Regressionsanalysen für Einsteiger

- 20. November 2018, DOAG2018, Nürnberg

ORACLE

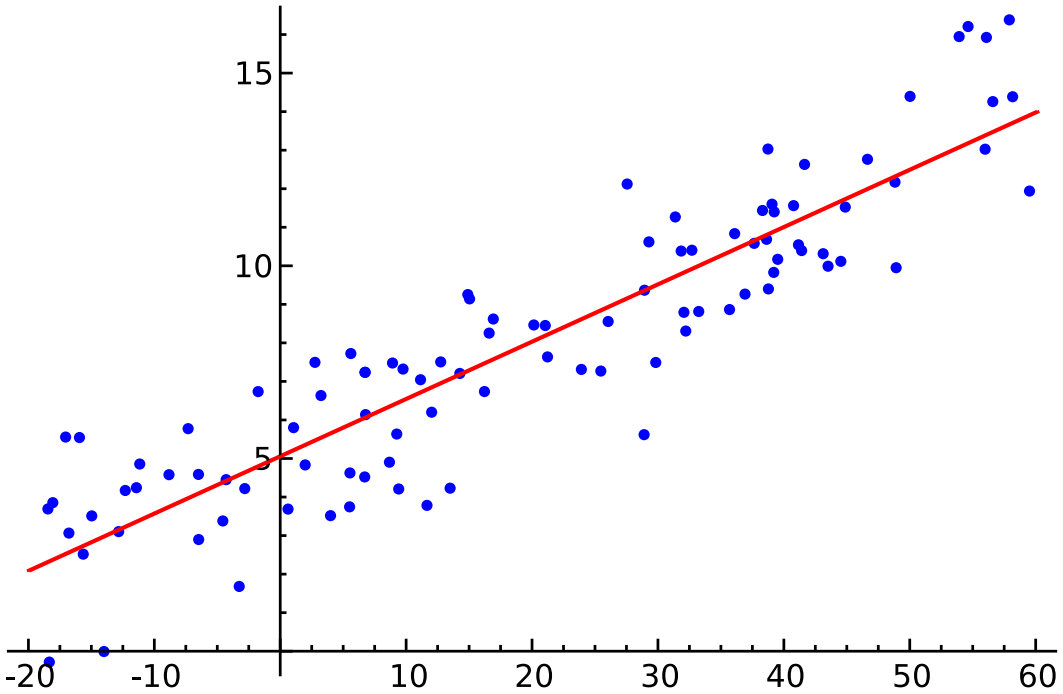
Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

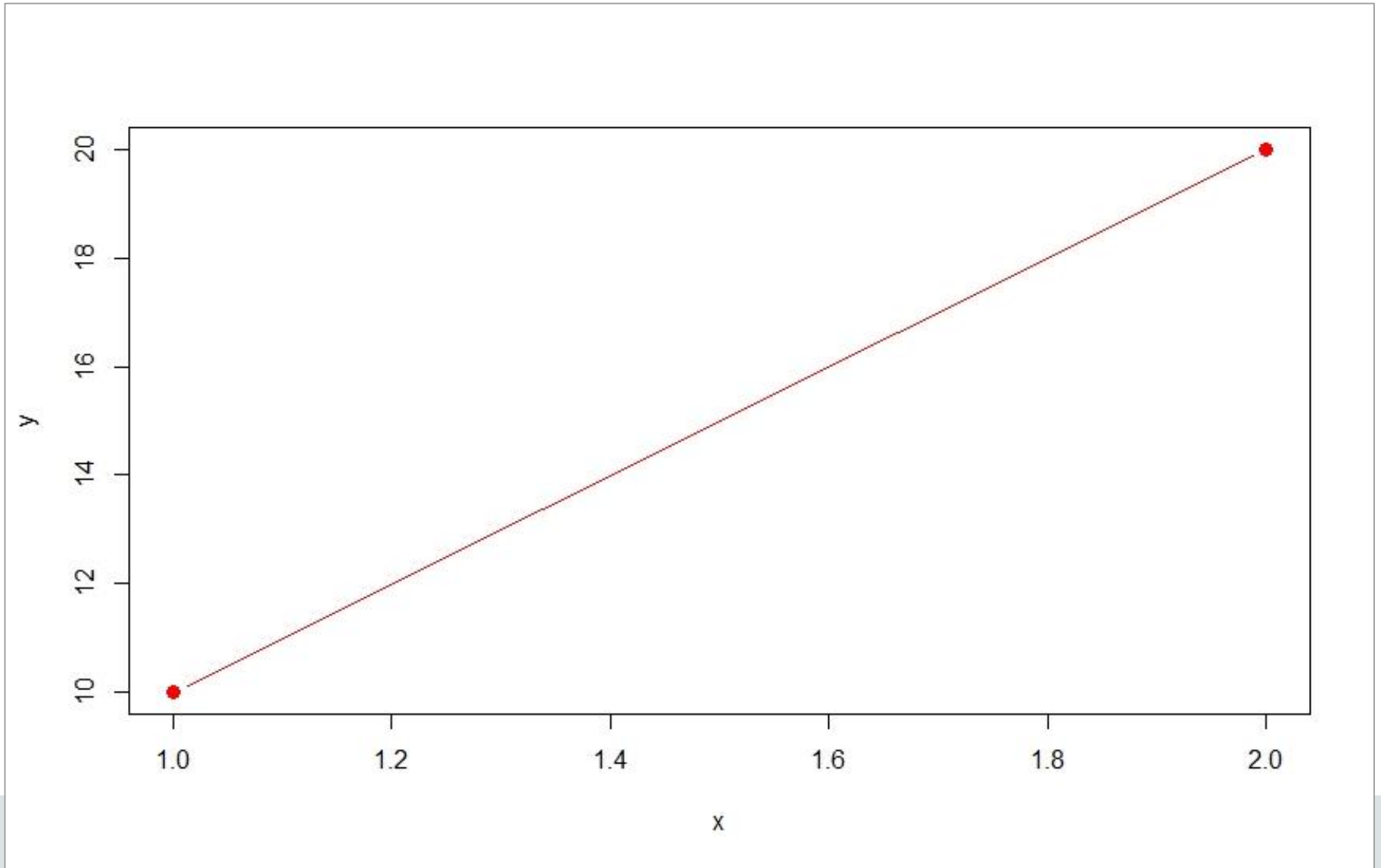
Bekanntes Problem / Fragestellung

Lege eine Gerade möglichst geschickt durch eine Punktewolke!

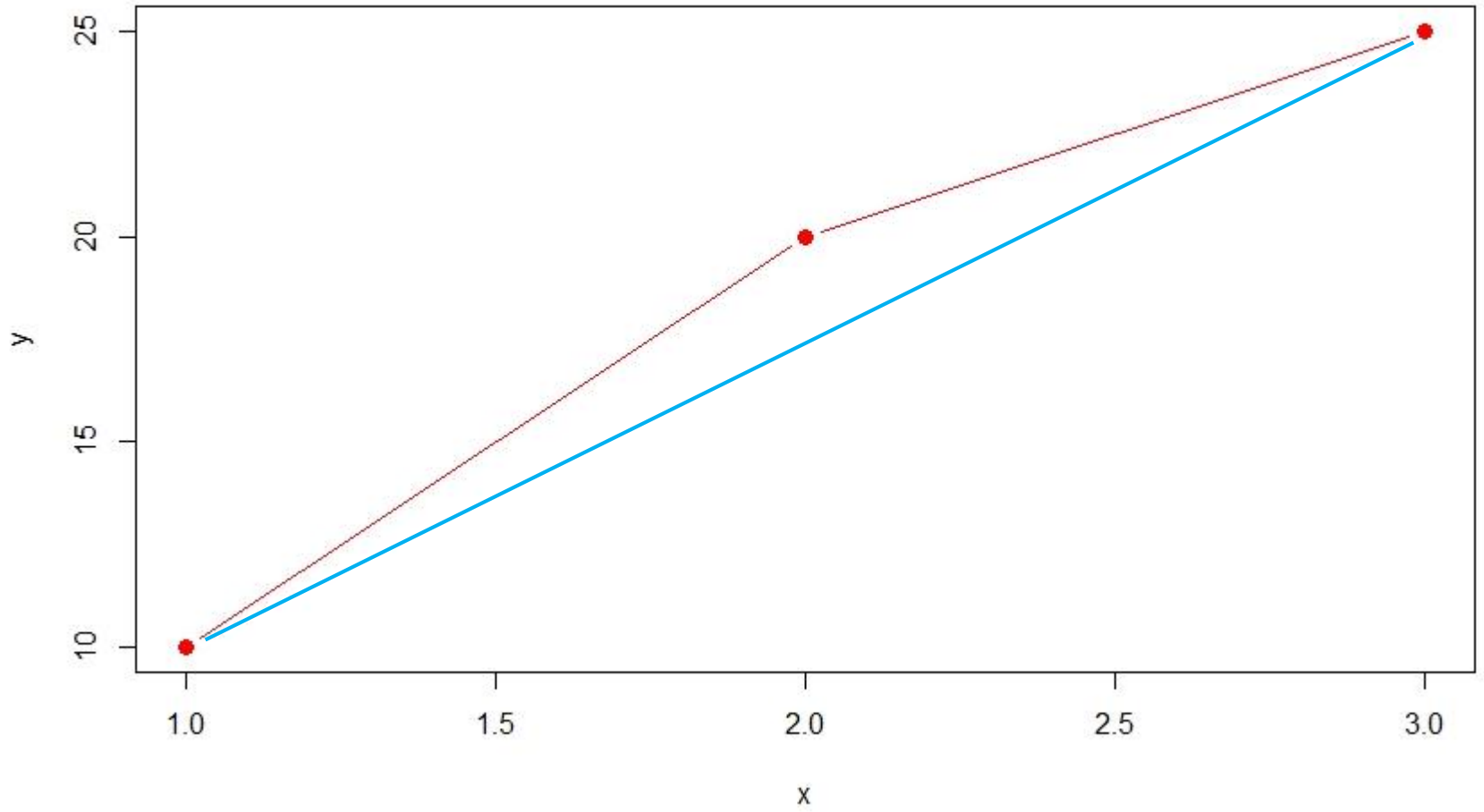


Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

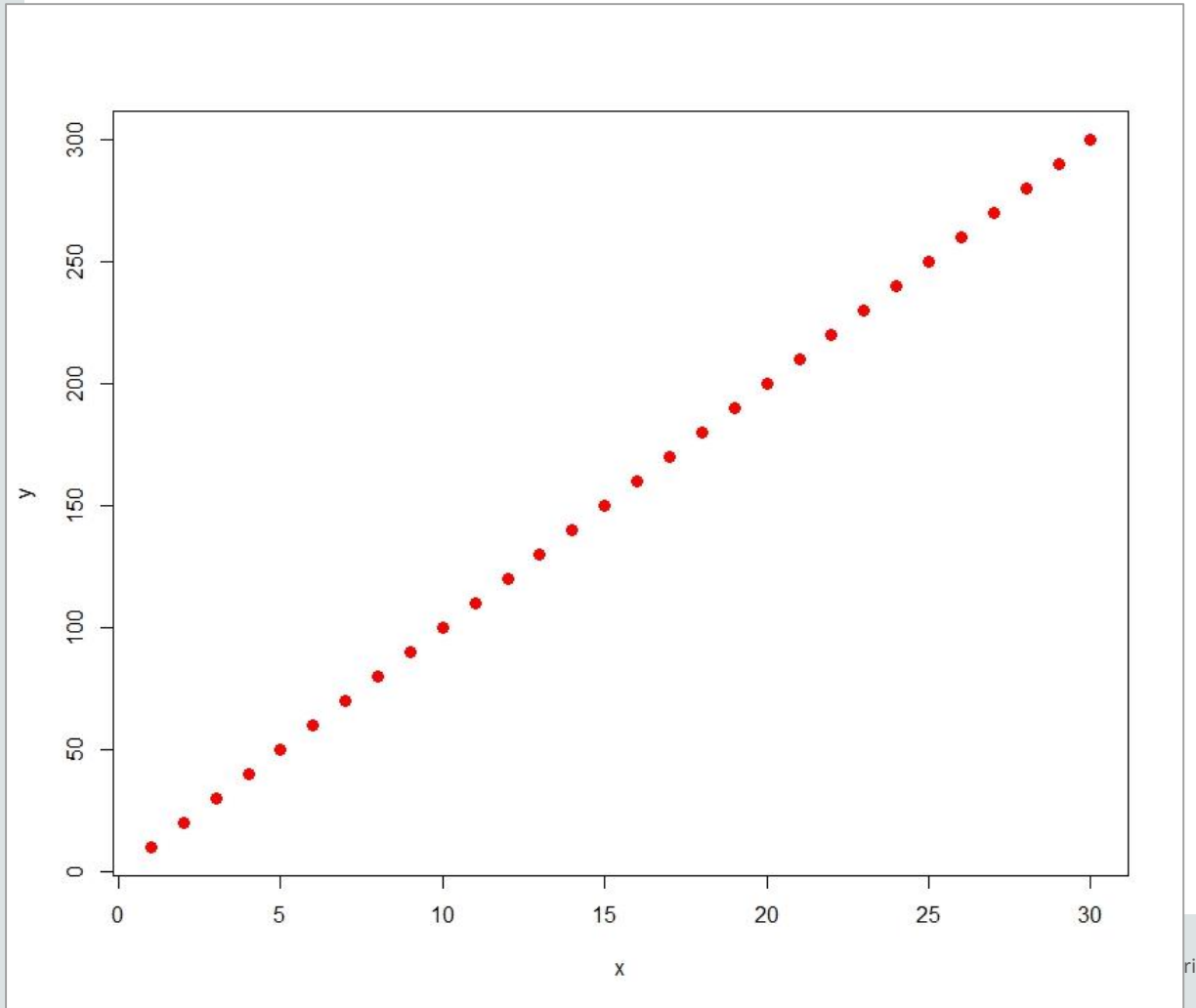
2 Variablen, 2 Punkte, eine Grade

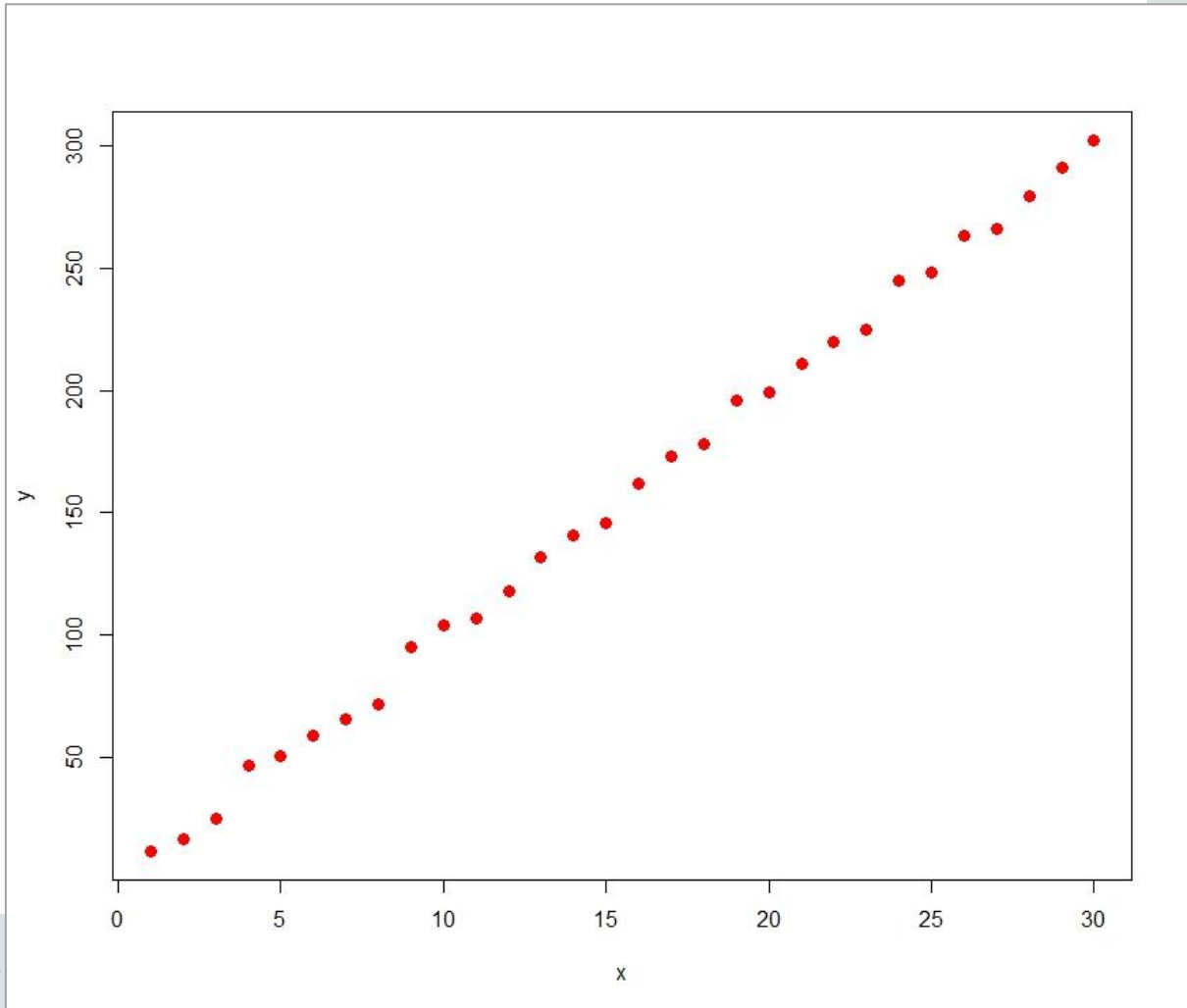


2 Variablen, 3 Punkte, eine Grade?



2 Variablen, 30 Punkte, eine Grade?

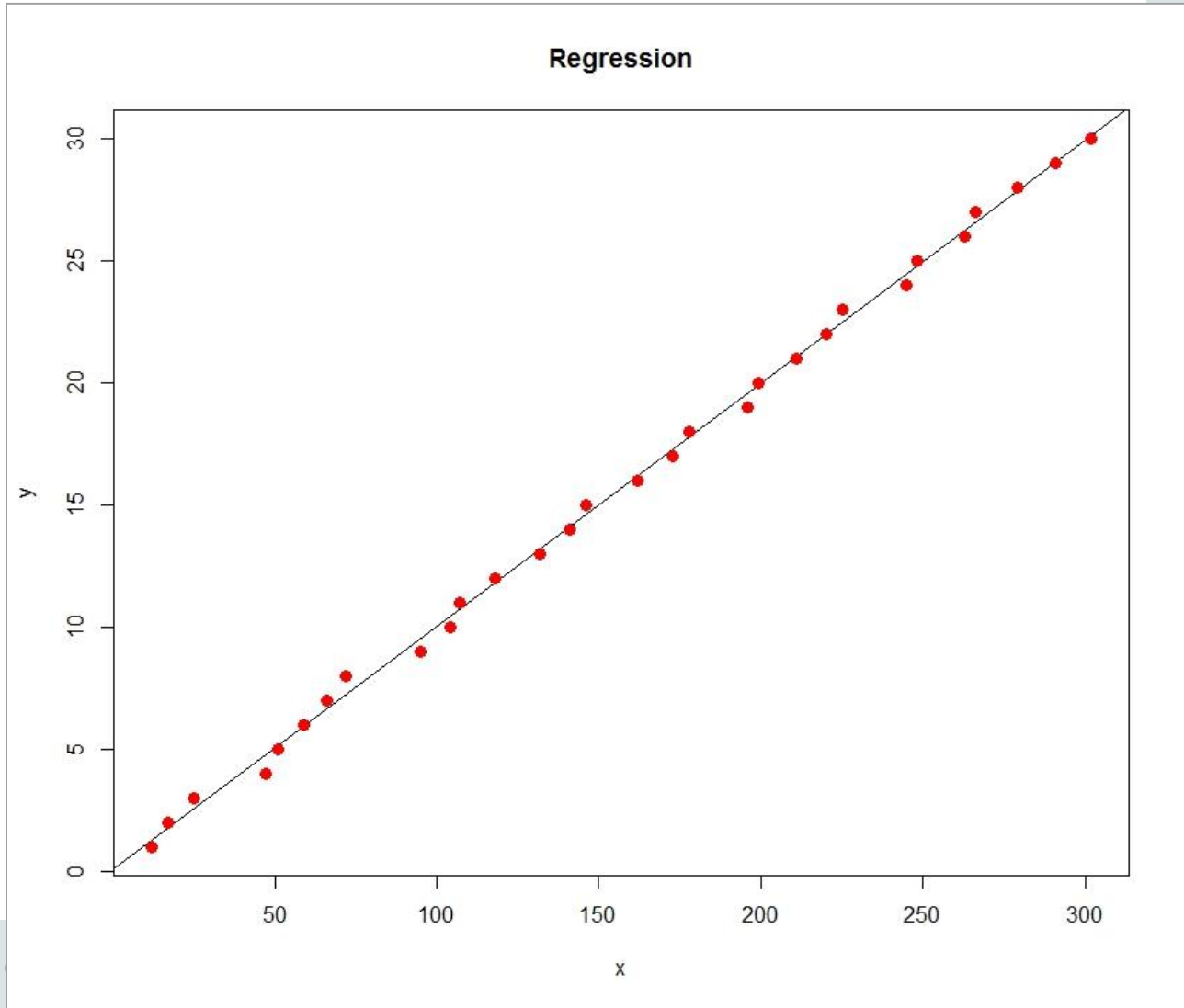




right © 2018,

Lösungsansatz und Fragestellungen

Minimiere den Abstand zwischen den Messwerten und der Gerade!



Copyright

Ordentlich aufgeschrieben….

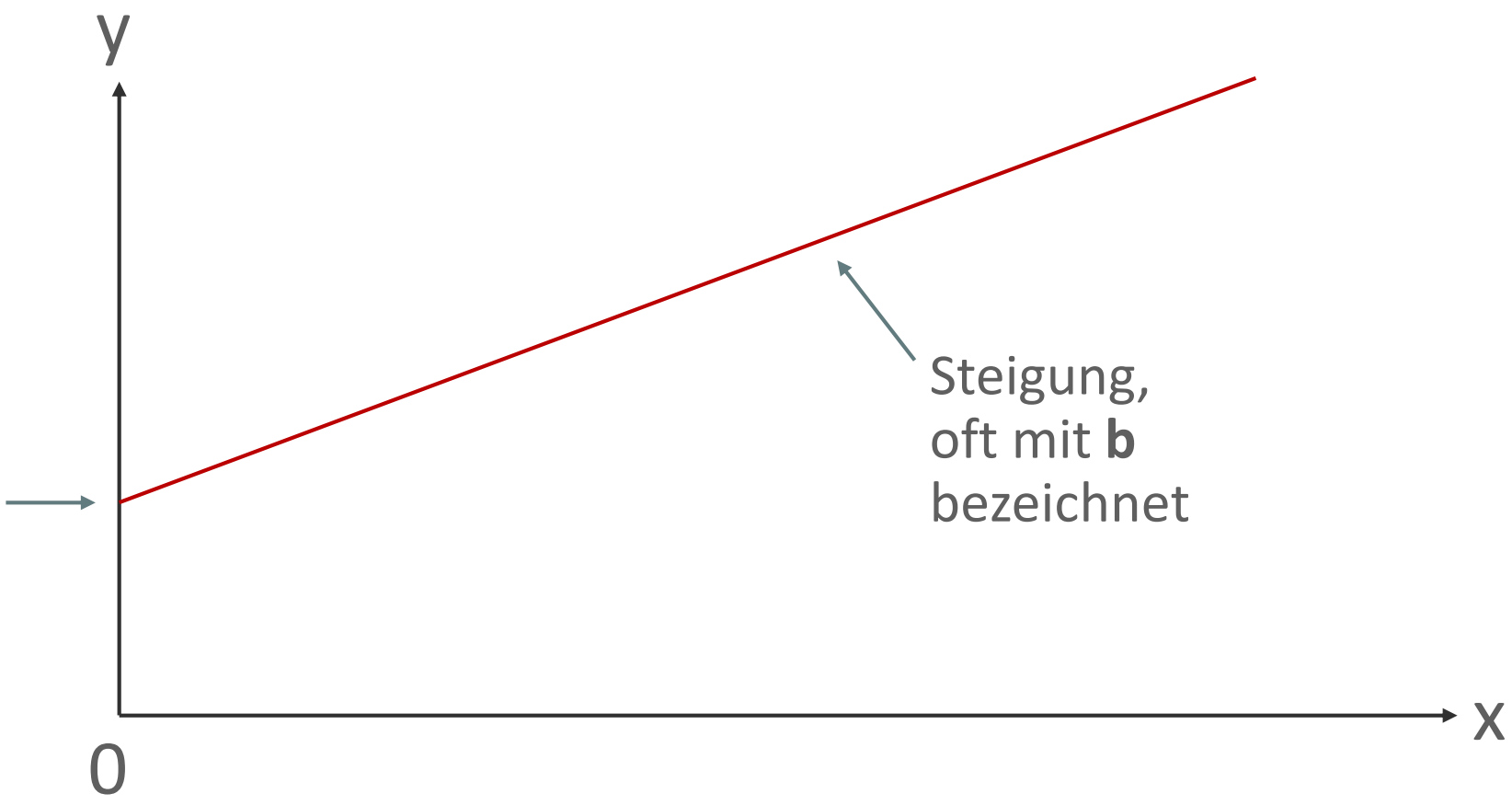
x: unabhängige Variable(n)

y: abhängige Variable(n)

Annahme: eine Geradengleichung ist ein gutes Modell für diese Abhängigkeit

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Geradengleichungen in 2-D



Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Ordentlich aufgeschrieben….

x: unabhängige Variable(n)

y: abhängige Variable(n)

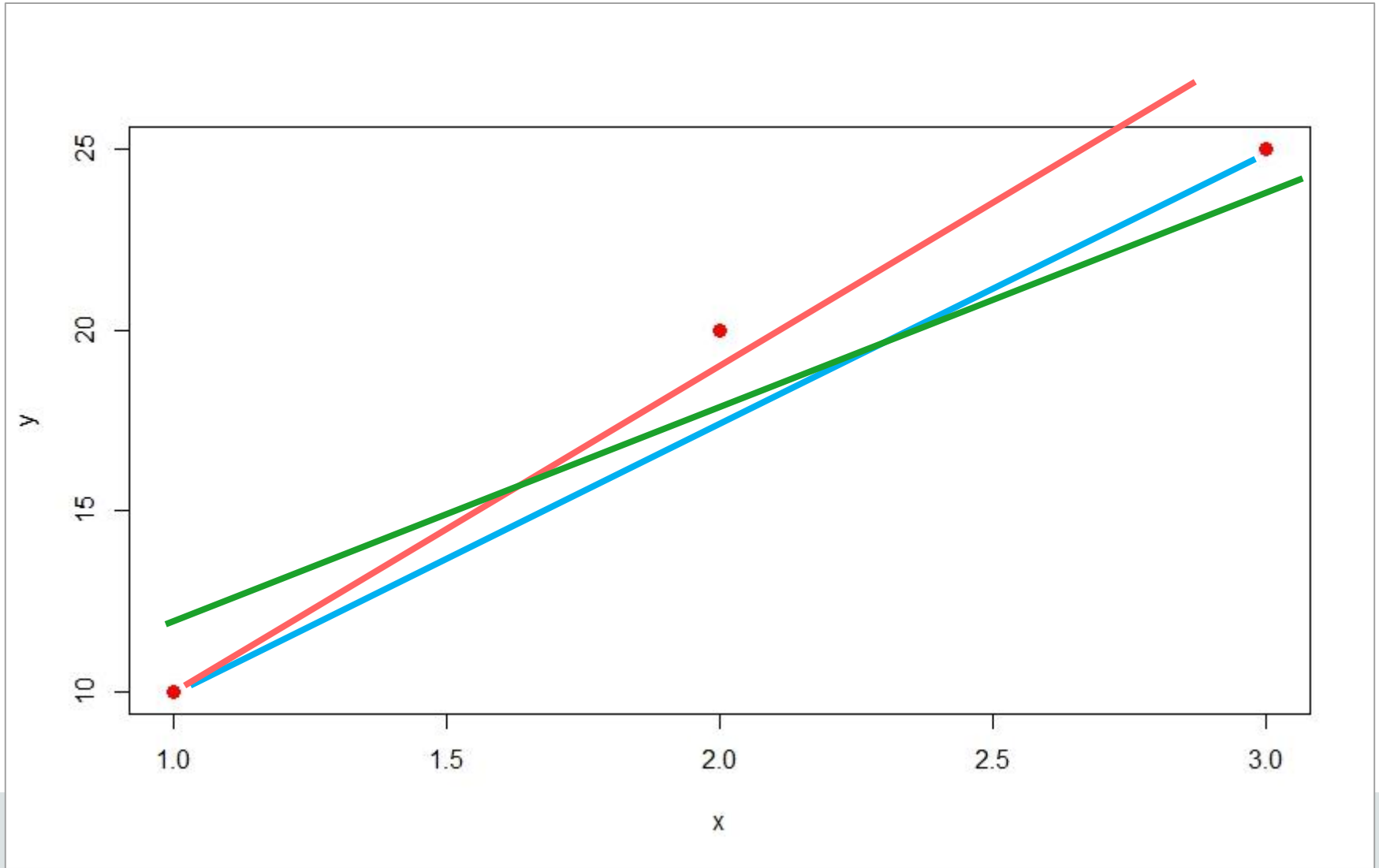
Annahme: eine Geradengleichung ist ein gutes Modell für diese Abhängigkeit

$$y = a + bx + \text{Fehler}$$

Wähle **a** und **b** so, dass der Fehler minimiert wird!

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Fehler = Summer der quadrierten Abstände



Beispiel mit R (1)

```
> zusammenhang <- lm(y~x)
> print(zusammenhang)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    -0.4966      10.0320
```

a **b**

x: unabhängige Variable

y: abhängige Variable

y = a + bx + Fehler

Modell:

y = -0,4966 + 10,032x

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Beispiel mit R (2)

x: unabhängige Variable
y: abhängige Variable

y = a + bx + Fehler

```
> print(summary(zusammenhang))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7597 -2.5017  0.1796  2.3684  7.3684

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.49655    1.38261  -0.359    0.722
x           10.03204    0.07788 128.813   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.692 on 28 degrees of freedom
Multiple R-squared:  0.9983,    Adjusted R-squared:  0.9983
F-statistic: 1.659e+04 on 1 and 28 DF,  p-value: < 2.2e-16
```

a

b

Beispiel mit R (3)

```
> b <- data.frame(x=4)
>
> voraussage <- predict(zusammenhang,b)
> print(voraussage)
       1
39.63159
```

Im Datensatz:

x	y
4	47

Modell:

y = -0,4966 + 10,032x

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

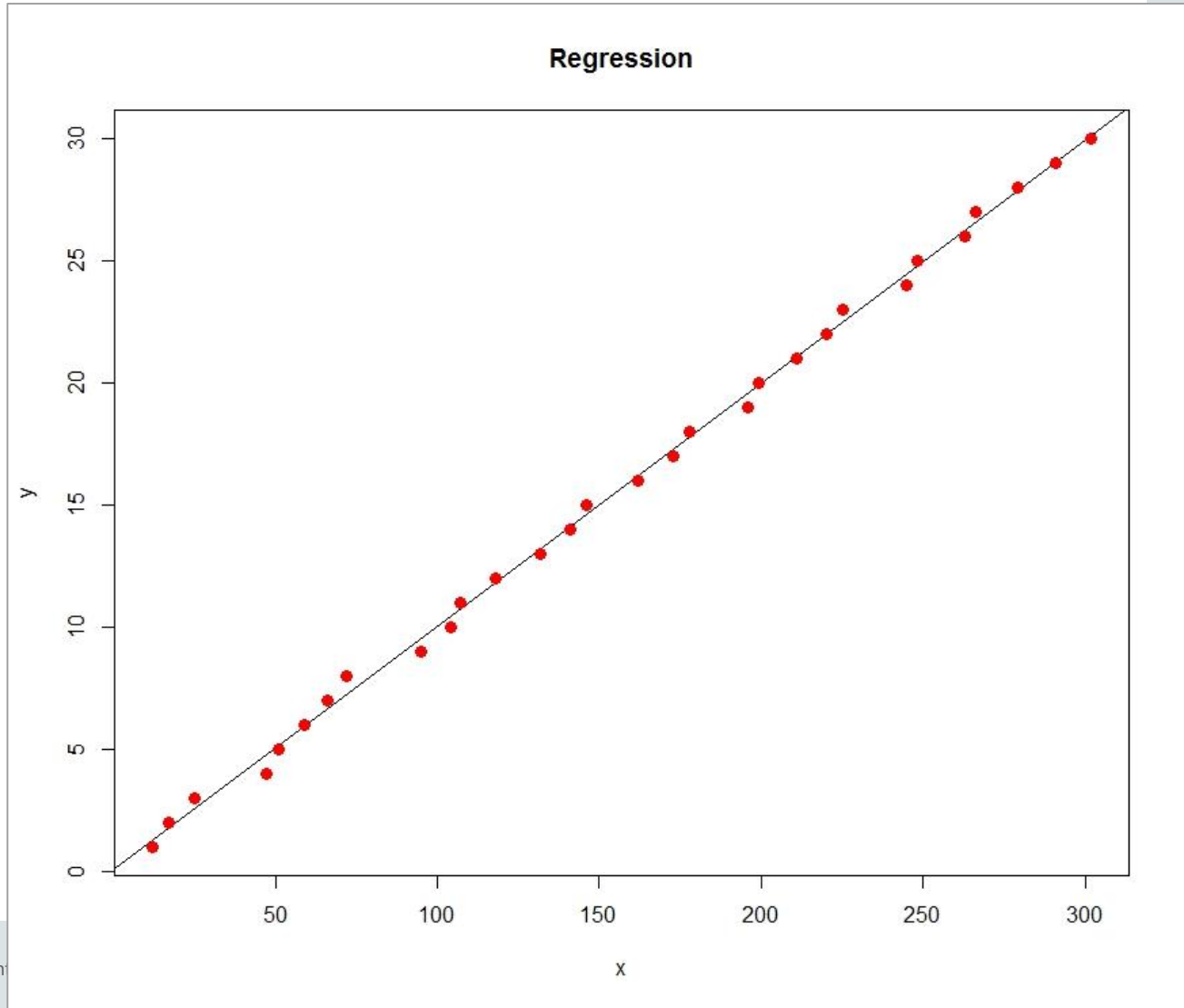
Beispiel mit R (4)

```
> plot(y,x,col = "red",main = "Regression",
+      abline(lm(x~y)),cex = 1.3,pch = 16,xlab = "x",ylab = "y")
> .
```

```
Residual standard error: 3.692 on 28 degrees of freedom
Multiple R-squared:  0.9983,    Adjusted R-squared:  0.9983
F-statistic: 1.659e+04 on 1 and 28 DF,  p-value: < 2.2e-16
```

p-value: unabhängige Variable ist statistisch signifikant (hier: nur x)

R-squared: Quadrat des Pearson-Koeffizienten

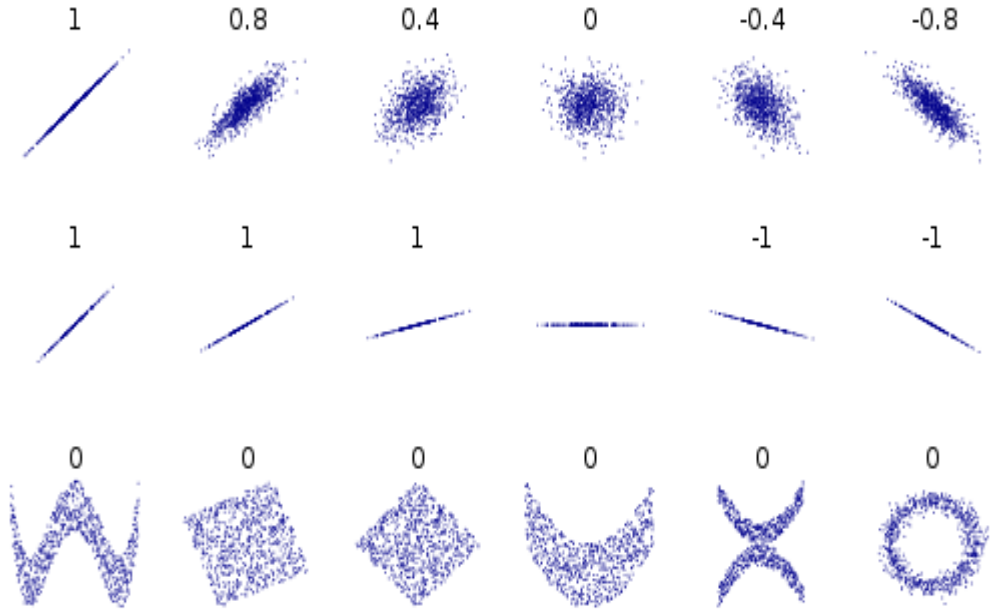


ORACLE
Copyright

Pearson Coefficient

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

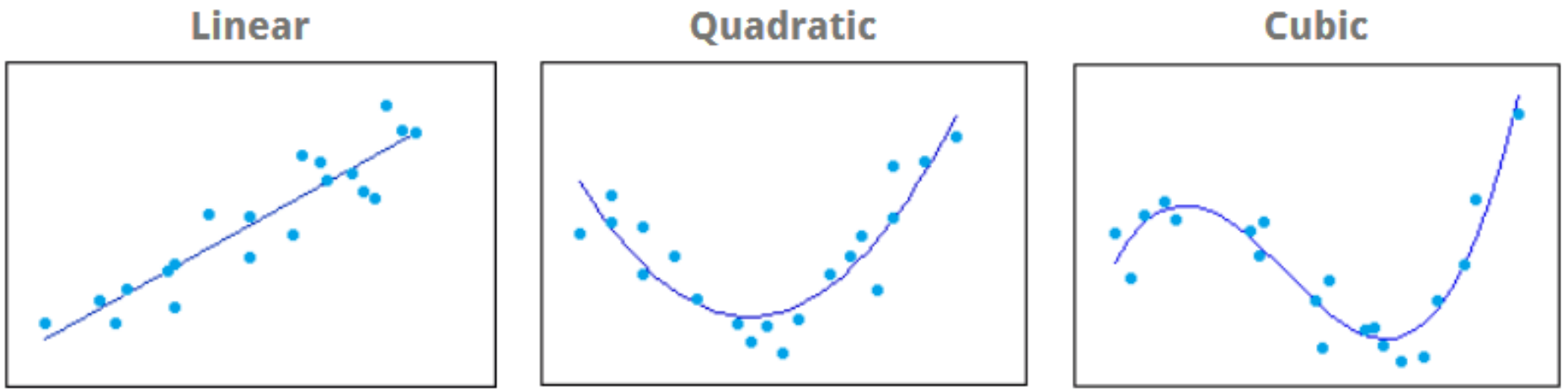
- Liegt zwischen -1 und +1
- Ist ein Maß für Stärke und Richtung einer Linearen Korrelation*
- R-squared bei linearer Regression ist das Quadrat des Pearson Coefficients => beschreibt die durchschnittliche Veränderung der abhängigen Variablen, wenn die unabhängige Variable sich um 1 verändert
- *Korrelation ist keine Kausalität!



Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Lineare Funktion oder was sonst?

http://statisticsbyjim.com/regression/curve-fitting-linear-nonlinear-regression/



⇒ Trotzem Lineare Regression, da die Parameter linear sind!

$$y = a + bx + cx^2 + dx^3 + \ldots + \text{Fehler}$$

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Verallgemeinerungen

x: unabhängige Variable

y: abhängige Variable

nicht nur je eine, sondern mehrere Variablen
(also nicht nur 2D, sondern höherdimensional)

lineare oder nicht-lineare Gleichungen

Annahme: eine ~~Geradengleichung~~ ist ein gutes Modell für diese Abhängigkeit

y = a + bx + Fehler

andere
Fehlerdefinition/Metriken

mehrere Koeffizienten, auch nicht-linear

Wähle **a** und **b** so, dass der Fehler minimiert wird!

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Linear Regression:

Lineare Modelle

General Linear Models:

Multivariate Regressionsmodelle

(ANOVA, ordinary linear regression, t-test, F-test)

Fehler folgt multivariater Normalverteilung

Generalized Linear Models:

Lineare Modelle

Erlauben auch Fehler mit anderen Verteilungen

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Wann machen Regressionsanalysen Sinn?

- Es gibt abhängige und unabhängige Variablen
- Zielstellung: möglichst genaue Voraussagen treffen (unabhängige Variablen sind gegeben, abhängige Variablen sollen vorausgesagt werden)
- Man hat eine Idee vom Modell (Gerade, Polynom…)
- Man hat eine Idee welche Metrik sinnvoll sein könnte
- Metrische Daten
- Möglichst wenig Korrelation zwischen abhängigen Variablen (oder spezielle Regressionsmethoden nutzen)

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Skalenniveaus

Nominalskala: Häufigkeit von Ausprägungen kann gezählt werden

⇒Modalwert (häufigster Wert)

Ordinalskala: Ausprägungen können geordnet werden: <, =, >

=> Quartile (Quantile, 2. Quantil = Median)

⇒min, max

metrische Werte (Kardinalskala): Abstandsbestimmung möglich: |a-b|

=> Spannweite (max-min)

=> Mittelwerte, Streuungsmaße

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Zusammenfassung

- Unabhängigke Variablen (möglichst wenig korreliert)
- Abhängige Variablen
- **Beschreibe die abhängigen Variablen m.H. der unabhängigen Variablen**:
  - Wähle ein Modell aus
    Modelle: Linear (Parameter) oder nicht-linear
    (oder Auswahl über Stepwise oder BestSubset Regression)
  - Fitte die Modellparameter so, dass die Fehler minimiert werden
    Ordinary Least Squares (OLS)= minimiere die Summe der Fehlerquadrate
- Plotte das Ergebnis (Bewertung wie gut das Modell fitted)
- Nutze das Modell für Voraussagen

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Wann man seine Daten nicht Log-transformieren sollte

Nicht-lineare Daten „linear zu machen“ hat auch Nebenwirkungen:

- Paper von O`Hara und Kotze, *Do not log-transform count data*: http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2010.00021.x/abstract
- Blogeintrag bei R-Bloggers: https://www.r-bloggers.com/do-not-log-transform-count-data-bitches/
- Paper zu den Implikationen einer Log-Transformation: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Kontakt

Dr. Nadine Schöne| Senior Systemberaterin

Email: nadine.schoene@oracle.com

Tel: +49 331 200 7190

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

DOAG 2018 - Dienstag

wann	wo	wer	was
08:30	Oslo	Daniela Nicola	Verhagelt das Wetter ihren Geschäftserfolg? Eine KI Analyse
11:00	Oslo	Harald Erb	Machine Learning - eine Challenge für Architekten
13:00	Oslo	M. Braun, A. Etyemez	Machine Learning mit Keras und Tensorflow: Praxisbeispiel
14:00	Oslo	Heli Helskyaho	The Basics of Machine Learning
17:00	Oslo	Björn Ständer	Panel KI/Machine Learning

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

DOAG 2018 - Mittwoch

wann	wo	wer	was
10:00	Kopenhagen	Douglas Hood	Database Driven Machine Learning
10:00	Budapest	Anton Thome	Predictive Analytics: Ein Projektbericht
12:00	Oslo	Peter Czerner	Big Data im Nanobereich: Halbleiter ICs mit Oracle, R & Co.

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

DOAG 2018 - Donnerstag

wann	wo	wer	was
13:00	Seoul	J.-C. Pokolm	Proaktives Problemmanagement mit Cluster Health Advisor
15:00	Helsinki	Oliver Röniger	Self Service Analytics

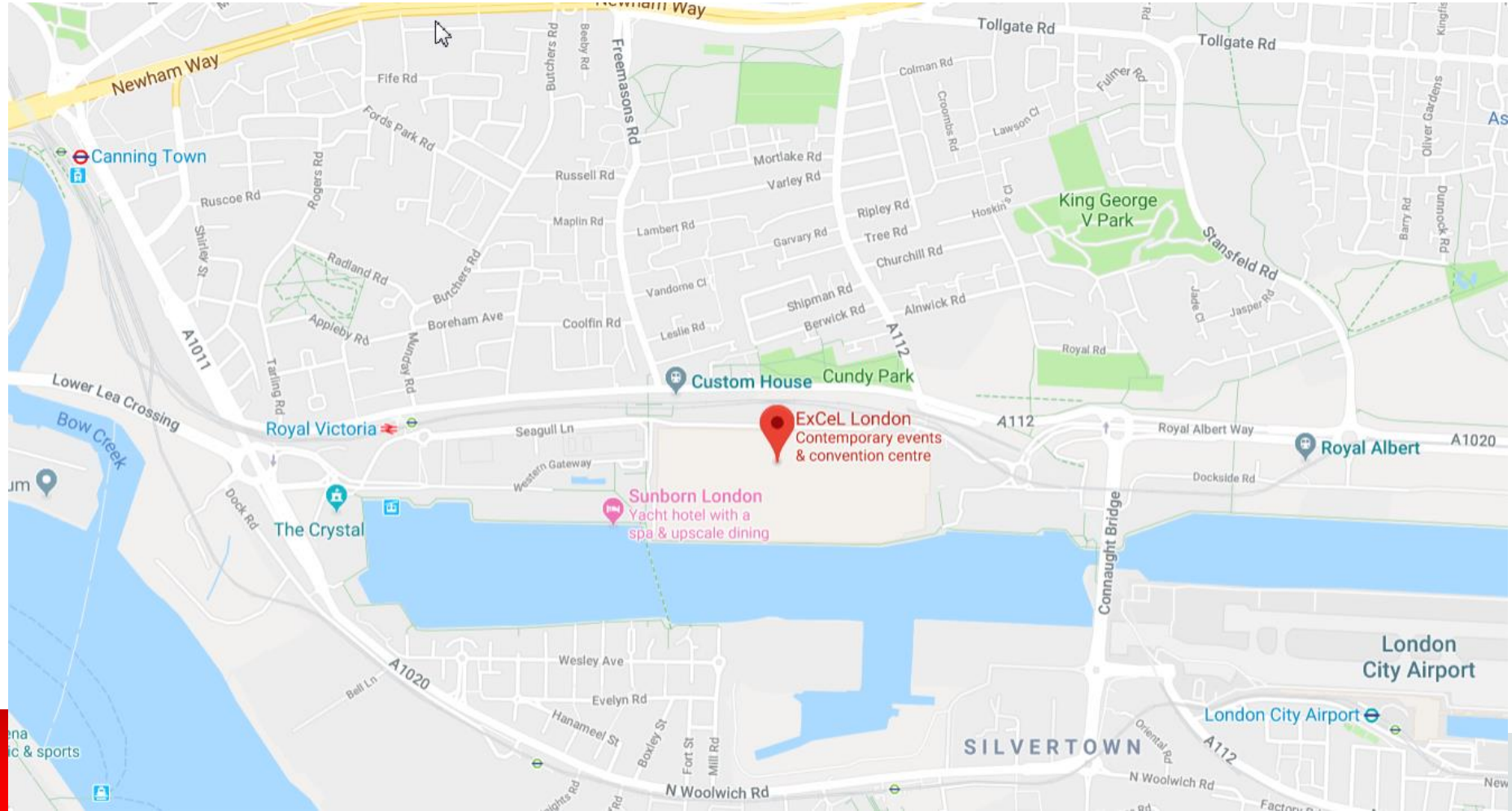
Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Oracle OpenWorld Europe 2019

https://www.oracle.com/uk/openworld/

ORACLE OPENWORLD EUROPE

16–17 January, 2019
LONDON



Integrated Cloud

Applications & Platform Services

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |


ORACLE®