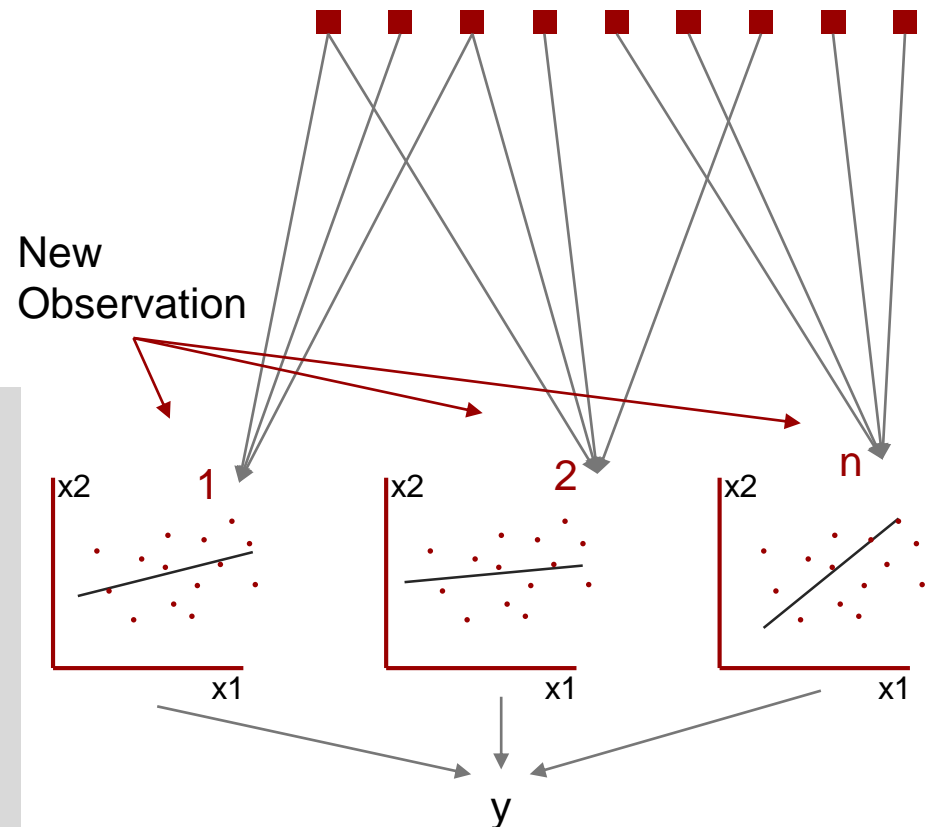
# Different Prediction Methods VI/VIII

## Bagging / Bootstrap Aggregation (ensemble Method\*)

- From the available data  $n$  bootstraps are taken (samples, so that an observation can be included in several samples)
- $n$  classifications are created (e.g. decision trees, logistic regression) → jointly estimate the class of the new observation

- Methodology for classifications ("classifying") is not restricted
- A new observation is presented to all  $n$  bootstrap classifiers → Usually majority vote to estimate the class

Training observations



\*Multiple models can be used simultaneously instead of a single model

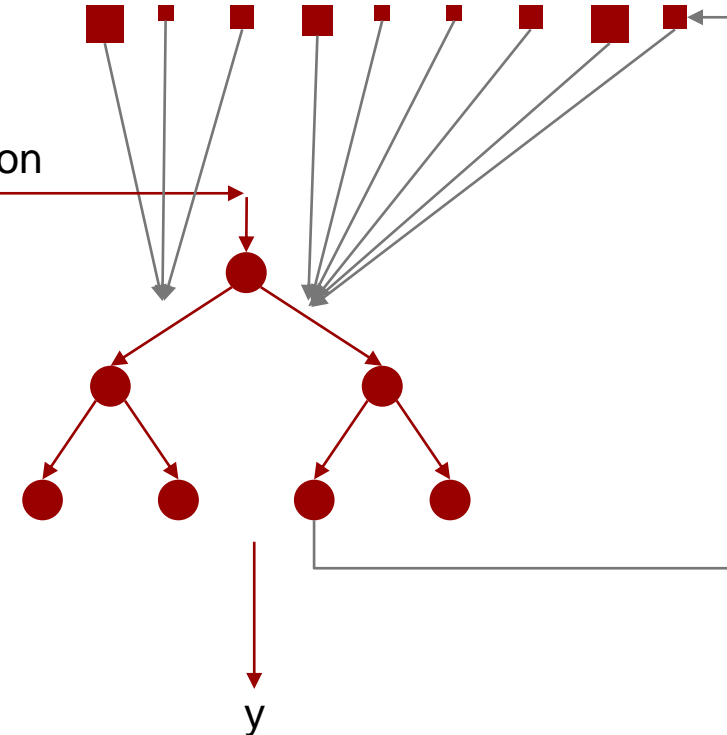
# Different Prediction Methods VII/VIII

## Boosting (ensemble Method\*)

- Several models are used iteratively in boosting.
  - In the iteration steps, training observations that are difficult to estimate are given a higher weighting
- Start with a model with uniform weights → Error values between observed values and estimated value of the target variable ( $y$ )
  - Observations with high errors → higher weightings in the next iteration step
  - The number of iterative steps, "Boosting runs" is fixed or determined by independent validation data
  - There is a strong risk of "overfitting".

Training observations

New Observation



\*Multiple models can be used simultaneously instead of a single model

# Different Prediction Methods VIII/VIII

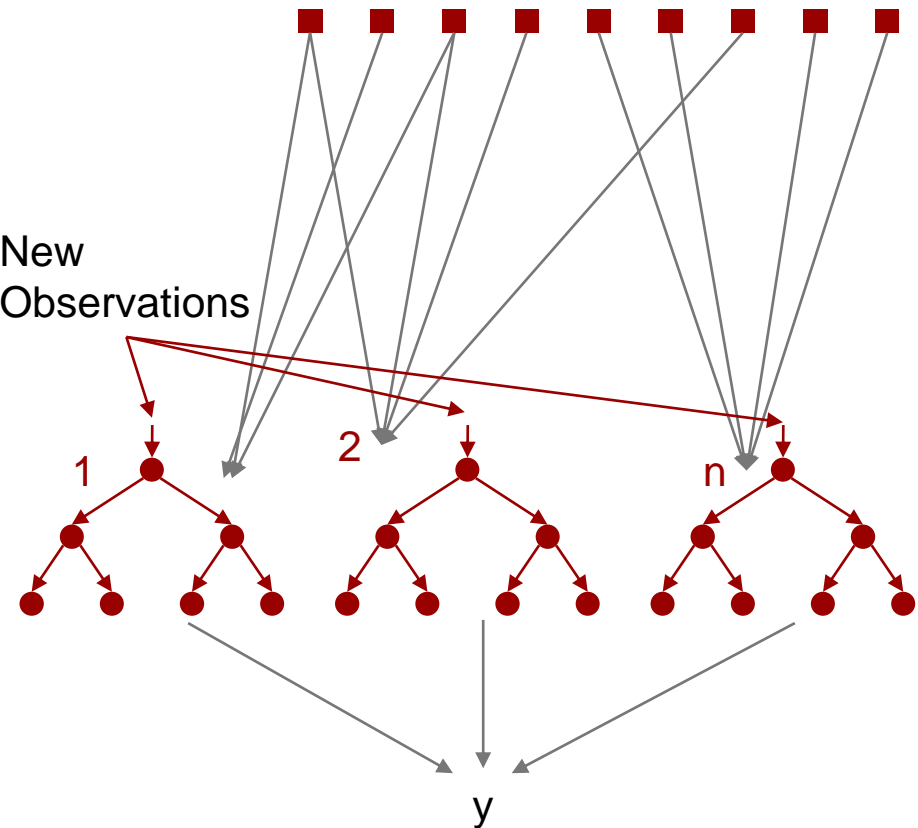
## Random Forrest (ensemble Method\*)

- A "forest" of decision trees is created
- Based on bootstrapping (i.e. drawing random samples so that one element can be in several samples) decision trees/decision trees are created

- The accuracy of Random Forrest (with *multiple* decision trees, based on a smaller sample of a large amount of data) is *often* superior to a large amount of data compared to a *single* decision tree.

Training observations

New Observations



\*Multiple models can be used simultaneously instead of a single model

# Elements of a prediction

- Predictive analytics applications are defined by two elements:

## What is predicted?

- The type of behavior (e.g. actions, events, incidents) that should be predicted for individuals, stocks or other entities.



## What will be done with the prediction?

- The decisions controlled by the forecast; the actions performed by the organizations in response to or informed by the forecasts.



## Summary: Predictive Analytics

---

- The value of a prediction is determined by the fact that the prediction leads to a decision or action
- A distinction can be made between the prediction of a categorical value (classification) and continuous values (regression)

# Agenda

1. Introduction to Predictive Analytics
2. Introduction to Linear Regression with Spreadsheets
  1. Linear regression
  2. Evaluation of linear regression
  3. Non-linear influences

# Regression Definition

” Regression tries to explain what influence a set of variables has on the result of a target variable.

## Variables

- The target variable is referred to as a "**dependent**" variable
- The input variables are referred to as "**independent**" variables

## Linear Regression is an explanatory tool

- Identifies the input variables that have the greatest statistical influence on the target variable (output).
- The regression model assumes that the dependent variable tends to fluctuate with the independent variables in a systematic manner.

## Attention

- Predictions should not be used outside the observed values of the independent variables used in the forecasting model.

# Linear regression

## Characteristics

Linear regression ...

- is applied to **continuous**, dependent variables
- assumes a **linear relationship** between the dependent variable and the independent variable
  - However, this assumption will not apply in some cases
- Assumes an **additive relationship** between variables
  - $\hat{y} = f(x) = \alpha + \beta_1 * x_1 + \beta_2 * x_2 + \dots + \beta_n * x_n$
  - $y = f(x) = \alpha + \beta_1 * x_1 + \beta_2 * x_2 + \dots + \beta_n * x_n + e$

### Nomenclature

$\hat{y}$  – Prediction of the dependent variable  
 $y$  – Observation of the dependent variable  
 $x$  – Value of the independent value  
 $\alpha$  – Intersection with the Y-axis, when all  $x = 0$  („intercept“)  
 $\beta$  – Regression coefficient / slope of the dependent variable ("slope")  
 $e$  – Random error ("residuals")

## Linear regression

# Calculation of intersection and regression coefficients

The function for calculating the prediction is called the regression function of the population ("population regression function")

- Between each observation ( $y$ ) and prediction ( $\hat{y}$ ) lies the difference of the error ( $e$ ) with corresponding combination of the independent variables ( $x$ )

- $e = y - \hat{y}$

- Intersection and regression coefficients are determined by minimizing the sum of the square error

- $$\min z = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (\alpha + \beta_1 * x_{1i} + \beta_2 * x_{2i} + \dots + \beta_n * x_{ni}))^2$$

### Nomenclature

$\hat{y}$  – Prediction of the dependent variable

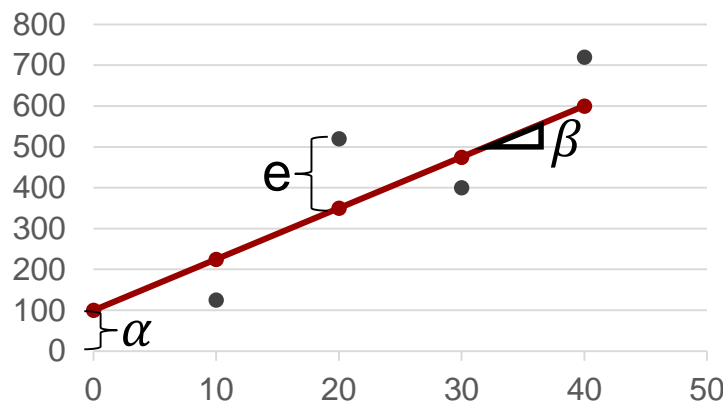
$y$  – Observation of the dependent variable

$x$  – Value of the independent value

$\alpha$  – Intersection with the Y-axis, when all  $x = 0$  („intercept“)

$\beta$  – Regression coefficient / slope of the dependent variable ("slope")

$e$  – Random error ("residuals")



# Assumptions for the calculation of the forecast

### Error ( $e$ )

- The distribution for a value of an independent variable ( $x$ ) has an average value of 0:  $\mu_e=0$ .
- The standard deviation of the error ( $\sigma_e$ ) is the same for all values of the independent variable ( $x$ ).\*
- The error is normally distributed for a value of an independent variable ( $x$ )
- The errors ( $e_i$ ) for different observations ( $y_i$ ) are independent of each other

In summary → the errors are independent and identically distributed (iid)

### Consideration of assumptions about the error

- Individual variables of a regression model are evaluated by the p-value
- The p-value is a measure of the fulfilment of the assumptions about the error  
→ therefore a p-value is given for each individual independent variable

### Nomenclature

$\hat{y}$  – Prediction of the dependent variable  
 $y$  – Observation of the dependent variable  
 $x$  – Value of the independent variable  
 $\alpha$  – Intersection with the Y-axis, when all  $x = 0$  („intercept“)  
 $\beta$  – Regression coefficient / slope of the dependent variable ("slope")  
 $e$  – Random error ("residuals")

\*Note: there can be several observations  $\{x, y\}$  with different  $y$  at the same  $x$ !

BEAR Food and Beverages supplies food to several customers in the region. A few weeks ago BEAR started to order a new Logistics Service Provides (LSP) for transport. The LSP are paid after the order has been fulfilled, the invoice amount takes into account actual working time, fuel consumption, toll costs and other factors. You are asked to develop a model for estimating transport costs and identify the influence of cost drivers in order to make the costs of LSP comparable with others.

- a) Create a regression model that estimates costs using the following variables
  - Distance
- b) Make a forecast for the following shipment: 553 km, 5.2 t, refrigerated, Wednesday, toll route

# Solution 4-1A

Example - Solution is presented

**1** Click on **Datenanalyse** in the ribbon.

**2** Select **Regression** in the **Analyse-Funktionen** dialog box.

**3** Configure the **Regression** task pane:

- Eingabe:**
  - Y-Eingabebereich: **\$B\$1:\$B\$51**
  - X-Eingabebereich: **\$C\$1:\$C\$51**
  - ☒ **Beschriftungen**
  - ☐ **Konstante ist Null**
  - ☐ **Konfidenzniveau:** 95 %
- Ausgabe:**
  - ☒ **Ausgabebereich:** **\$L\$1**
  - ☐ **Neues Tabellenblatt:**
  - ☐ **Neue Arbeitsmappe:**
  - Residuen:**
    - ☐ **Residuen**
    - ☐ **Standardisierte Residuen**
    - ☐ **Residuenplots**
    - ☐ **Kurvenanpassung**
  - Normalverteilte Wahrscheinlichkeit:**
    - ☐ **Quantilsplot**

	A	B	C	D	E	F	G
1	TransportNr	Transportkosten	Entfernung(km)	Transportkosten(km)	Gekauft	Montatresk	Wochentag(00-01)
2	291221	590,9073792	617				
3	291222	824,260505	999				
4	291222	278,594226	450				
5	291223	792,1827852	404				
6	291223	233,2116417	369				
7	291224	497,4058434	808				
8	291224	505,3892834	242				
9	291225	590,2600665	573				
10	291225	479,549527	372				
21	291231	709,1085748	955				
22	291231	698,7960109	494				
23	291231	468,8294026	501				
24	291232	284,7648354	233				
25	291232	495,0318488	483				
26	291232	623,0943033	790				
27	291233	638,2296495	1032				
28	291233	451,0698134	574				
29	291233	576,9972402	998				
30	291234	826,4823084	548				
31	291234	585,299949	672				
32	291234	380,8294884	608				
33	291232	1184,797378	969				
34	291232	842,3120039	899				
35	291233	678,3093541	823				
36	291233	971,6325473	605				
37	291234	195,1291222	389				
38	291234	608,1596073	654				
39				14,64	0	1	1
40				6,37	0	0	7
41				4,88	0	1	5
42				16,11	1	0	6
43				17,45	0	1	6
44				12,89	0	1	3
45				17,84	1	1	3
46				6,8	0	1	2
47				15,26	1	0	6



# Solution 4-1A+B

Example - Solution  
is presented

AUSGABE: ZUSAMMENFASSUNG								
<i>Regressions-Statistik</i>								
Multipler Korrelationskoeffizient	0,59788028							
Bestimmtheitsmaß	0,35746083							
Adjustiertes Bestimmtheitsmaß	0,3440746							
Standardfehler	193,989375							
Beobachtungen	50							
ANOVA								
		<i>Freiheitsgrade</i>	<i>(dratsummen</i>	<i>Quadratsumme</i>	<i>Prüfgröße (F)</i>	<i>F krit</i>		
Regression	1	1004907,25	1004907,25	26,7036172	4,5455E-06			
Residue	48	1806330,12	37631,8776					
Gesamt	49	2811237,37						
	<i>Koeffizienten</i>	<i>Standardfehler</i>	<i>t-Statistik</i>	<i>P-Wert</i>	<i>Untere 95%</i>	<i>Obere 95%</i>	<i>Untere 95,0%</i>	<i>Obere 95,0%</i>
Schnittpunkt	241,940695	73,2608184	3,3024569	0,00181511	94,6399473	389,241443	94,6399473	389,241443
Entfernung(km)	0,55100288	0,1066274	5,16755427	4,5455E-06	0,33661412	0,76539164	0,33661412	0,76539164

Prediction (553 km):  $241,91 + 0,551 \cdot 553 = 546,6452877$

# Simple selection of independent variables (simple feature selection) I/III

## Correlation

### Meaning of the correlation

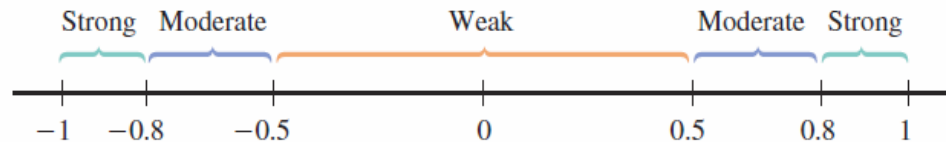
- a numerical evaluation of the strength of the relationship between variables

### Calculation of the correlation

- Pearson's sample correlation coefficient  $r$

$$r = \frac{\sum z_x * z_y}{n-1} \quad \text{with } z_x = \frac{x - \bar{x}}{s_x} \text{ and } z_y = \frac{y - \bar{y}}{s_y}$$

- Excel-Funktion: KORREL()



### Consideration in modeling

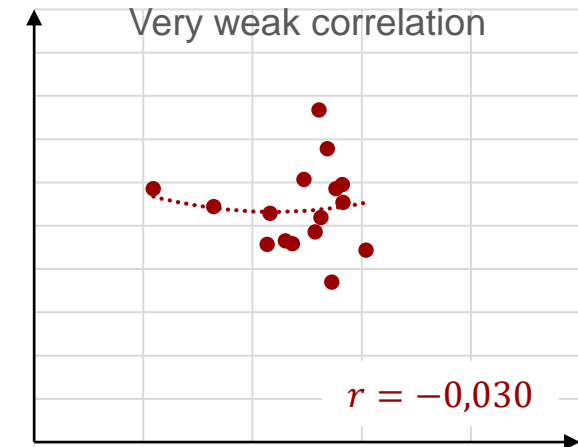
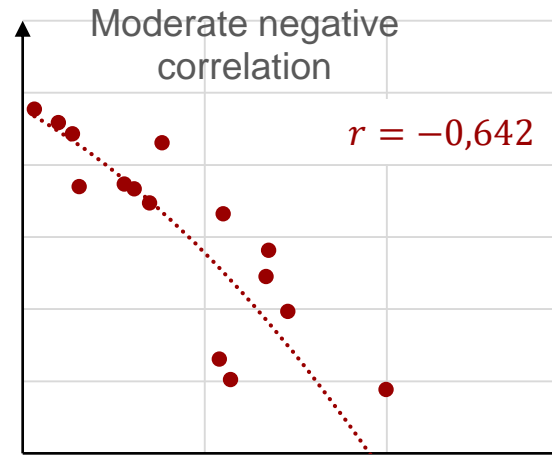
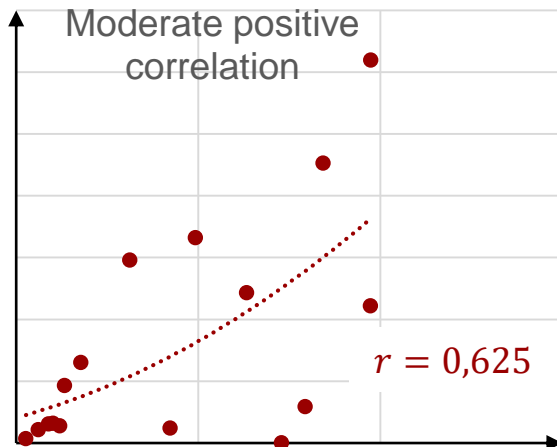
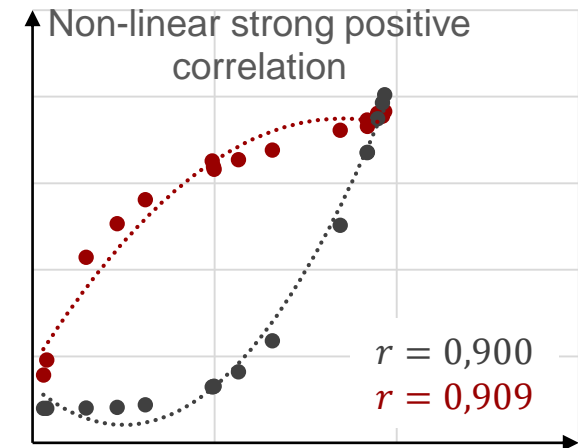
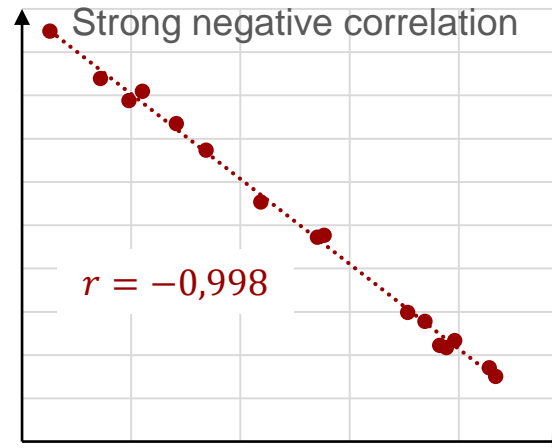
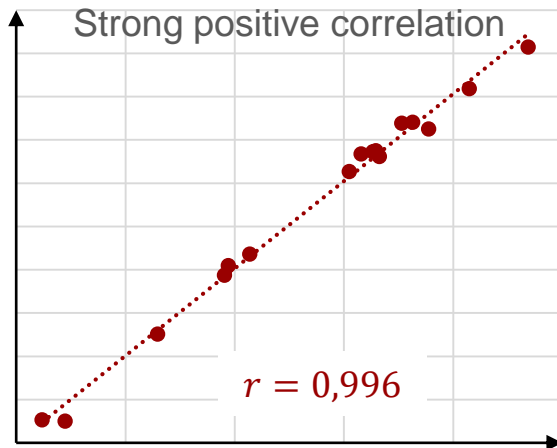
- Multicollinearity
  - Predominance of strong correlation between the independent variables of a regression model
  - tends to increase the uncertainty of beta factors
  - should be avoided → **independent variables should correlate only weakly with each other**

### Nomenclature

$\hat{y}$  – Prediction of the dependent variable  
 $y$  – Observation of the dependent variable  
 $x$  – Value of the independent value  
 $\alpha$  – Intersection with the Y-axis, when all  $x = 0$  („intercept“)  
 $\beta$  – Regression coefficient / slope of the dependent variable ("slope")  
 $e$  – Random error ("residuals")  
 $r$  – Correlation coefficient  
 $z$  – "z-score" of the observations of the variables  $x$  or  $y$   
 $\bar{x}$  – Estimated value of the mean value  
 $s$  – Standard deviation based on a sample  
 $n$  – Number of observations

# Simple selection of independent variables (simple feature selection) II/III

## Correlation Visualized



# Correlation is not causality

Correlation between two variables measures the statistical relationship

- No statement as to whether one variable causes the behavior of the other variable - that causality is present
- A third variable could affect both variables that is not captured, viewed or known

It is difficult to determine causal influences

- The classification of correlating variables as causal influences and the derivation of decisions based on this classification can have serious consequences
- Causal influences are determined by precise control of variable values

## Selection of independent variables - Ideally

- When creating a regression model, select independent variables for which a causality with regard to the dependent variable can be assumed (e.g. explainable by experts).
- Correlation can be an indicator and serve as a filter when selecting variables → however, causality should be questioned

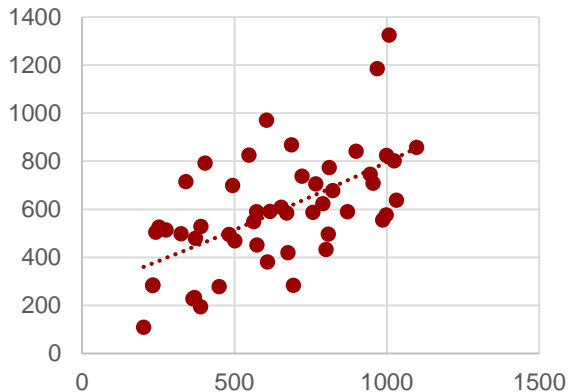
BEAR Food and Beverages supplies food to several customers in the region. A few weeks ago BEAR started to order a new LDL for transport. The LDL are paid after the order has been fulfilled, the invoice amount takes into account actual working time, fuel consumption, toll costs and other factors. They should develop a model for estimating transport costs and identify the influence of cost drivers in order to make the costs of LDL comparable with others.

- c) Visualize the relationship between the available independent variables and the dependent variable (transport costs)
- d) Calculate the correlations between the available independent variables and the dependent variable (transport costs)
- e) Add another variable to the model to improve the prediction

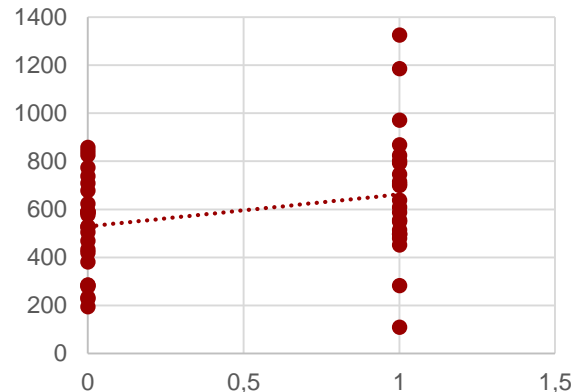
# Solution 4-1C+D

Example - Solution  
is presented

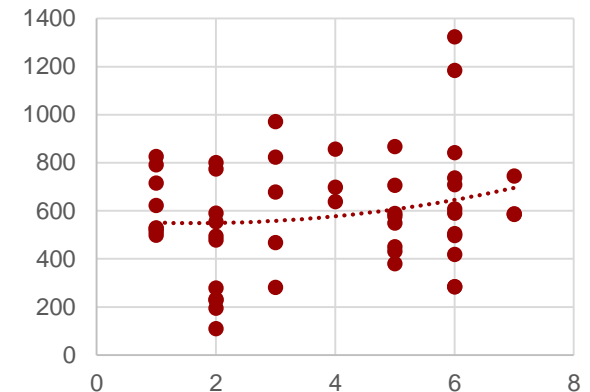
Distance(km)



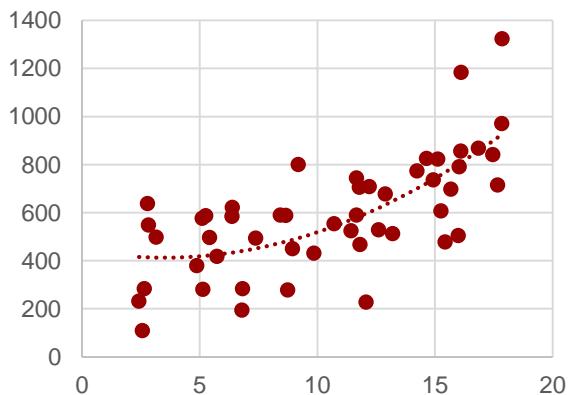
Cooled



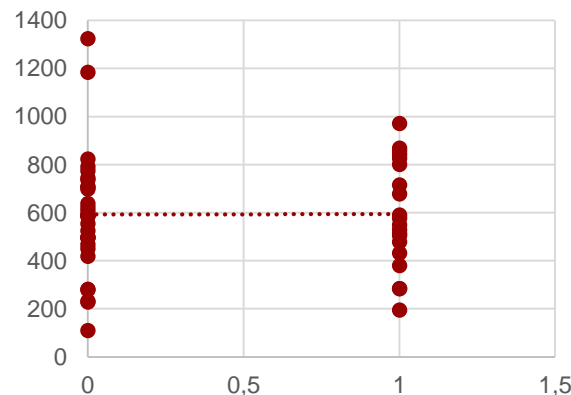
Day of the week(SO-SA)



Transport volume(t)



Toll route



Correlation	Transport costs
<i>Distance(km)</i>	0,597880283
<i>Transport volume(t)</i>	0,673773221
<i>Cooled</i>	0,28172568
<i>Toll route</i>	0,004099884
<i>Day of the week (SO-SA)</i>	0,186083212

# Solution 4-1E

Example - Solution is presented

	A	B	C	D	E	F	G	H	I
1	TransportNr	Transportkosten	Entfernung(km)	Transportvolumen(t)	Get				
2	291221	590,9073792	617	11,67					
3	291222	824,260505	999	15,12					
4	291222	278,594226	450	8,74					
5	291223	792,1827852	404	16,03					
6	291223	233,2116417	369	2,42					
7	291224	497,4058434	808	5,42					
8	291224	505,3892834	242	15,99					
9	291225	590,2600665	573	8,43					
10	291225	479,549527	372	15,43					
11	291226	709,1085748	955	12,21					
12	291226	698,7960109	494	15,68					
13	291227	468,8294026	501	11,81					
14	291227	284,7648354	233	6,82					
15	291228	495,0318488	483	7,38					
16	291228	623,0943033	790	6,39					
17	291229	638,2296495	1032	2,78					
18	291229	451,0698134	574	8,94					
19	291230	576,9972402	998	5,11					
20	291230	826,4823084	548	14,64					
21	291231	585,299949	672	6,37					
22	291231	380,8294884	608	4,1					

Regression

Y-Eingabebereich:

\$B\$1:\$B\$51

X-Eingabebereich:

\$C\$1:\$D\$51

☒ Beschriftungen
 ☐ Konstante ist Null

☐ Konfidenzniveau:
 

95 %

Ausgabe

☒ Ausgabebereich:
 

\$L\$1

☐ Neues Tabellenblatt:

☐ Neue Arbeitsmappe

Residuen

☐ Residuen
 ☐ Residuenplots

☐ Standardisierte Residuen
 ☐ Kurvenanpassung

☐ Normalverteilte Wahrscheinlichkeit
 

Quantilsplot

OK


Abbrechen

Hilfe

For the calculation of the model, the independent variables must be in adjacent columns!

Linear Regression 1 – 29.07.2018  
Quantitative Decision Making

– 31 –


  
TU  
berlin

# Evaluation of Linear Regression I/XI

## Simple methods

### Evaluation of the model ("Goodness of Fit")

Make general considerations before creating a model

- (1) Is a linear relationship appropriate to assume the relationship between the independent and dependent variables?
- (2) Do irregularities in the data have to be considered before the regression model can be created and used?

Visual inspection

- (1) Scatter cloud of residuals: Scatter diagram of (x, residuals) pairs
- (2) The point cloud should not show any patterns if possible!

Evaluation of key figures

- (1) What is the accuracy of the model's predictions? **MFE, MAE, MSE, MAPE**
- (2) Which values do the regression-specific key figures reach?
- (3) How does the forecast compare?

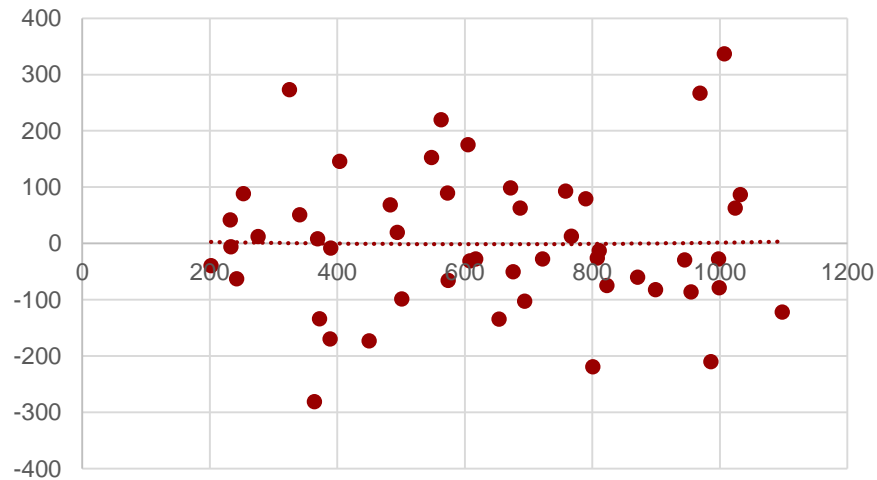


# Evaluation of Linear Regression II/XI

## Visual Evaluation

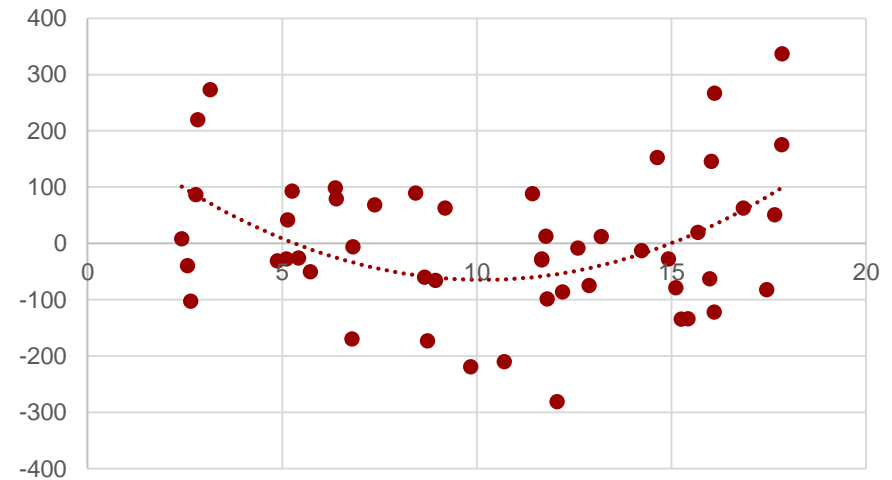
- Point cloud of residuals of problem 4-1E

Distance, residuals



⇒ No abnormality

Transport volume, residues



⇒ Initially speaks against linear correlation

# Forecast Error

## Mean Forecast Error (MFE)

- Average error.
- Positive and negative errors can balance each other out with this measure!

$$MFE = \frac{\sum_{t=k+1}^n e_t}{n - k}$$

## Mean absolute Error (MAE)

- Average absolute error
- Prevents compensation of positive and negative errors, but depending on scaling of the data and therefore difficult to compare
- All errors equally weighted

$$MAE = \frac{\sum_{t=k+1}^n |e_t|}{n - k}$$

## Nomenclature

### Values:

$y_t$  – observed value in  $t$   
 $e$  – Error

### Indices:

$n$  – Number of observations  
 $k$  – Number of observations, for which no prediction is possible  
 $t$  – Control variable of the sum

# Forecast Error

## Mean squared Error (MSE)

- Average square error
- Prevents compensation of positive and negative errors, but depending on scaling of the data and therefore difficult to compare
- Larger errors are weighted more heavily due to squaring

$$MSE = \frac{\sum_{t=k+1}^n e_t^2}{n - k}$$

The square error has a unit that cannot be interpreted, therefore the root mean squared error (RSME) is also used:

$$RSME = \sqrt{\frac{\sum_{t=k+1}^n e_t^2}{n - k}}$$

## Mean Absolute Percentage Error (MAPE)

- Average percentage error
- Enables comparability with different scaling
- All errors equally weighted

$$MAPE = \frac{\sum_{t=k+1}^n \left| \left( \frac{e_t}{y_t} \right) * 100 \right|}{n - k}$$

For observations with value 0, the percentage error cannot be calculated. Therefore the symmetric mean absolute percentage error (sMAPE) is also used:

$$sMAPE = \frac{1}{n - k} * \sum_{t=k+1}^n \frac{|e_t|}{\frac{y + \hat{y}}{2}}$$

## Nomenclature

### Values:

$y_t$  – observed value in  $t$   
 $e$  – Error

### Indices:

$n$  – Number of observations  
 $k$  – Number of observations, for which no prediction is possible  
 $t$  – Control variable of the sum

# Key figures: coefficient of determination

### Coefficient of determination ( $R^2$ )

- Measures "proportion of variability in the dependent variable", which can be explained by the "linear relationship between the dependent and the independent variables"
- Compares the (squared) residuals with the (squared) distance between the observation and the estimated mean of the dependent variable.

### Sum of Squares / Error Sum of Squares (ESS)

- Total squared distance between observations ( $y$ ) and predictions ( $\hat{y}$ )
- $$ESS = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \dots + (y_n - \hat{y}_n)^2 = \sum (y - \hat{y})^2$$

### Total Square Sum / Total Sum of Squares (TSS)

- Summed (squared) distance between observations ( $y$ ) and the estimated value of the mean value ( $\bar{y}$ )
- $$TSS = (y_1 - \bar{y}_1)^2 + (y_2 - \bar{y}_2)^2 + \dots + (y_n - \bar{y}_n)^2 = \sum (y - \bar{y})^2$$

### Calculation of the coefficient of determination

- $$R^2 = 1 - \frac{ESS}{TSS}$$

### Nomenclature

$\hat{y}$  – Prediction  
 $y$  – Observations  
 $\bar{y}$  – Estimated value of the mean value ( $y$ )  
 $x$  – Value of the independent value  
 $\bar{x}$  – Estimated value of the mean value ( $x$ )  
 $\alpha$  – Intersection with the Y-axis, when all  $x = 0$  („intercept“)  
 $\beta$  – Regression coefficient  
 $e$  – Random error ("residuals")  
 $r$  – Correlation coefficient  
 $z$  – "z-score" of the observations of the variables  $x$  or  $y$   
 $s$  – Standard deviation based on a sample  
 $n$  – Number of observations  
 $R^2$  – Coefficient of determination  
 $ESS$  – Sum of squares of the residuals  
 $TSS$  – Total square sum

# Key figures: Adjusted coefficient of determination

### Susceptibility of coefficient of determination

- In addition to the accuracy of the prediction, the aim is also to use as few independent variables as possible for the prediction
  - Data collection effort is reduced
  - Influence of independent variables can be generalized and interpreted
- By adding additional independent variables  $R^2$  always increases

### Effects of the adjusted $R^2$

- The adjusted coefficient of determination penalizes the addition of further independent variables
  - if the adjusted  $R^2$  increases as additional independent variables are added, this is an indicator of additional predictive performance of the additional variable
  - If the adjusted  $R^2$  decreases, this is an indicator of lack of predictive performance.
- $$\text{adjusted } R^2 = 1 - \left[ \frac{n-1}{n-(k+1)} \right] * \frac{ESS}{TSS}$$

### Nomenclature

$\hat{y}$  – Prediction  
 $y$  – Observations  
 $\bar{y}$  – Estimated value of the mean value ( $y$ )  
 $x$  – Value of the independent value  
 $\bar{x}$  – Estimated value of the mean value ( $x$ )  
 $\alpha$  – Intersection with the Y-axis, when all  $x = 0$  („intercept“)  
 $\beta$  – Regression coefficient  
 $e$  – Random error ("residuals")  
 $r$  – Correlation coefficient  
 $z$  – "z-score" of the observations of the variables  $x$  or  $y$   
 $s$  – Standard deviation based on a sample  
 $n$  – Number of observations  
 $R^2$  – Coefficient of determination  
 $ESS$  – Sum of squares of the residuals  
 $TSS$  – Total square sum

# Evaluation of Linear Regression V/XI

## Key figures: Calculated coefficient of determination

AUSGABE: ZUSAMMENFASSUNG		
Regressions-Statistik		
Multipler Korrelationskoeffizient	0,597880283	
Bestimmtheitsmaß	0,357460832	
Adjustiertes Bestimmtheitsmaß	0,3440746	
Standardfehler	193,9893749	
Beobachtungen	50	
ANOVA		
	Freiheitsgrade (df)	Quadratsummen (SS)
Regression	1	1004907,251
Residue	48	1806330,123
Gesamt	49	2811237,374
	Koeffizienten	Standardfehler
Schnittpunkt	241,9406951	73,26081837
Entfernung(km)	0,55100288	0,106627401

AUSGABE: ZUSAMMENFASSUNG		
Regressions-Statistik		
Multipler Korrelationskoeffizient	0,84658925	
Bestimmtheitsmaß	0,716713358	↗
Adjustiertes Bestimmtheitsmaß	0,704658607	↗
Standardfehler	130,1705915	
Beobachtungen	50	
ANOVA		
	Freiheitsgrade (df)	Quadratsummen (SS)
Regression	2	2014851,379
Residue	47	796385,9955
Gesamt	49	2811237,374
	Koeffizienten	Standardfehler
Schnittpunkt	-23,02277814	59,95436872
Entfernung(km)	0,476654664	0,072194213
Transportvolumen(t)	29,76094702	3,854880777

$$R^2 = 1 - \left( \frac{ESS}{TSS} \right) = 1 - \left( \frac{796385,9955}{2811237,374} \right) = 1 - 0,2833 = 0,7167$$

$$adjusted R^2 = 1 - \left[ \frac{n-1}{n-(k+1)} \right] * \frac{ESS}{TSS} = 1 - \left[ \frac{50-1}{50-(2+1)} \right] * 0,2833 = 1 - 1,0426 * 0,2833 = 0,7046$$

# Key figures: significance level („p-Value“)



*The significance level (p-Value) is a measure of the inconsistency between the hypothetical value of a population and the observed sample.*

### The level of significance in general

- A "test statistic" is calculated which compares the inconsistency between the null hypothesis  $H_0$  and the sample (inconsistency  $\rightarrow$  distance between the value of the null hypothesis and the actual characteristics of the sample).
- The "test statistics" assume a statistical distribution (usually normal distribution) and thus evaluate the inconsistency (e.g. how many standard deviations the value of  $H_0$  is away from the observed value).
- The p-value is the probability that the test statistics will reach this value if the null hypothesis were true.

### Significance level of a regression

- The null hypothesis  $H_0$  of a regression states that the beta values are zero ( $\beta_i=0, \forall i$ ), i.e. the independent variables used have no predictive value.
- The smaller the p-value, the more reliably this null hypothesis can be rejected

## Key figures: Significance level (calculated)

- P-Values from Problem 4-1E

AUSGABE: ZUSAMMENFASSUNG				
<i>Regressions-Statistik</i>				
Multipler Korrelationskoeffizient	0,84658925			
Bestimmtheitsmaß	0,716713358			
Adjustiertes Bestimmtheitsmaß	0,704658607			
Standardfehler	130,1705915			
Beobachtungen	50			
ANOVA				
	<i>Freiheitsgrade (df)</i>	<i>Quadratsummen (SS)</i>	<i>Mittlere Quadratsumme (MS)</i>	<i>Prüfgröße (F)</i>
Regression	2	2014851,379	1007425,689	59,45484685
Residue	47	796385,9955	16944,38288	
Gesamt	49	2811237,374		
	<i>Koeffizienten</i>	<i>Standardfehler</i>	<i>t-Statistik</i>	<i>P-Wert</i>
Schnittpunkt	-23,02277814	59,95436872	-0,384005013	0,702706562
Entfernung(km)	0,476654664	0,072194213	6,602394368	3,28138E-08
Transportvolumen(t)	29,76094702	3,854880777	7,720328783	6,67982E-10

⇒ Both variables seem significant



# Key figures: confidence intervals



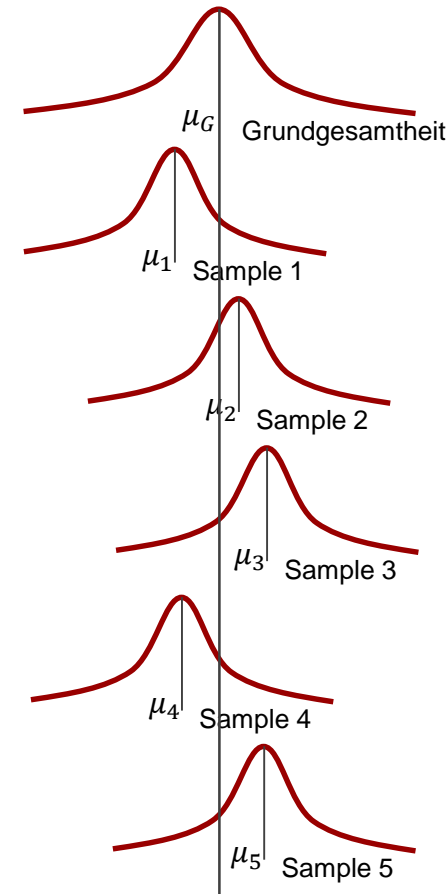
*The confidence interval is an interval of plausible values for a characteristic of the population of all study units.*

## Confidence intervals

- Different samples usually result in different estimates (different sets of observations of the same variable, e.g. at different times)
- The statistical methods serve to generate an estimated value that is as close as possible to the value of the "population of all study units"
- The values within a confidence **interval** around the estimated value are also plausible values.

Construction of the confidence interval:

- Depending on a confidence **level**, the actual value lies between the upper and lower limits of the interval



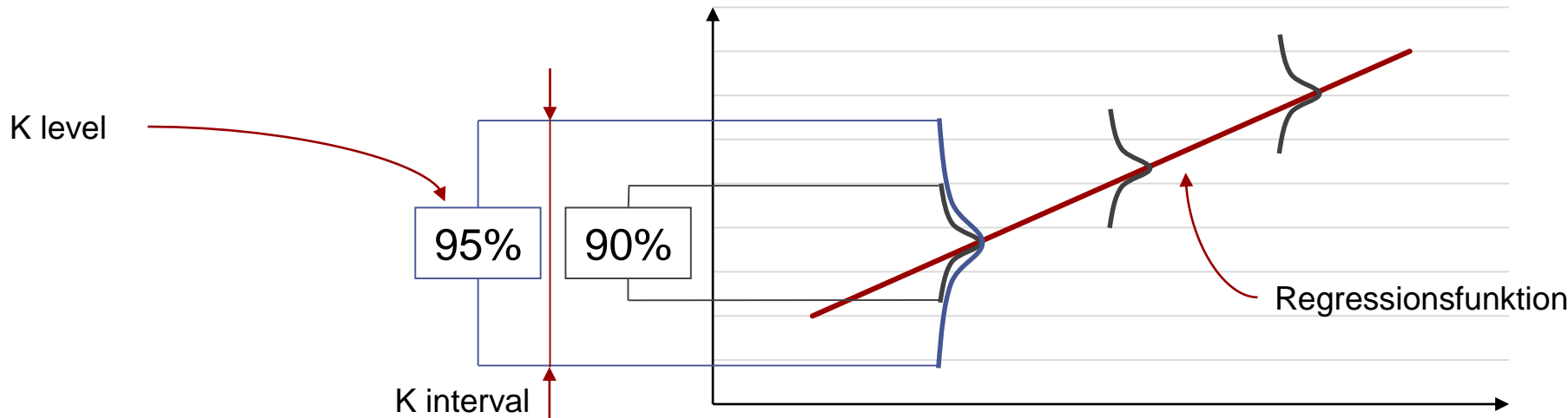
# Evaluation of Linear Regression IX/XI

## Key figures: confidence level

” The confidence level is the success rate of the method used to determine the confidence interval.

### Confidence level

- How much "confidence" exists in the method used to estimate the interval (not confidence in the estimated interval)
  - Example for 95% level: in an infinite number of samples, the actual value would lie in 95% of the generated intervals with the method used
- If the sample size increases, the confidence **interval** becomes smaller at a constant confidence **level**



# Evaluation of Linear Regression X/XI

## Key figures: Confidence Level (calculated)

### ■ Problem 4-1E Confidence Intervals

	Koeffizienten	Standardfehler	t-Statistik	P-Wert	Untere 95%	Obere 95%	Untere 80,0%	Obere 80,0%
Schnittpunkt	-23,02277814	59,95436872	-0,384005013	0,702706562	-143,635411	97,5898544	-100,9529623	54,90740602
Entfernung(km)	0,476654664	0,072194213	6,602394368	3,28138E-08	0,33141864	0,62189069	0,382814825	0,570494503
Transportvolumen(t)	29,76094702	3,854880777	7,720328783	6,67982E-10	22,0059272	37,5159668	24,75027681	34,77161722

	Intersection	Distance (km)	Transport capacity (t)
<i>P value</i>	0,702706562	3,28138E-08	6,67982E-10
<i>Upper 95%</i>	97,58985439	0,621890687	37,51596685
<i>Upper 80,0%</i>	54,90740602	0,570494503	34,77161722
<i>Coefficients</i>	-23,02277814	0,476654664	29,76094702
<i>Lower 80,0%</i>	-100,9529623	0,382814825	24,75027681
<i>Lower 95%</i>	-143,6354107	0,331418641	22,00592718

An interval of approx. 240 related to transport costs at 95% confidence level

⇒ Ø Distance = 637  
 $637 * 0,622 = 396$   
 $637 * 0,331 = 210$

An interval of approx. 186 related to transport costs at 95% confidence level

⇒ Ø Capacity = 10,49  
 $10,49 * 37,5 = 393$   
 $10,49 * 22,01 = 230$

An interval of approx. 163 related to transport costs at 95% confidence level

# Evaluation of Linear Regression XI/XI

## Baseline

The previous concepts perform the valuation in isolation to determine the added value of a model, however, a comparative concept is required - a baseline

- The baseline should be appropriate for the problem ("simple but not simplistic" → therefore random selection of the prediction value is often excluded)

### Appropriate baseline concepts

- Any form of data-driven model: e.g. model based on human intuition
- Time Series: e.g. Prediction = A Naive Model
- Classification: e.g. prediction = random choice or most frequent category
- Regression: e.g. prediction = median or mean value; prediction with regression model with few independent variables (e.g. the independent variable with highest correlation)
- Advanced techniques (e.g. neural networks): e.g. prediction with regression model
- Models with multiple data sources: e.g. prediction with only one data source
  - If data sources should generate recurring costs, these can be evaluated with it

# Features and Feature Engineering

## The term of the feature

” A **feature** is an "explanatory" input variable (independent variable, predictor) that uses an algorithm for prediction.

*"A piece of information that is potentially useful for prediction"*

*"The items, that represent this knowledge suitable for Machine Learning algorithms"*

### Features...

- ...represent a quantitative and machine-readable representation of the underlying characteristics of the prediction problem to be modeled
- ...are therefore strongly application-related
- ...often have to be obtained first by suitable transformation and combination of variables of the raw data

# Features and Feature Engineering

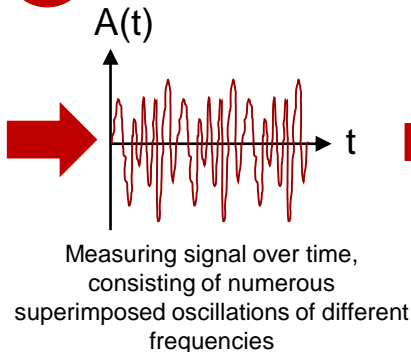
## Examples of Feature Generation

### Example 1: Predictive maintenance of electric motors

Damage to motor rolling bearings can be predicted by increased amplitudes at very specific critical frequencies. The current amplitudes of these critical frequencies must first be extracted from the vibration data.

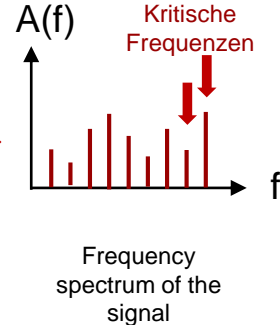
Measurement  
of engine  
vibrations

1



Transformation of the  
signal into frequency  
range (Fast Fourier)

2



Amplitudes for  
critical frequencies

3

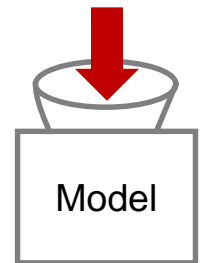
$$A(f_1) = 1,5$$
$$A(f_2) = 0,7$$

Division by engine  
speed during  
measurement

4

$$M_1 = \frac{A(f_1)}{n} = 0,03$$
$$M_2 = \frac{A(f_2)}{n} = 0,014$$

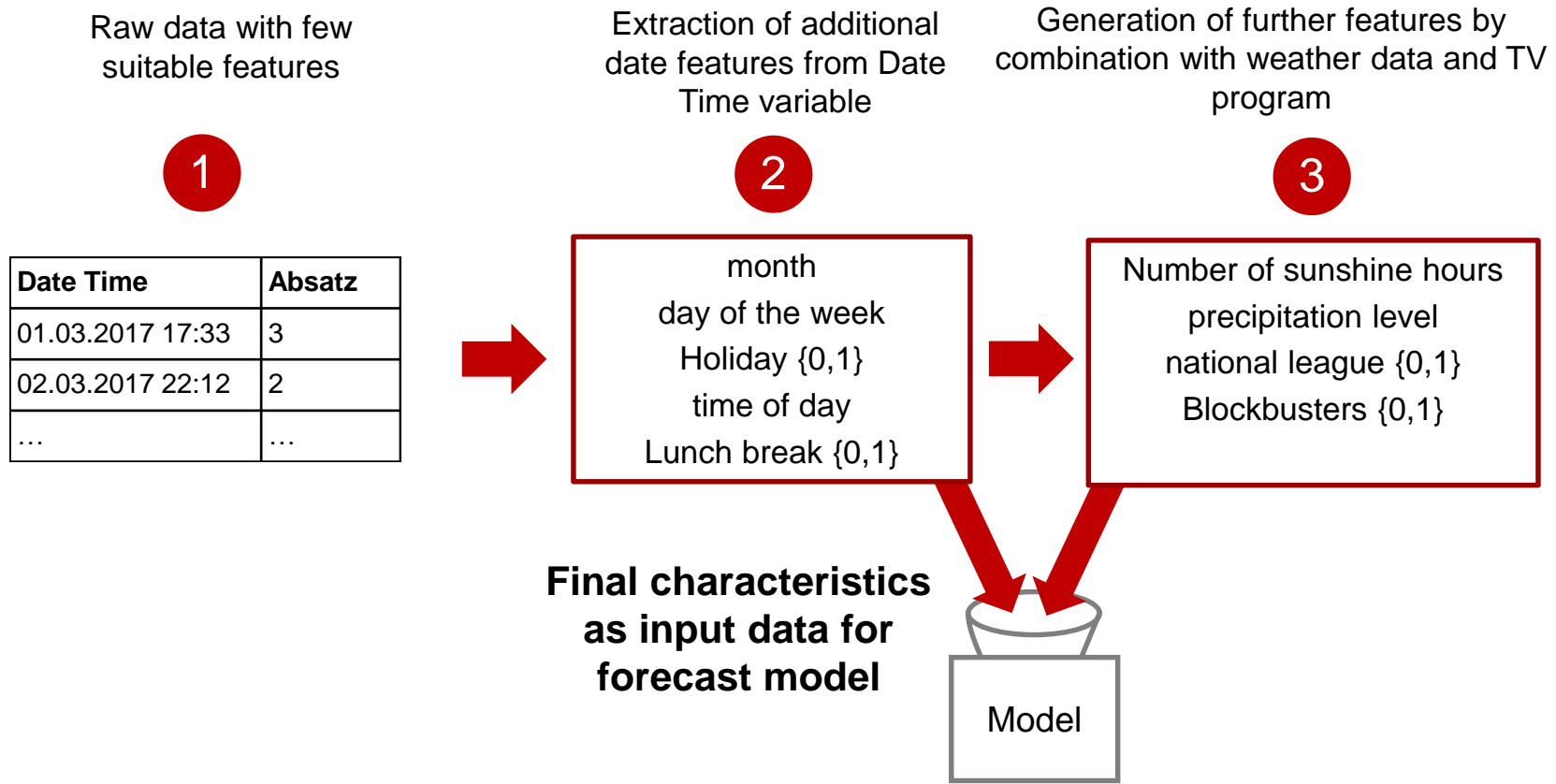
Final characteristics  
as input data for  
forecast model



# Features and Feature Engineering

## Examples of Feature Formation

**Example 2: Forecast of sales figures of an eCommerce retailer**  
**Customers prefer to shop when they are free - especially in bad weather.**



# Features and Feature Engineering

## Feature Engineering

” **Feature engineering** is the process of transforming raw data into characteristics that better represent the underlying problem of the model and thus lead to better forecast results.

### Feature Engineering ...

- ...is fundamental for building good and understandable prediction models ("Actually the success of all Machine Learning algorithms depends on how you present the data.")
- ...is a largely manual, iterative and time-consuming process
- ...requires in-depth knowledge of data, field of application and ML procedures

### Why is feature engineering important?

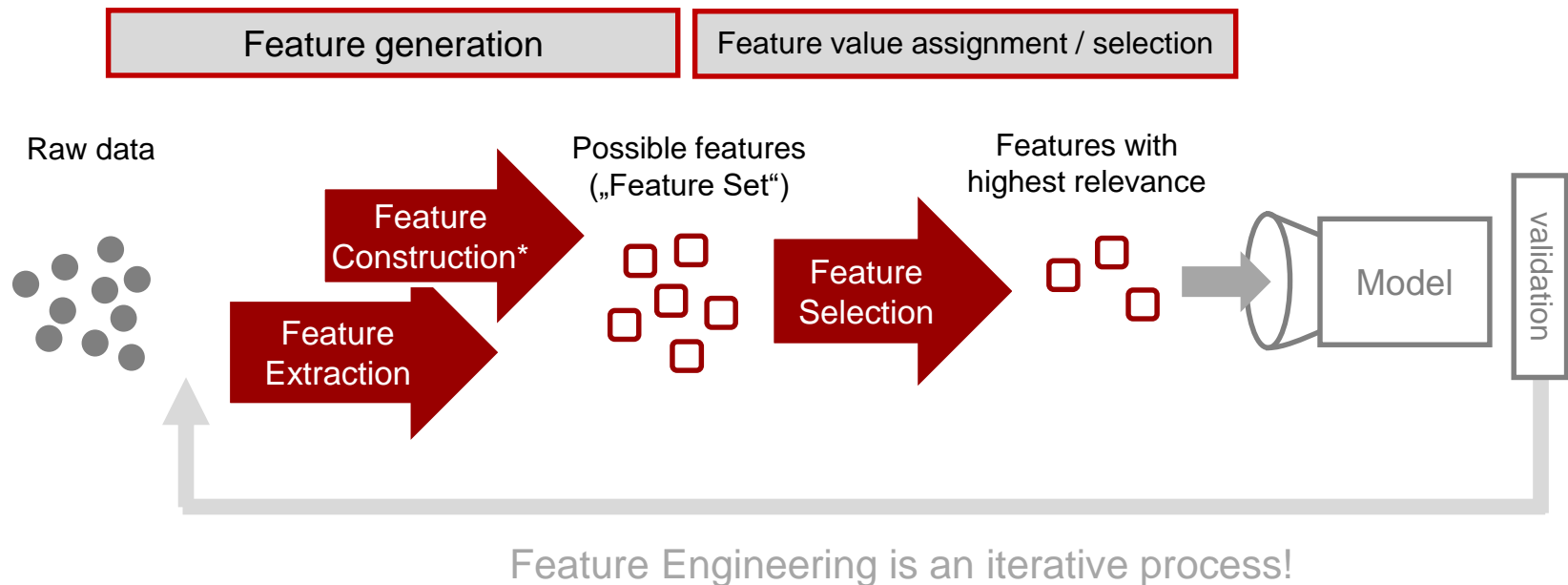
- Raw data often not in a suitable form to be interpreted by forecast models
- Raw data often do not represent the problem to be solved in a sufficiently meaningful way
- Process-specific requirements for data structure (e.g. categorical or binary variables)
- Not all variables in the raw data set are relevant for forecast
- Raw data often too complex or too heterogeneous (e.g. images, audio, texts, sensor data)



# Features and Feature Engineering

## Process of Feature Engineering

Feature Engineering comprises several consecutive sub-processes:



\*Feature Engineering does not have a clear definition. Feature engineering in the narrower sense is often understood above all as "feature construction".

# Features and Feature Engineering

## Feature Construction

” A **feature construction** is the manual generation of characteristics by transforming the raw data including specific knowledge about the application area.

### Insertion of:

- Intuition
- Creativity
- Logic

### Necessary knowledge about:

- Raw data on which it is based
- Application field (characteristics from problem context)
- Process-specific requirements for input data

Raw data



Feature Construction




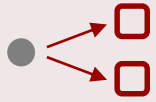

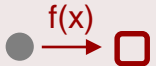

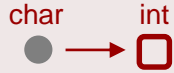
Features

### Supporting methods:

- Brainstorming of features
- Data visualization (e.g. scatterplots)
- Correlation analyses (e.g. correlation coefficients...)

# Features and Feature Engineering

## Feature Construction

Transformation		Examples
Combining variables		<ul style="list-style-type: none"> <li>▪ Addition, multiplication etc. of numerical variables</li> <li>▪ Linking strings to a string</li> </ul>
Split Variables/ Decomposing		<ul style="list-style-type: none"> <li>▪ Splitting a date into day, month, year</li> <li>▪ Extracting the owner key from the container number</li> <li>▪ Splitting a categorical variable into several binary variables</li> </ul>
Aggregating Variables		<ul style="list-style-type: none"> <li>▪ Combining table rows with averages or totals</li> </ul>
Transform variables		<ul style="list-style-type: none"> <li>▪ Mathematical transformation (logarithmicizing, exponentiation...)</li> <li>▪ Statistical transformation (mean value, variance...)</li> <li>▪ Fourier or laplacet transformation</li> </ul>
Normalize / standardize variables		<ul style="list-style-type: none"> <li>▪ Scaling to a uniform value range, e.g. (0.1)</li> <li>▪ Standardization to standard normal distributions Variable</li> <li>▪ Standardization of non-uniform into uniform strings</li> </ul>
Change Data Type		<ul style="list-style-type: none"> <li>▪ Conversion of continuous variables into categorical, discrete characteristics</li> <li>▪ Conversion of strings into numeric variables</li> </ul>

### Binary Variables

A certain condition influences the dependent variable (e.g. refrigerated transports, weekend trips,...)

- This condition can be formulated as a binary variable:

- $$x \begin{cases} 1, & \text{if the condition is true} \\ 0, & \text{otherwise} \end{cases}$$

### Factor variables

A certain value of a variable with several conditions (categorical variable, "factor variable") ( $f$ ) influences the dependent variable (for example, product color).

- The categories are mutually exclusive:  $f-1$  variables ("dummy variables") are required (if  $f-1$  states do not apply, state  $f$  applies)
- The categories are not mutually exclusive:  $f$  variables ("dummy variables") are required

### Nomenclature

$\hat{y}$  – Prediction

$y$  – Observations

$\bar{y}$  – Estimated value of the mean value ( $y$ )

$x$  – Value of the independent value

$\bar{x}$  – Estimated value of the mean value ( $x$ )

$\alpha$  – Intersection with the Y-axis, when all  $x = 0$  („intercept“)

$\beta$  – Regression coefficient

$e$  – Random error ("residuals")

$n$  – Number of observations

$f$  – Number of categories of an independent variable

$R^2$  – Coefficient of determination

ESS – Sum of squares of the residuals

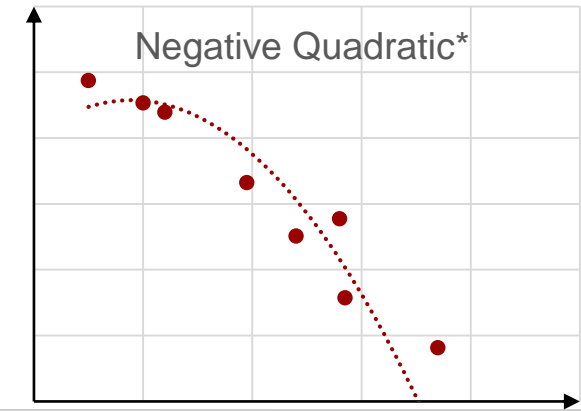
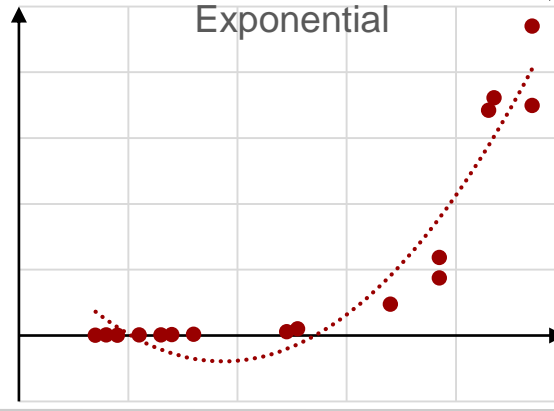
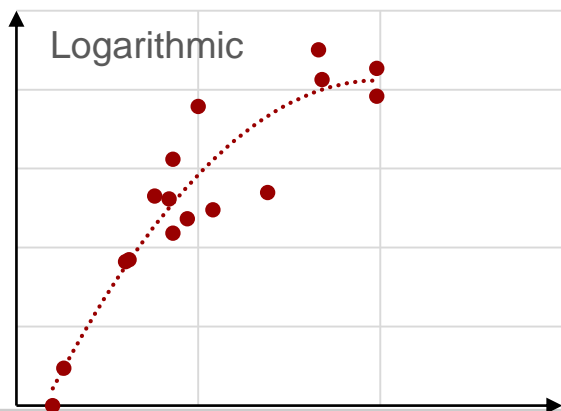
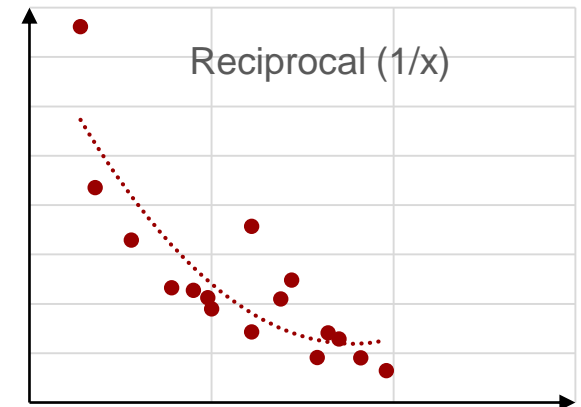
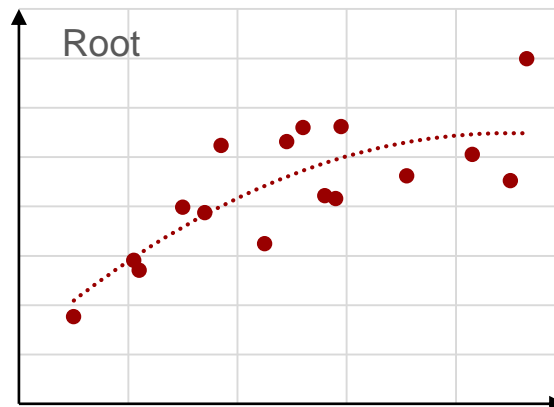
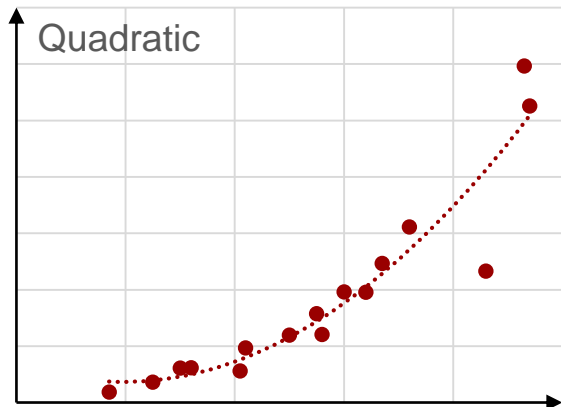
TSS – Total square sum

## Regression with non-linear influences - Continuous non-linear influences

Examples of simple continuously non-linear influences

- Meaning:  $y$  behaves non-linear for values of  $x$

\* representative of the fact that all interrelationships can be negative



# Regression with non-linear influences - Transformation of continuous variables

## Linear regression can only calculate linear dependencies

However, new variables can be created that approximate non-linear relationships into a linear one:

→ If y behaves non-linear, x has to behave non-linear in the same manner!

- Square relationship:  $\hat{y} = \alpha + \beta_1 * x'$  with  $x' = x^2$
- Root relationship:  $\hat{y} = \alpha + \beta_1 * x'$  with  $x' = \sqrt{x}$
- Reciprocal relationship:  $\hat{y} = \alpha + \beta_1 * x'$  with  $x' = \frac{1}{x}$
- Logarithmic relationship:  $\hat{y} = \alpha + \beta_1 * x'$  with  $x' = \ln(x)$
- Exponential relationship:  $\hat{y} = \alpha + \beta_1 * x'$  with  $x' = e^x$

## Interaction of independent variables (e.g. multiplicative relationship between two or more independent variables:

This can be approximated by a new variable that maps the interaction

- multiplicative relationship :  $\hat{y} = \alpha + \beta_1 * x'$  with  $x' = x_1 * x_2$

### Nomenclature

$\hat{y}$  – Prediction

y – Observations

$\bar{y}$  – Estimated value of the mean value (y)

x – Value of the independent value

$x'$  – Transformed variable

$\bar{x}$  – Estimated value of the mean value (x)

$\alpha$  – Intersection with the Y-axis, when all x = 0 („intercept“)

$\beta$  – Regression coefficient

e – Random error ("residuals")

n – Number of observations

f – Number of categories of an independent variable

$R^2$  – Coefficient of determination

ESS – Sum of squares of the residuals

TSS – Total square sum

# Features and Feature Engineering

## Regression with non-linear influences – A simple way to transform continuous variables

### Tukey's Ladder of Transformations

Moving along the ladder

- If the bulge of the function goes to the left, go left of the ladder
- If the bulge of the function goes to the right, go right of the ladder

Applying the ladder movement

- Either use  $y = x^\lambda$  or substitute the formula for  $y = x'$  (the transformation would be the same for most cases)
- $\lambda$  does not have to jump in the steps below, more gradual movement is recommended!

$\lambda$	-2	-1	-0,5	0	0,5	1	2	
$x'$	$\frac{-1}{x^2}$	$\frac{-1}{x}$	$\frac{-1}{\sqrt{x}}$	$\log(x)$	$\sqrt{x}$	$x$	$x^2$	$e^x$

Starting point

Not included in the original ladder

### Nomenclature

$\hat{y}$  – Prediction  
 $y$  – Observations  
 $\bar{y}$  – Estimated value of the mean value ( $y$ )  
 $x$  – Value of the independent value  
 $x'$  – Transformed variable  
 $\bar{x}$  – Estimated value of the mean value ( $x$ )  
 $\alpha$  – Intersection with the Y-axis, when all  $x = 0$  („intercept“)  
 $\beta$  – Regression coefficient  
 $e$  – Random error ("residuals")  
 $n$  – Number of observations  
 $f$  – Number of categories of an independent variable  
 $R^2$  – Coefficient of determination  
 $ESS$  – Sum of squares of the residuals  
 $TSS$  – Total square sum

# Features and Feature Engineering

## Feature Construction



**Suitability of features is strongly determined by the selected forecasting method**

**Features scaled differently:** Support Vector Machines, artificial neural networks and regressions can hardly handle features with different value ranges

**Steady characteristics:** Due to their discrete rules and regulations, decision trees and random forests often perform worse with continuous input variables than artificial neural networks, for example

**Categorical variables:** Decision trees and Random Forest may perform better when categorical variables are broken down into multiple binary variables

**Redundant features:** Linear and logistic regressions as well as artificial neural networks are not robust against strongly correlated input variables

**Noise:** Decision trees can handle noisy variables very well



# Features and Feature Engineering

## Feature Extraction

- Method for automatically generating features from raw data
- Application with very extensive, unstructured or unquantified raw data such as:
  - Very extensive tables
  - Image and audio data
  - text data
  - sensor data
- The aim is to reduce dimensions and generate features of smaller dimensions that are more manageable for the model.
- Various methods for automatic dimension reduction and extraction of features of smaller dimensions

### Examples of procedures:

- Main Component Transformation (PCA)
- Discriminant Analysis (LDA)
- Factor analysis
- Edge detection for image data

Original data	Possible features
Time series	<ul style="list-style-type: none"><li>▪ Minima, Maxima</li><li>▪ tendencies</li><li>▪ seasonalities</li></ul>
Image data	<ul style="list-style-type: none"><li>▪ RGB color values</li><li>▪ contours</li></ul>
Audio and vibration data	<ul style="list-style-type: none"><li>▪ signal cut-outs</li><li>▪ frequency range</li><li>▪ amplitudes</li></ul>
Text data	<ul style="list-style-type: none"><li>▪ word frequencies</li><li>▪ Word positions in text</li></ul>

# Features and Feature Engineering

## Feature Selection

” A feature selection is a process that is used to select the characteristics with the highest relevance for a forecast model and remove irrelevant and redundant characteristics.

### Why is Feature Selection important?

- Negative influence on prediction quality: Many prediction methods have only little robustness against too many or unsuitable input variables
- Long training duration for models with many input variables and possibly additional costs by measuring these parameters (e.g. predictive maintenance - sensors)
- Interpretability decreases with model complexity
- If necessary, higher amount of training data necessary

### Methods:

Wrapper methods

Filter Methods

Integrated Feature  
Selection

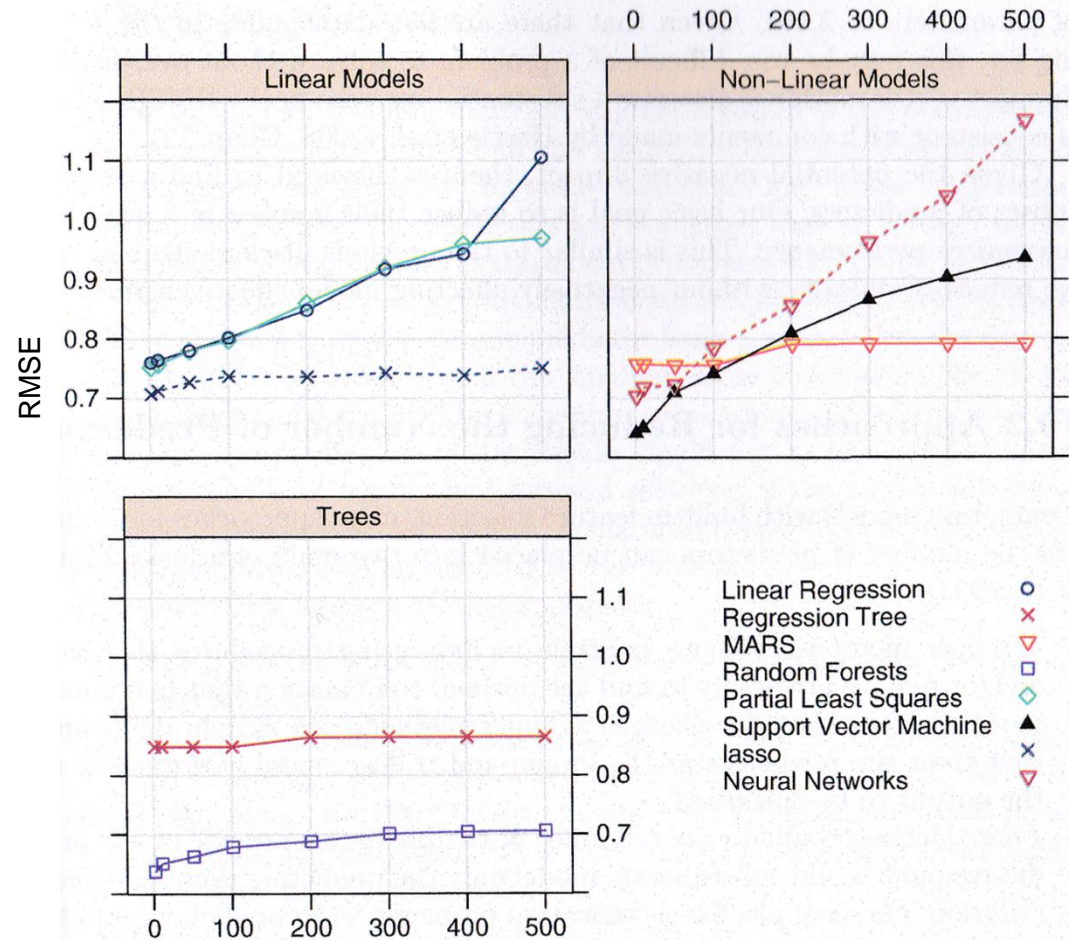
# Feature Selection - Importance for Forecast Quality

Some forecasting methods cannot handle too many or unsuitable input variables with a negative influence on the quality of the forecast.

Example:

- Linear regression: Includes all features built into the model and thus also unsuitable features
- Regression tree: Implicitly selects those characteristics that lead to the highest forecast quality and excludes unsuitable characteristics

Development of RMSE with successive addition of non-informative input variables for selected methods



# Features and Feature Engineering

## Feature Selection – Wrapper-Methods

### General principle (in principle the model selection process):

- Identification of the optimal characteristics is carried out by successively running through several sub-models with different combinations of characteristics and validated with regard to their prediction quality (e.g. MAPE, MSE/RMSE, p-value).
- The combination of characteristics with the highest prediction quality is used as the configuration for the final model.

### Variants for Characteristic Selection:

**Forward Selection:** Starting with a characteristic, another characteristic is included in the model in each iteration and the quality is evaluated until all characteristics have been tested; if the model quality has increased, the characteristic remains in the model

**Backward Selection:** Starting with all characteristics in the model, those with the least relevance are successively removed

**Stepwise Selection:** Special form of forward selection with repeated validation of each parameter in the final model

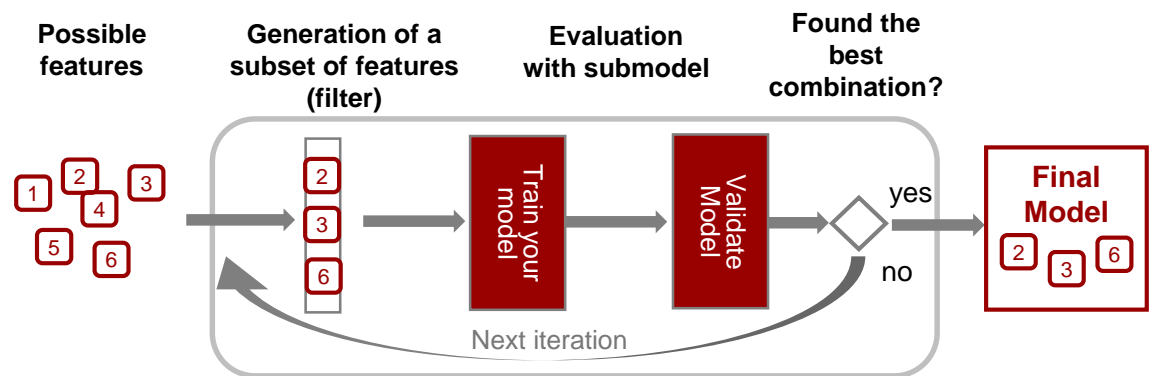
**Random Selection:** Different randomly selected combinations of characteristics are tested iteratively

#### Advantages:

- Evaluates the actual predictive quality of the model
- Characteristic is valued in context with other characteristics in the model
- Prevention of redundant features

#### Disadvantages:

- long calculation time, as a high number of models must be estimated



# Features and Feature Engineering

## Feature Selection – Filter-Methods

### General principle:

- The variables are evaluated and selected independently and in advance of the model estimation
- A measure of suitability (usually univariate) is determined separately for each characteristic
- Characteristics are ranked according to suitability level and then the  $n$  variables with the highest suitability are selected

### Advantages:

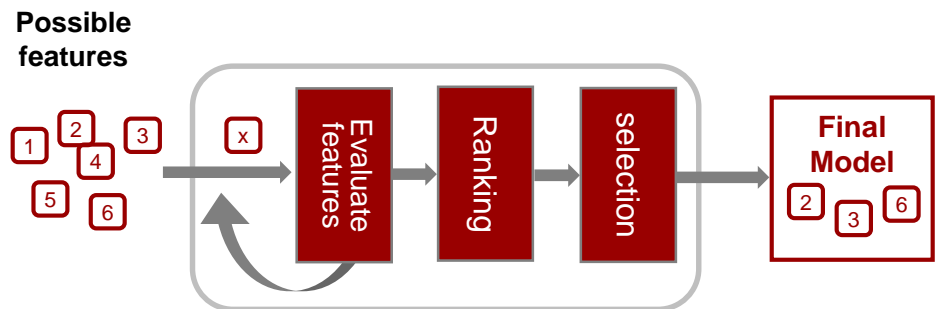
- less computationally intensive and methodically simple (common method in practice)

### Disadvantages:

- Final quality of the model is not evaluated
- Variables are not checked in context to each other, so redundancies between variables may not be detected

### Possible suitability dimensions:

- Correlation coefficient (see introduction)
- ANOVA
- Hypothesis tests for significance (e.g. t-test)



# Features and Feature Engineering

## Feature Selection – Integrierte Verfahren

### General principle:

- Some classification and forecast procedures have an implicit characteristic value assignment
- In the context of the model estimation, characteristics with low relevance for the prediction are ignored

Procedure	Integr. Feature Selection
Linear regression	✗
Logistic regression	✗
Artificial neural networks	✗*
Partial Least Squares	✓
Support Vector Machines	✗
K-Nearest Neighbors	✗
Decision trees	✓
Random Forest	✓
Gradient Boosted Machines	✓
Naive Bayes Classifier	✗

✓ integrated

✗ not integrated

\* Approaches for integration exist

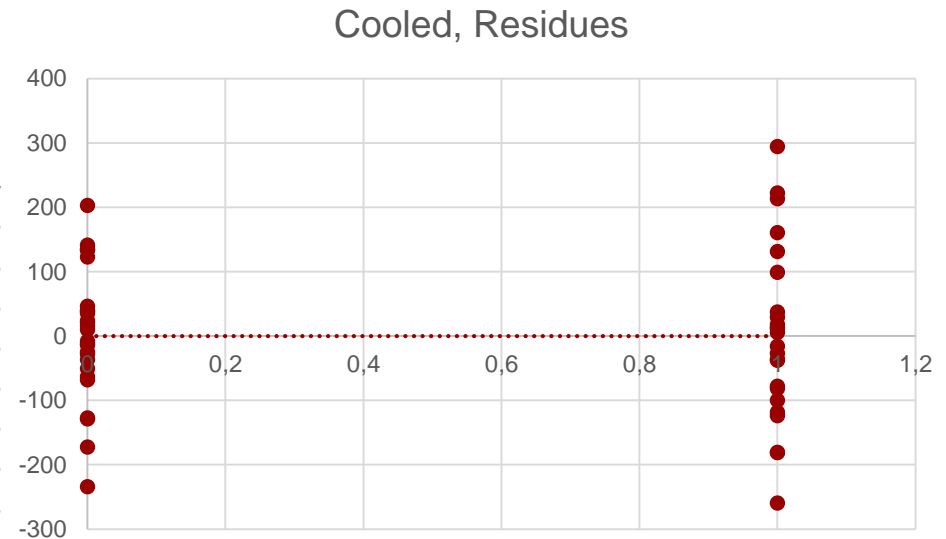
BEAR Food and Beverages supplies food to several customers in the region. A few weeks ago BEAR started to order a new LDL for transport. The LDL are paid after the order has been fulfilled, the invoice amount takes into account actual working time, fuel consumption, toll costs and other factors. They should develop a model for estimating transport costs and identify the influence of cost drivers in order to make the costs of LDL comparable with others.

- f) Add the following variable(s) to the model and evaluate the prediction model
  - „cooled“
- g) Transform the following variable(s) and evaluate the prediction model
  - „Transport volume“
- h) Transform the following variable(s) and evaluate the prediction model
  - „cooled“
- i) Add the following variable(s) to the model and evaluate the prediction model
  - „Toll“

# Solution 4-1F

Example - Solution is presented

AUSGABE: ZUSAMMENFASSUNG	
Regressions-Statistik	
Multipler Korrelationskoeffizient	0,871086034
Bestimmtheitsmaß	0,758790879 ↗
Adjustiertes Bestimmtheitsmaß	0,743059849 ↗
Standardfehler	121,4134238
Beobachtungen	50



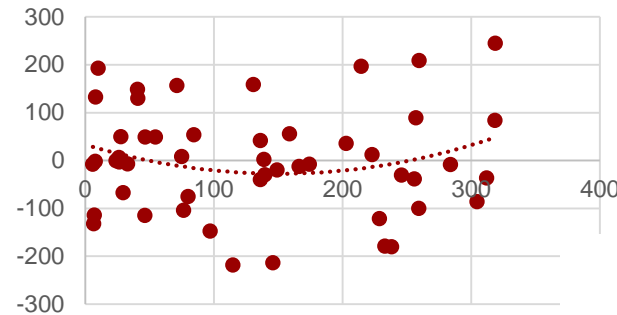
ANOVA				
	Freiheitsgrade (df)	Quadratsummen (SS)	Mittlere Quadratsumme (MS)	Prüfgröße (F)
Regression	3	2133141,278	711047,0928	48,23529655
Residue	46	678096,0958	14741,21947	
Gesamt	49	2811237,374		
	Koeffizienten	Standardfehler	t-Statistik	P-Wert
Schnittpunkt	-57,16901047	57,20538711	-0,999364105	0,322846568
Entfernung(km)	0,472688564	0,067351934	7,018188464	8,55746E-09
Transportvolumen(t)	28,7793313	3,612205185	7,967247104	3,30537E-10
Gekühlt	97,86379063	34,54736337	2,832742678	0,006827913



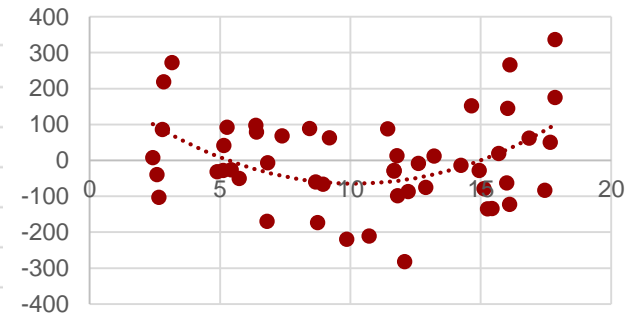
# Solution 4-1G

Example - Solution is presented

Transp.vol^2, Residues



Transport volume, Residues



## AUSGABE: ZUSAMMENFASSUNG

### Regressions-Statistik

Multipler Korrelationskoeffizient	0,886727213
Bestimmtheitsmaß	0,78628515
Adjustiertes Bestimmtheitsmaß	0,772347225
Standardfehler	114,2844634
Beobachtungen	50



## ANOVA

	Freiheitsgrade (df)	Quadratsummen (SS)	Mittlere Quadratsumme (MS)	Prüfgröße (F)
Regression	3	2210434,2	736811,4001	56,41335776
Residue	46	600803,1741	13060,93857	
Gesamt	49	2811237,374		

	Koeffizienten	Standardfehler	t-Statistik	P-Wert
Schnittpunkt	56,82117088	47,79231294	1,188918623	0,240573108
Entfernung(km)	0,477184616	0,063293736	7,539207646	1,42444E-09
Transp.vol^2	1,454047459	0,165103546	8,80688207	1,96901E-11
Gekühlt	79,6733638	32,76256754	2,431841268	0,018974057

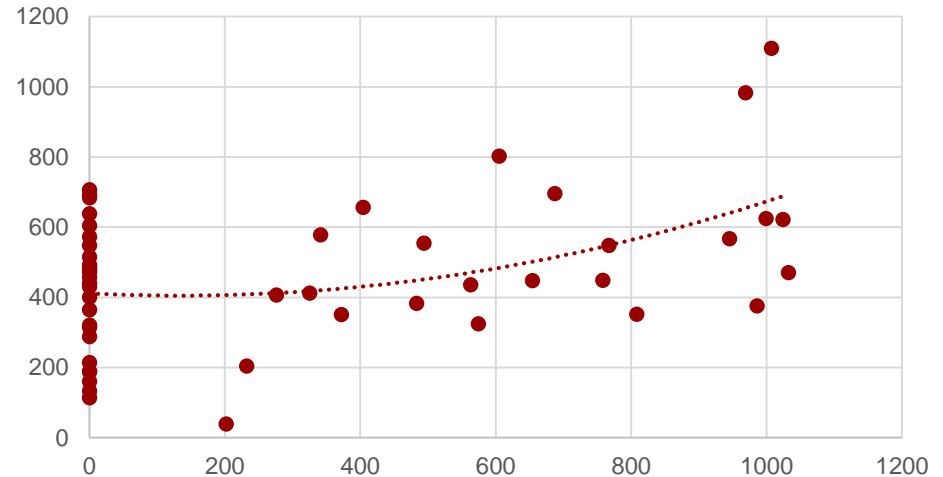
# Solution 4-1H

Example - Solution is presented

AUSGABE: ZUSAMMENFASSUNG	
Regressions-Statistik	
Multipler Ko	0,887579396
Bestimmthe	0,787797184
Adjustiertes	0,77395787
Standardfeh	113,879464
Beobachtung	50




Cooled\*Dist, Residues



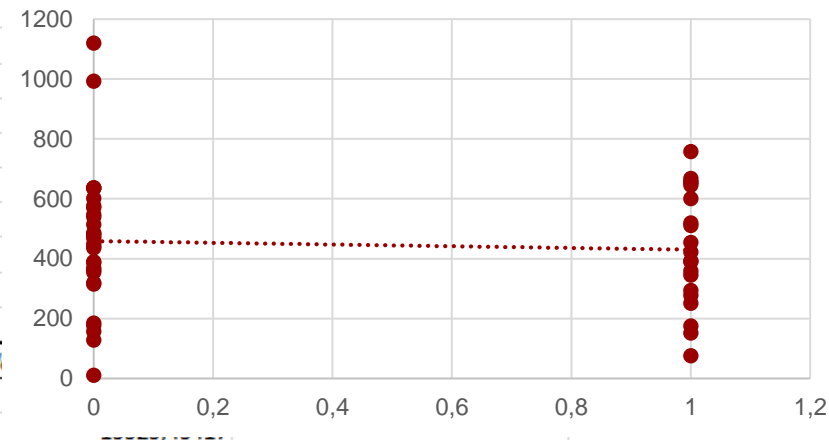
ANOVA				
	Freiheitsgrade (df)	Quadratsummen (SS)	mittlere Quadratsumme (MS)	Prüfgröße (F)
Regression	3	2214684,888	738228,2959	56,92458314
Residue	46	596552,4864	12968,53231	
Gesamt	49	2811237,374		
	Koeffizienten	Standardfehler	t-Statistik	P-Wert
Schnittpunkt	98,67222	46,47455179	2,123145167	0,039153339
Entfernung(l	0,412013505	0,068565924	6,009012685	2,80707E-07
Transp.vol^2	1,461927402	0,163972072	8,915709744	1,37305E-11
Gek*Entf	0,118851316	0,047412647	2,506742907	0,01578152

# Solution 4-11

Example - Solution  
is presented

AUSGABE: ZUSAMMENFASSUNG		
Regressions-Statistik		
Multipler Ko	0,886977626	
Bestimmthe	0,786729309	
Adjustiertes	0,767771914	
Standardfeh	115,4271813	
Beobachtung	50	
ANOVA		
	Freiheitsgrade (df)	Quadratsummen (SS,tl)
Regression	4	2211682,836
Residue	45	599554,5379
Gesamt	49	2811237,374
	Koeffizienten	Standardfehler
Schnittpunkt	62,21473659	51,38499858
Entfernung(l	0,474909615	0,064357102
Transp.vol^2	1,466043648	0,171296788
Mautstrecke	-10,76365724	35,16009112
Gekühlt	77,09316243	34,14668286

Toll, Residues



⇒ The legal department has informed you no longer to use the new LDL because it is under investigation for evasion of the toll.

# Overfitting

## Overfitting

- Adapt a model to a sample/data set, so that the data set can be estimated particularly well, but does not achieve realistic results outside the sample.
- Cause: The model contains systematic components for the random error

## Solving

- In linear regression, pay particular attention to the development of the adjusted  $R^2$
- Division\* into training (80% of observations) and test set (20% of data) and comparison of error values
  - Special case cross-validation: The data set is divided into  $K$  containers (folds).  $K-1$  vessels are used as training and one vessel as a test set. This is repeated until all combinations of vessels have been validated as a model. The results are finally averaged.

\*Note: This does not work for predecessor and successor dependencies of observations (e.g. demand of previous period or what a customer last bought). Accordingly, this is only cumbersome for time series analyses.

## Problem 4-2

Capital Bike Sharing (CBS) from Washington, D.C., USA provides bicycles as public transport. These bicycles are supposed to relieve other means of transport. Therefore, the expected demand for bicycles should be predicted in order to adjust the planning of the other means of transport if necessary. The demand of flexible users in particular is to serve as an indicator.

### ■ Data Set

- instant – Index
- dteday – Day
- season - season (1 = winter, 2 = spring, 3 = summer, 4 = autumn)
- yr - year (0 = 2011, 1 = 2012)
- mnth - month (1 = January ...)
- holiday - holiday yes/no
- weekday - weekday (0 = Sunday, ...)
- workingday - working day yes/no
- weathersit - Weather conditions (1 = clear to partly cloudy, 2 = foggy and cloudy, 3 = light snow, light rain and thunderstorm, 4 = heavy snow, heavy rain and thunderstorm)
- temp - normalized temperature (0 = -8°C, ..., 1 = 39°C)
- atemp - normalized temperature (0 = -16°C, ..., 1 = 50°C)
- hum - Humidity
- windspeed - Wind speed
- casual - number of unregistered users
- registered - number of registered users
- cnt - sum of registered and unregistered users

## Problem 4-2

- a) Load the data from "04\_Problem Set 4-2.csv" in Excel and determine the correlations of the variables of time and weather (temp, humidity,...) in connection to the variable casual.
- b) Create a point cloud of the most strongly correlating variable and compare the linear model with an adapted model with a better linear relationship, whose relationship you estimate from the point cloud. Continue with the more suitable model.
- c) Add the variables workingday and hum to the best model in b) and rate the models. Continue with the more suitable model.
- d) Look at the point cloud from hr to casual. It is difficult to represent the connection mathematically. Therefore, create several factor variables for 4 day segments instead of one linear variable. Add the new factor variable to the best model in c) and rate the model.

## Problem 4-3

BEAR Electronics receives various components from its supplier Cybernetics Corp in the fields of switching technology, automation technology and drive technology. Cybernetics Corp's delivery reliability is 96.5%. However, the shortfalls in recent months have often led to production problems. Cybernetics has announced that it is working on the problem, but this has not yet led to an improvement. The planner has started to add a flat rate of 5% to all orders, but this has not solved the problem. You should now try to predict the shortfall using a model. Use the April orders in the data record "04\_Problem Set 4-3.csv "

- Variables

- OrderNo – Order number
- ArticleNo – Article No.
- OrderQuantity – ordered quantity
- ShippingQuantity – delivered quantity
- Categorie – Switchgear, Automation, Propulsion or other
- Weigh\_g – Weight in grams
- Length\_mm – Length in mm
- Width\_mm – Width in mm
- Height\_mm – Height in mm
- Containersize – Quantity per packaging unit

## Problem 4-3

- a) Determine the correlations of the variables Order quantity, Weight\_g, Length\_mm, Width\_mm and Container quantity to the variable of the missing parts. Create a point cloud of the strongest correlating variable.
- b) Create a linear regression model for the Missing Parts variable using the variable you identified in 4-3a. Compare the linear model with a model that contains transformed variables according to the influences you estimate from the point cloud (create a linear version of the variable). Continue with the more suitable model.
- c) After interviewing the production staff, you learn that some components seem to be missing more often than others. Compare the average shortfall (missing parts) per product category. Based on the best model of 4-3b, you create one model with dummy variables for each product category and for the most conspicuous category. Continue with the more suitable model.
- d) The procurement department will inform you that the orders are picked and dispatched 2 days before delivery. They are also casually informed that the supplier is "always in chaos at the beginning of the week". Compare the average shortfall per shipping date. (NOTE: look for a weekday function in excel to apply to the data). Based on the model of 4-3c, create a model that takes into account the conspicuous days of the week.
- e) Create a forecast for the following orders and make a recommendation of how to change the orders

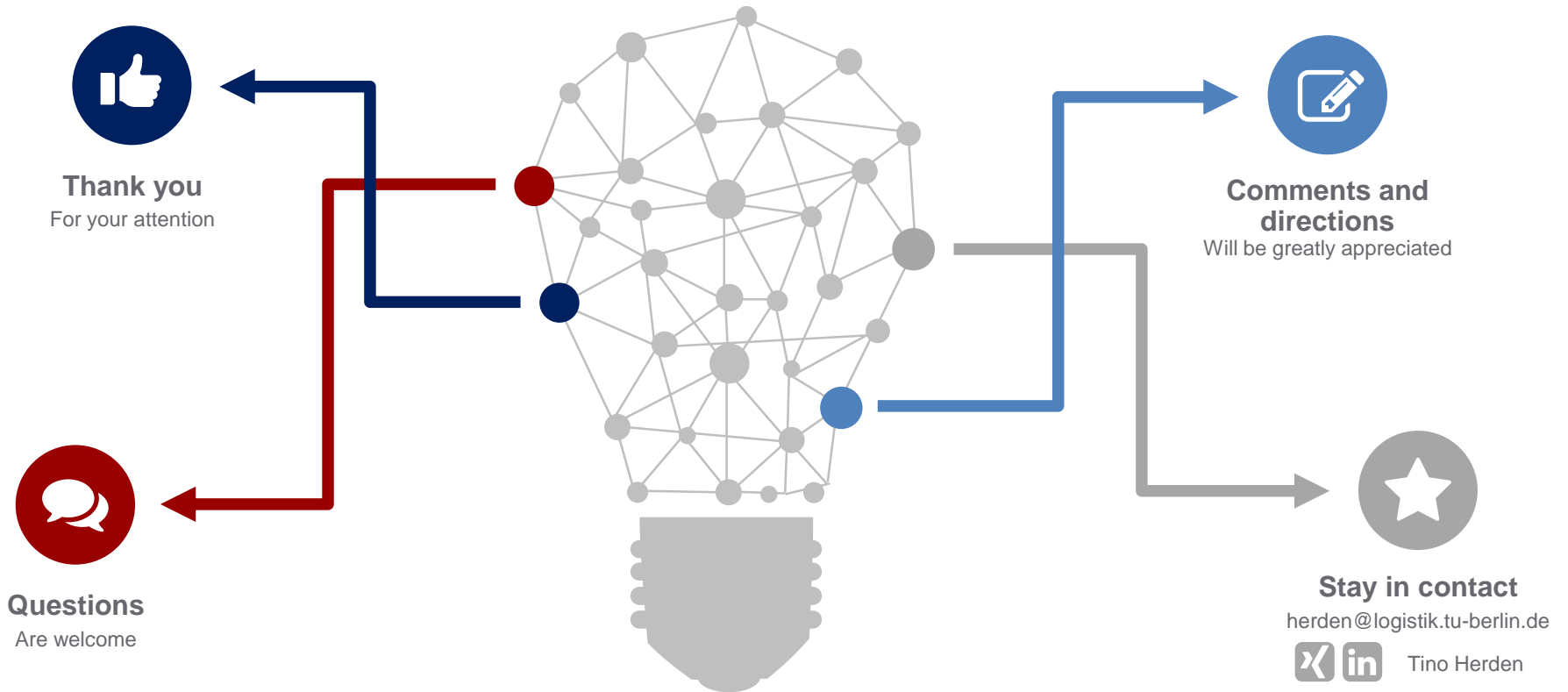
OrderNo	ArticleNo	OrderQuantity	OrderDate	Exp. ShippingDate	Category	Weight(g)	Length(mm)	Width(mm)	Height(mm)	Containersize
327109	STK45265	22	03.06.2017	08.06.2017	Switchgear	493	78	45	102	2
327110	ATK90100	1	04.06.2017	09.06.2017	Propulsion	4543	871	495	653	1



## Summary: Linear Regression

---

- Regression is a method for estimating values in continuous or categorical form.
- Linear regression is used to estimate the continuous values of the dependent variable for which a linear relationship to the respective independent variables is assumed.
- The linear relationship of the variables is evaluated by the coefficient of determination  $R^2$  and the probability that the predictive value of an included variable is zero is evaluated by the significance level



# References

- Baesens, B. (2014): Analytics in a Big Data World.
- Peck, R.; Olsen, C.; Devore, J.L. (2015): Introduction to Statistics and Data Analysis. 5<sup>th</sup> Edition.
- Provost, F., Fawcett, T. (2013): Data Science for Business.
- Ragsdale, C. (2012): Spreadsheet Modeling & Decision Analysis. 6<sup>th</sup> Edition.
- Schutt, R., O'Neil, C. (2013): Doing Data Science – Straight Talk from the Frontline.
- Alpaydin, E. (2010): Introduction to Machine Learning. 2.Auflage. London: The MIT Press.
- Bishop, C. M. (2006): Pattern Recognition and Machine Learning. New York: Springer Science + Business Media.
- Domingos, P. (2012): A Few Useful Things to Know about Machine Learning. Online verfügbar: <https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf>.
- Heaton, J. (2017): An Empirical Analysis of Feature Engineering for Predictive Modeling.
- Kern, R. (2017): Feature Engineering. TU Graz. Online verfügbar: <http://kti.tugraz.at/staff/denis/courses/kddm1/featureengineering.pdf>.
- Kuhn, M., Johnson K. (2013): Applied Predictive Modeling. New York: Springer Science + Business Media.
- Machine Learning Mastery (2018): Discover Feature Engineering, How to Engineer Features and How to Get Good at It. Online verfügbar: <https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/>.
- Wierse, A., Riedel, T. (2017): Smart Data Analytics. Berlin, Boston: De Gruyter Oldenbourg.
- Zabokrtsky, Z. (o.J.): Feature Engineering in Machine Learning. Institute of Formal and Applied Linguistics, Charles University in Prague. Online verfügbar: [https://ufal.mff.cuni.cz/~zabokrtsky/courses/npfl104/html/feature\\_engineering.pdf](https://ufal.mff.cuni.cz/~zabokrtsky/courses/npfl104/html/feature_engineering.pdf).