Lineare Regression

Peter Büchel

HSLU I

Stat: Block 11

Lineare Regression

- Verallgemeinerung der Varianzanalyse
- Fortsetzung von Block 3: Jetzt mit Hypothesentest
- Lineare Regression ist der (oder einer der Startpunkte) in Machine Learning

Einführung, Beispiel

- Auftrag als Statistiker einer Firma: Analyse, Strategie auszuarbeiten, wie Verkauf eines bestimmten Produktes gesteigert werden kann
- Firma stellt Daten von Werbebudget und Verkauf zur Verfügung
- Datensatz `Werbung` besteht aus:
  - Dem `Verkauf` dieses Produktes in 200 verschiedenen Märkten und den Werbebudgets für dieses Produkt in diesen Märkten
  - Werbebudget für die drei verschiedenen Medien `TV`, `Radio` und `Zeitung`

- Code:

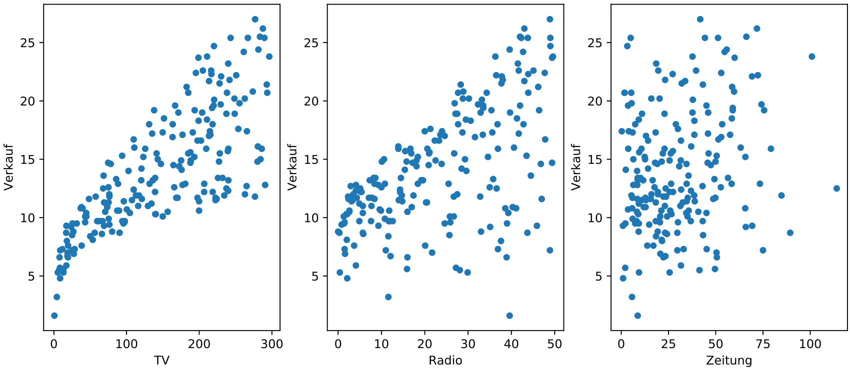
```
import pandas as pd

werbung = pd.read_csv("../Data/Werbung.csv").drop(["Unnamed:
0"], axis=1)

werbung.head()
```

```
##        TV  Radio  Zeitung  Verkauf
## 0  230.1   37.8     69.2     22.1
## 1   44.5   39.3     45.1     10.4
## 2   17.2   45.9     69.3      9.3
## 3  151.5   41.3     58.5     18.5
## 4  180.8   10.8     58.4     12.9
```

- Daten in Streudiagrammen dargestellt:



- Code:

```
import matplotlib.pyplot as plt

plt.subplot(131)
werbung.plot(kind="scatter",x="TV", y="Verkauf",
ax=plt.gca())


plt.subplot(132)
werbung.plot(kind="scatter",x="Radio", y="Verkauf",
ax=plt.gca())

plt.subplot(133)
werbung.plot(kind="scatter",x="Zeitung", y="Verkauf",
ax=plt.gca())

plt.show()
```

- Für Firma nicht möglich, Verkauf des Produktes direkt zu erhöhen
- Aber sie kann Werbeausgaben in den drei Medien kontrollieren
- Ziel: Zusammenhang zwischen Werbung und Verkauf herstellen, damit Firma ihre Werbebudgets anpassen kann, damit sie den Verkauf indirekt erhöhen kann
- Ziel: Möglichst genaues *Modell* zu entwickeln, damit auf Basis der drei Medienbudgets der Verkauf des Produkts *vorhersagt* werden kann
- Abbildung oben links: Deutlicher Zusammenhang zwischen dem Werbebudget und dem Verkauf des Produktes

- Je mehr in Werbung investiert wird, desto grösser Verkaufszahlen
- Frage: Welche *Form* dieser Zusammenhang?
- Möglichkeit: Datenpunkte folgen einer Gerade siehe später
- Abbildung oben rechts: überhaupt keinen Zusammenhang
- Folglich kann man die Zeitungswerbung hier sein lassen

- Mathematische Sichtweise: Gesucht Funktion f, die Werbebudgets X_1 (`TV`), X_2 (`Radio`) und X_3 (`Zeitung`) den Verkauf Y ermittelt:

$$Y \approx f(X_1, X_2, X_3)$$

- Beziehung oben: Kein Gleichheitszeichen, da Streudiagramme keine Graphen einer Funktion darstellen

- Funktion f kann Zusammenhang zwischen X_1, X_2, X_3 und Y nur *approximativ* darstellen

- Bezeichnung:
  - Variable Y: *Zielgrösse*, *Outputvariable*,
  - X_1, X_2 und X_3: *Prädiktoren,erklärende Variable*

- Allgemein: Quantitative Zielgrösse Y und p verschiedene Prädiktoren $X_1, X_2, \ldots, X_p$

- Annahme: Es besteht irgendein Zusammenhang zwischen Y und $X_1, X_2, \ldots, X_p$

- Allgemeine Form:

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

- f irgendeine feste, aber *unbekannte* Funktion von $X_1, X_2, \ldots, X_p$

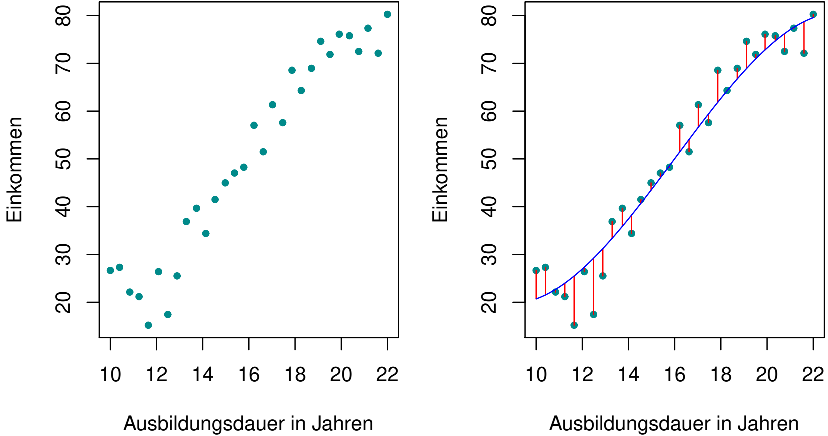
- Grösse ε: *Zufälliger Fehlerterm* unabhängig von $X_1, X_2, \ldots, X_p$ mit Mittelwert 0

- Bedeutung Fehlerterm $\varepsilon \quad \rightarrow \quad$ Folgendes Beispiel

Beispiel: Einkommen

- Abbildung links: `Einkommen` von 30 Individuen in Abhängigkeit der `Ausbildungsdauer` (in Jahren)

- Graphik deutet an: `Einkommen` kann aus `Ausbildungsdauer` berechnet werden



- Aber: Funktion f, die die Prädiktoren und die Zielgrösse miteinander in Verbindung bringt, in der Regel unbekannt
- In dieser Situation: f aus den Daten *schätzen*
- Datensatz simuliert: Funktion f bekannt (blaue Kurve) in Abb. rechts
- Einige Beobachtungen liegen überhalb, andere unterhalb der blauen Kurve
- Die roten vertikalen Linien repräsentieren den Fehlerterm ε
- Insgesamt haben Fehler einen empirischen Mittelwert annähernd 0
- Ziel der Regression: Funktion f zu *schätzen*

- Schätzen in der Stochastik: Berechnung
- Schätzung ist Annäherung (Approximation) an wahre Grösse
- Geschätzte Grösse wird mit Hut $\widehat{}$ gekennzeichnet
- $\widehat{Y}$: Schätzung der unbekannten Grösse Y
- $\widehat{f}$: Schätzung der unbekannten Funktion f

Warum soll f geschätzt werden?

- Hauptgründe, warum man unbekannte Funktion f schätzen will:
  - Datenpunkte *vorherzusagen* (*Prognose*)
  - *Rückschlüsse* auf Funktion selbst zu ziehen
- Prognose: Oft Prädiktoren $X_1, X_2, \ldots, X_p$ einfach verfügbar, aber die Zielgrösse nicht
- In so einem Fall: Y schätzen durch

$$\widehat{Y} = \widehat{f}(X_1, X_2, \ldots, X_p)$$

- Fehlerterm im Mittel 0

Beispiel

- Prädiktoren $X_1, X_2, \ldots, X_p$ seien die Werte von verschiedenen Charakteristiken einer Blutentnahme, die der Hausarzt des Patienten in seinem Labor bestimmen kann

- Zielgrösse Y: Mass für Risiko, dass der Patient starke Nebenwirkungen bei der Anwendung eines bestimmten Medikamentes erleidet

- Arzt möchte bei Verschreibung eines Medikamentes Y aufgrund von $X_1, X_2, \ldots, X_p$ vorhersagen können, damit er nicht ein Medikament Patienten verschreibt, die ein hohes Risiko für Nebenwirkungen bei diesem Medikament haben - d.h. bei denen Y gross ist

- Genauigkeit von $\widehat{Y}$ als Vorhersage von Y hängt von zwei Grössen ab:
  - *Reduzibler Fehler*
  - *Irreduzibler Fehler*
- Allgemein: $\widehat{f}$ keine perfekte Schätzung von f und diese Ungenauigkeit führt zu einem Fehler
- *Reduzibler Fehler*: Schätzung mit statistischen Methoden verbessern
- Aber auch für perfekte Schätzung von f: Outputvariable hat Form

$$\widehat{Y} = f(X_1, X_2, \ldots, X_p)$$

- Vorhersage Y enthält immer noch Fehler
- Liegt am Fehlerterm ε: Hängt nicht von $X_1, X_2, \ldots, X_p$ ab
- Variablilität von ε beeinflusst die Genauigkeit der Vorhersage

- *Irreduzibler Fehler*: Fehler kann nicht beeinflusst werden, wie gut auch die Schätzung von f ist
- Woher kommt nun dieser Fehler ε, der grösser als Null ist?
- Grösse kann Variablen enthalten, die nicht gemessen wurden, die aber für die Vorhersage von Y wichtig sind
- Da diese Variablen nicht gemessen wurden $\quad\rightarrow\quad$ Für die Vorhersage auch nicht verwendbar
- Grösse ε kann aber auch nicht messbare Grössen enthalten
- Bsp: Stärke der Nebenwirkungen eines Medikamentes abhängig sein von der Tageszeit der Einnahme des Medikamentes oder auch einfach vom allgemeinen Wohlbefinden des Patienten

Rückschlüsse auf f: Fragestellungen

- *Welche Inputvariablen werden mit dem Output assoziert?*
  - Natürlich alle, denkt man zuerst
  - Aber oft sind es einige wenige Variablen, die auf Y einen substantiellen Einfluss haben
  - Sehr viele Inputvariablen $\rightarrow$ *wichtige* Inputvariablen identifizieren
  - Beispiel `Werbung`:
    - Ausgaben bei TV-Werbung grosser Einfluss auf die Verkaufszahlen
    - Zeitungswerbung aber nicht
    - Auf die TV-Werbung konzentrieren

- *Wie sieht der Zusammenhang zwischen Outputvariable und jeder Inputvariable aus?*
  - Einige Inputvariablen haben einen positiven Zusammenhang mit der Outputvariable
  - Eine Vergrösserung der Inputvariable hat in diesem Fall eine Vergrösserung von Y zur Folge
  - Andere Inputvariablen haben einen negativen Zusammenhang mit Y
  - In Abhängigkeit von der Komplexität von f kann der Zusammenhang zwischen der Zielvariablen und einer erklärenden auch von den Werten der anderen erklärenden Variablen abhängen (Interaktion)

- *Kann der Zusammenhang zwischen der Outputvariable und jeder Inputvariable durch eine lineare Gleichung angemessen beschrieben werden oder ist der Zusammenhang komplizierter?*
  - Historisch sind die meisten Schätzungen von f linear
  - Dies hat damit zu tun, dass solche Schätzungen sehr einfach sind
  - In vielen Situationen: Annahme Linearität ausreichend oder gar wünschenswert
  - Aber oft ist der wahre Zusammenhang komplizierter und das lineare Modell liefert keinen angemessenen Zusammenhang zwischen Input- und Outputvariablen

Fragen für Beispiel der `Werbung`

- Welche Medien tragen zum Verkauf des Produktes bei?
- Welche Medien haben den grössten Einfluss auf den Verkauf?
- Welchen Zuwachs im Verkauf hat eine bestimmte Vergrösserung der TV-Werbung zur Folge?

Schätzung von f?

- Mehrere Verfahren um zu f schätzen
- Hier nur *parametrische Methode*
- Vorgehen:
  - *Annahme* über die funktionale Form von f
  - Einfachste Annahme: f linear in $X_1, X_2, \ldots, X_p$:

$$f(X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

  - Nach Wahl des Modells: Verfahren, das die Daten in das Modell *passt*
  - Lineares Modell: Parameter $\beta_0, \beta_1, \ldots \beta_p$ schätzen
  - Parameter so bestimmen, dass

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

  - Häufigste Methode zur Bestimmung von $\beta_0, \beta_1, \ldots \beta_p$: *Methode der kleinsten Quadrate*

Beispiele

- Beispiel `Werbung`: Lineares Modell:

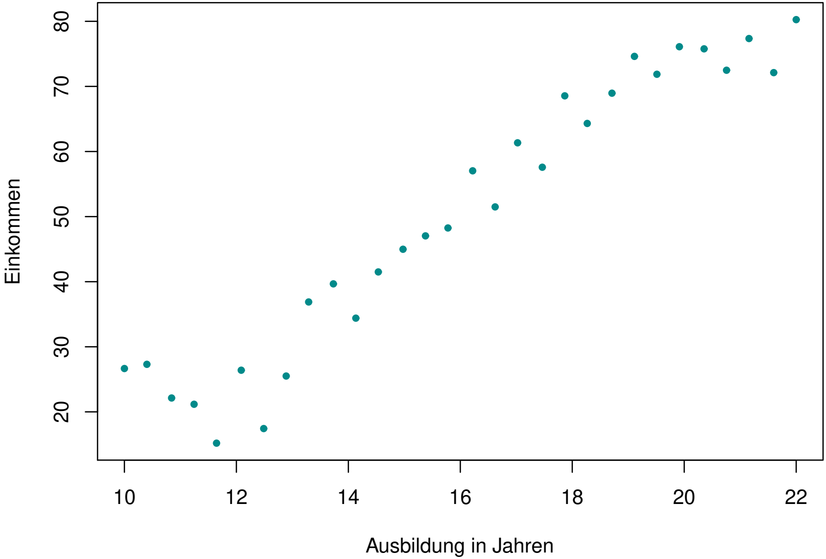
$$\texttt{Verkauf} \approx \beta_0 + \beta_1 \cdot \texttt{TV} + \beta_2 \cdot \texttt{Radio} + \beta_3 \cdot \texttt{Zeitung}$$

- Beispiel `Einkommen`: Lineares Modell:

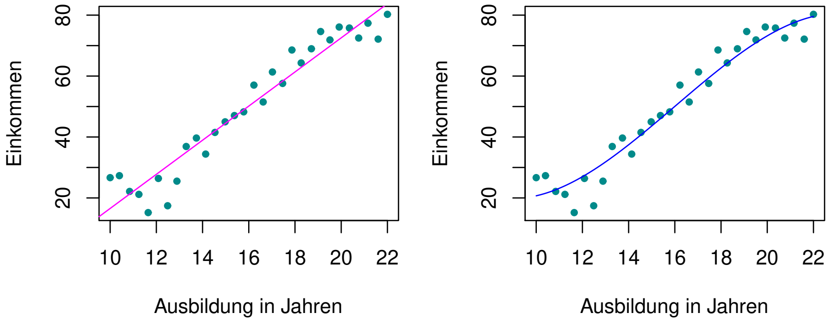
$$\texttt{Einkommen} \approx \beta_0 + \beta_1 \cdot \texttt{Ausbildung}$$

Beispiel

- Datensatz `Einkommen`:



- Frage: Welches *Modell* wählen, oder welche Form soll f haben



- Aus Daten: Lineares Modell (oben links):

$$f(X) = \beta_0 + \beta_1 X$$

- Auch kubisches Modell (Polynom 3. Grades) möglich (oben rechts):

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

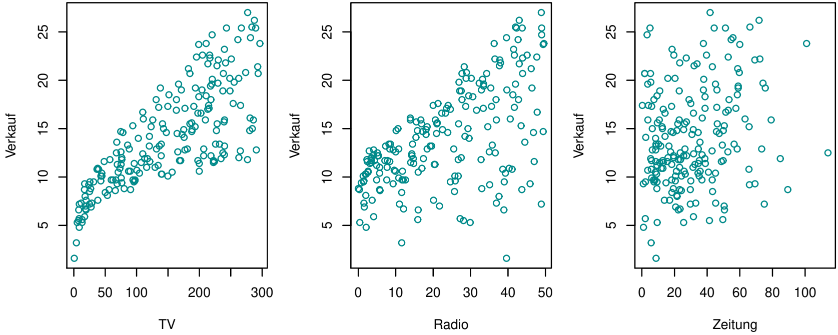
- Viele weitere Modelle denkbar
- Aber welches ist nun das „richtige"?
- Dies lässt sich in dieser Absolutheit nicht entscheiden
- Funktion f i. A. unbekannt: Liegt an uns „bestes" Modell zu wählen
- Statistik: Bei Entscheidungsfindung behilflich
- Welches Modell ist in unserem Beispiel das „bessere"?
- Kubisches Modell scheint besser zu passen, aber auch komplizierter
- Lineares Modell einfacher (etwas weniger genau) hat Vorteil: Die Parameter β_0 und β_1 lassen sich geometrisch interpretieren:
  - β_0 ist der y-Achsenabschnitt
  - β_1 die Steigung der Geraden

Bemerkungen

- Komplizierteres Modell muss *nicht* das bessere Modell sein
- Phänomen: *Overfitting*
- Fehler oder Ausreisser werden zu stark berücksichtigt
- In sehr vielen Fällen: Lineares Modell ausreichend

Lineare Regression

- Datensatz **`Werbung`**:



- **`Verkauf`** für ein bestimmtes Produkt (in Einheiten von tausend verkauften Produkten) als Funktion von Werbebudgets (in Einheiten von tausend CHF) für **`TV`**, **`Radio`** und **`Zeitung`**

- Aufgrund dieser Daten: Statistiker erstellen Marketingplan, der für nächstes Jahr zu höheren Verkäufen führen soll
- Welche Informationen sind nützlich, um solche Empfehlungen auszuarbeiten?

Fragestellungen

- *Gibt es Zusammenhang zwischen Werbebudget und Verkauf?*
  - Erstes Ziel: Entscheiden, ob die Daten genügend Hinweise für einen Zusammenhang zwischen Werbebudget und Verkauf liefern
  - Ist der Hinweis schwach, dann kann man argumentieren, dass auf die Werbung gänzlich verzichtet werden kann

- *Wie stark ist Zusammenhang zwischen Werbebudget und Verkauf?*
  - Annahme: Zusammenhang zwischen Werbebudget und Verkauf vorhanden
  - Möchten wissen, wie *stark* dieser Zusammenhang ist
  - Kann man für ein gegebenes Werbebudget den Verkauf mit hoher Genauigkeit vorhersagen?
  - In diesem Fall gäbe es einen starken Zusammenhang
  - Oder ist berechnete Vorhersage nur wenig besser als zufällige Vorhersage?
  - In diesem Fall würde ein schwacher Zusammenhang vorliegen

- *Welche Medien tragen zum Verkauf bei?*
  - Tragen alle drei Medien (TV, Radio, Zeitung) zum Verkauf bei oder ist es nur eines oder zwei?
  - Weg finden, um Einfluss jedes einzelnen Mediums auf den Verkauf separat zu ermitteln, auch wenn für alle drei Medien Geld ausgegeben wird
- *Wie genau kann man den Einfluss jedes einzelnen Mediums auf den Verkauf schätzen?*
  - Wie gross ist die Zunahme des Verkaufs für jeden zusätzlichen Franken, den wir für ein spezifisches Medium ausgeben?
  - Wie genau können wir diese Zunahme vorhersagen?

- *Wie genau können wir zukünftige Verkäufe vorhersagen?*
  - Welche Verkäufe können wir für beliebige Werbebudgets für TV, Radio und Zeitung vorhersagen und wie genau ist diese Vorhersage?
- *Ist der Zusammenhang linear?*
  - Ist der Zusammenhang zwischen Werbebudgets für die unterschiedlichen Medien und Verkauf annähernd linear, dann ist lineare Regression ein angebrachtes Modell
  - Falls nicht, so kann mit Hilfe von Variablentransformation lineare Regression unter Umständen trotzdem verwendet werden
- *Gibt es Synergie zwischen den verschiedenen Medien?*
  - Möglicherweise bewirkt CHF 50 000 für TV-Werbung und CHF 50 000 für Radiowerbung mehr Verkäufe als wenn wir CHF 100 000 für das eine oder andere Medium aufgewendet hätten
  - Marketing: *Synergieeffekt*; Statistik: *Interaktionseffekt*

- Mit linearer Regression lassen sich alle diese Fragen beantworten

Einfaches Regressionsmodell

- *Einfache lineare Regression*: Sehr einfaches Verfahren, um einen quantitativen Output Y auf der Basis einer einzigen Inputvariable X
- Annahme: Annähernd lineare Beziehung zwischen X und Y
- Mathematisch: Lineare Beziehung:

$$Y \approx \beta_0 + \beta_1 X$$

- Dabei steht „$\approx$" für „ist annähernd modelliert durch"

Beispiel

- Beispiel `Werbung`: X Grösse `TV` und Y Grösse `Verkauf`

- Nach dem linearen Regressionsmodell gilt dann

$$\texttt{Verkauf} \approx \beta_0 + \beta_1 \cdot \texttt{TV}$$

- Grössen β_0 und β_1 sind unbekannte Konstanten, die den y-Achsenabschnitt und die Steigung des linearen Modells darstellen

- β_0 und β_1 die *Koeffizienten* oder *Parameter* des Modells

- Koeffizienten werden aus den gegebenen Daten geschätzt

- Schätzungen $\widehat{\beta}_0$ und $\widehat{\beta}_1$ für die Modellkoeffizienten

- Sind diese Koeffizienten bekannt, so können zukünftige Verkäufe auf der Basis eines bestimmten Werbebudgets für TV vorhersagen

- Berechnung mittels:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

wobei $\widehat{y}$ die Vorhersage von Y auf Basis des Inputs $X = x$ bezeichnet.

Schätzung der Parameter

- Praxis: β_0 und β_1 unbekannt
- Bevor lineare Modell benützen $\quad \rightarrow \quad$ Koeffizienten schätzen
- Gehen von n Beobachtungspaaren aus:

$$(x_1, y_1), \quad (x_2, y_2), \quad \ldots, \quad (x_n, y_n)$$

- Jedes Paar besteht aus je einer Messung von X und Y
- Beispiel `Werbung`: $n = 200$ verschiedene Beobachtungspaare (Märkte)
  - x-Koordinate: TV-Budget
  - y-Koordinate: entsprechenden Produktverkäufen

- Ziel: $\widehat{\beta}_0$ und $\widehat{\beta}_1$ so zu bestimmen, dass die Gerade $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ möglichst gut zu den Daten passt

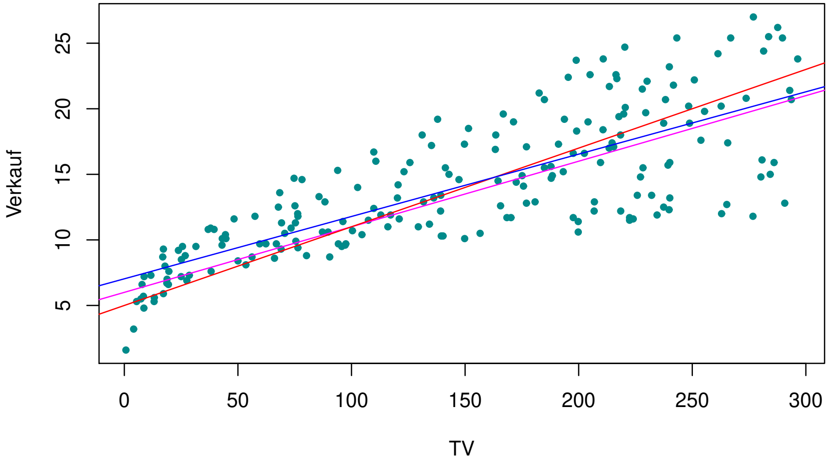
- Das heisst, dass
$$y_i \approx \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$
für alle $i = 1, \dots, n$

- Auf der linken Seite der obigen approximativen Beziehung steht der Messwert, auf der rechten der zugehörige y-Wert auf der Geraden

- Die Frage ist nun, was heisst „möglichst gut"?

Beispiel

- Abbildung: Einige Geraden eingezeichnet, die gut zu Datenpunkten passen



- Welche passt am besten?

- Punkte sollten möglichst nahe bei der gesuchten Geraden liegen
- Beispiel oben $\widehat{\beta}_0$ und $\widehat{\beta}_1$ so bestimmen, dass die resultierende Gerade so nahe wie möglich an den $n = 200$ Datenpunkten entlangläuft
- Was heisst aber so „nahe wie möglich"?
- Es gibt mehrere Methoden, um *Nähe* zu messen
- Die bei weitem gebräuchlichste: *Methode der kleinsten Quadrate*

Methode der kleinsten Quadrate

- Vorhergesagter Wert für Y abhängend vom i-ten Wert von X, also x_i:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

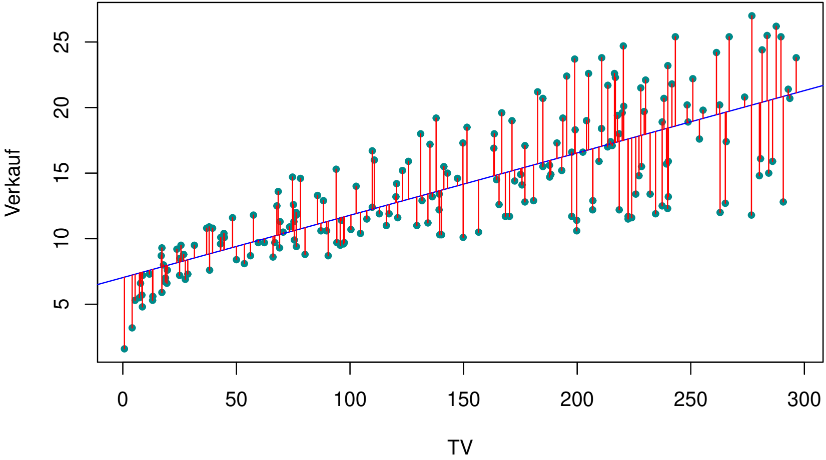
- i-tes *Residuum*:

$$r_i = y_i - \widehat{y}_i$$

- Differenz zwischen dem i-ten *beobachteten* Wert der Zielgrösse und dem i-ten von unserem linearen Modell *vorhergesagten* Wert der Zielgrösse

Beispiel

- Abbildung: Residuen als Strecken rot eingezeichnet



- Residuen oberhalb der Geraden positiv, unterhalb der Geraden negativ

- Summe der *Quadrate* der Residuen (RSS genannt)

- Es gilt dann

$$\mathrm{RSS} = r_1^2 + r_2^2 + \ldots + r_n^2$$

- Oder äquivalent:

$$\mathrm{RSS} = (y_1 - \widehat{\beta}_0 - \widehat{\beta}_1 x_1)^2 + (y_2 - \widehat{\beta}_0 - \widehat{\beta}_1 x_2)^2 + \ldots + (y_n - \widehat{\beta}_0 - \widehat{\beta}_1 x_n)^2$$

- Methode der kleinsten Quadrate: $\widehat{\beta}_0$ und $\widehat{\beta}_1$ so gewählt, dass RSS *minimal* wird

Für die, die es interessiert

- Mit Differentialrechnung: Für $\widehat{\beta}_0$ und $\widehat{\beta}_1$ gilt:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

mit

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{und} \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Mit Hilfe der Methode der kleinsten Quadrate geschätzte Koeffizienten für die einfache lineare Regression

Beispiel

- Beispiel `Werbung`: $\widehat{\beta}_0$ und $\widehat{\beta}_1$ und die Regressionsgerade bestimmen:

```
import statsmodels.api as sm

Y = werbung["Verkauf"]
X = werbung["TV"]
X = sm.add_constant(X)
## //usr/lib/python3/dist-packages/numpy/core/fromnumeric.py:24
##   return ptp(axis=axis, out=out, **kwargs)
fit = sm.OLS(Y,X).fit()

fit.params
## const    7.032594
## TV       0.047537
## dtype: float64
```

- Wert unter `const`: $\widehat{\beta}_0 \quad \rightarrow \quad y$-Achsenabschnitt

- Wert unter `TV`: $\widehat{\beta}_1 \quad \rightarrow \quad$ Steigung der Geraden

- Lineares Modell:

$$Y \approx 7.03 + 0.0475X$$

- `X = sm.add_constant(X)`: y-Achsenabschnitt wird auch berechnet

- Warnung: Kann man umgehen, wird aber unübersichtlicher (Verschlimmbesserung des Befehls)

- `fit.params` gibt nur einen Teil aus von

```
fit.summary()
```

- Output:

```
fit.summary()
```

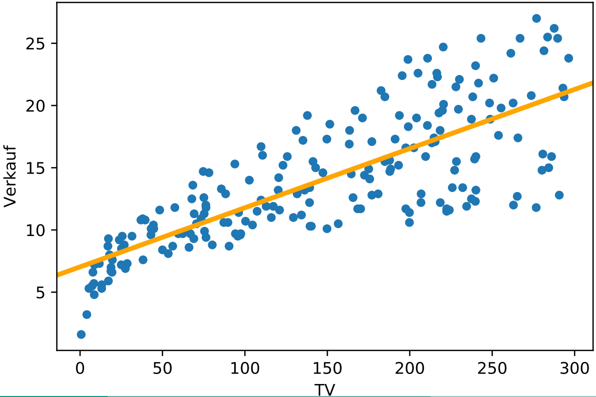
```
## <class 'statsmodels.iolib.summary.Summary'>
## """
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                Verkauf   R-squared:                       0.612
## Model:                            OLS   Adj. R-squared:                  0.610
## Method:                 Least Squares   F-statistic:                     312.1
## Date:                Mon, 04 May 2020   Prob (F-statistic):           1.47e-42
## Time:                        11:00:20   Log-Likelihood:                -519.05
## No. Observations:                 200   AIC:                             1042.
## Df Residuals:                     198   BIC:                             1049.
## Df Model:                           1
## Covariance Type:            nonrobust
## ==============================================================================
##                  coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------
## const          7.0326      0.458     15.360      0.000       6.130       7.935
## TV             0.0475      0.003     17.668      0.000       0.042       0.053
## ==============================================================================
## Omnibus:                        0.531   Durbin-Watson:                   1.935
## Prob(Omnibus):                  0.767   Jarque-Bera (JB):                0.669
## Skew:                          -0.089   Prob(JB):                        0.716
## Kurtosis:                       2.779   Cond. No.                         338.
## ==============================================================================
##
## Warnings:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
## """
```

- Gemäss Näherung: Für zusätzliche CHF 1000 Werbeausgaben werden 47.5 zusätzliche Einheiten des Produktes verkauft

- Abbildung mit Regressionsgerade

```
from statsmodels.graphics.regressionplots import abline_plot

ax = werbung.plot(kind="scatter",x="TV", y="Verkauf")
abline_plot(model_results=fit, ax = ax, color="orange", linewidth=3)
```



Wie genau sind Schätzungen für die Koeffizienten?

- Annahme: *wahrer* Zusammenhang von der Form

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

- f eine unbekannte Funktion

- ε ist ein zufälliger Fehlerterm mit Mittelwert 0

- Wird f durch eine lineare Funktion approximiert
  $\rightarrow$ Zusammenhang

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

- β_0 der y-Achsenabschnitt mit dem erwarteten Wert für Y, wenn $X = 0$

- β_1 ist die Steigung, also die mittlere "Anderung von Y bei einer Zunahme von X um eine Einheit

- Im Fehlerterm ε ist alles hineingepackt, was beim einfachen linearen Modell unterschlagen wurde:
  - Der wahre Zusammenhang ist selten linear
  - Es gibt vielleicht noch weitere Variablen, die Y beeinflussen
  - Es gab vielleicht Messfehler
- Für Summe von diesen zufälligen Variablen darf wegen des Zentralen Grenzwertsatzes eine Normalverteilung angenommen werden
- Weitere Annahme: Fehlerterm unabhängig von X ist

Beispiel

- Annahme: exakten Zusammenhang zwischen X und Y bekannt:

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

- Mit $f(X) = 2 + 3X$, also einer linearen Beziehung

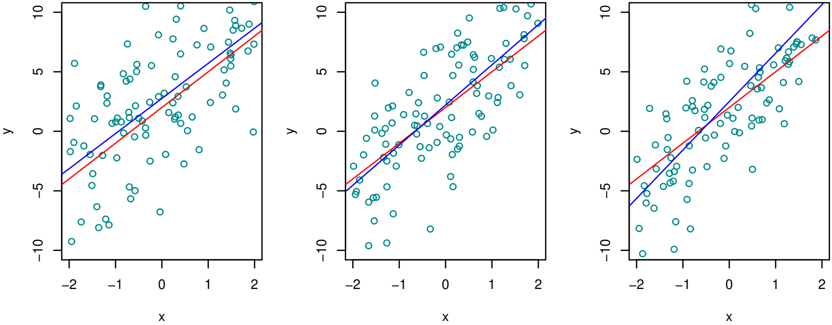
- Beobachteten Daten von Y simulieren durch

$$Y = 2 + 3X + \varepsilon$$

- ε normalverteilt mit Mittelwert 0 ist $\quad \rightarrow \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$

- Erzeugen 100 zufällige Werte von X mit zugehörigen Werte von Y

- Abbildung mit 3 solche Simulationen:

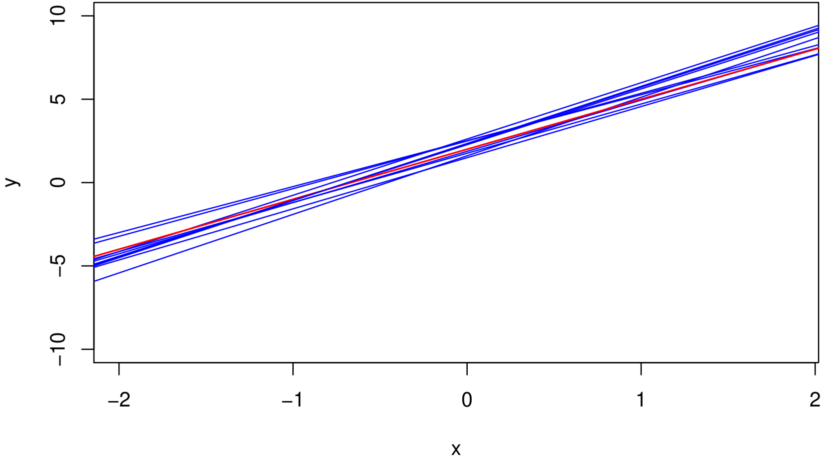


- Rote Gerade Graph der Gleichung $Y = 2 + 3X$ (in allen drei Simulationen gleich)

- Blaue Gerade: Regressionsgerade (Methode der kleinsten Quadrate)

- Diese ändert sich von Simulation zu Simulation

- Abbildung: sind die Regressionsgerade (blau) von 10 Simulationen eingezeichnet



- Zugrundeliegenden Gerade (rot) ähnlich sind, aber nie gleich

- Wahre Beziehung zwischen erklärender Variable X und Zielgrösse Y bei reellen Daten i.A. nie bekannt
- Regressionsgerade kann immer mit Hilfe der Methode der Kleinsten Quadrate bestimmt werden
- Anwendung: Daten, für welche Regressionsgerade bestimmt werden kann
- Wahre lineare Beziehung (sofern sie überhaupt existiert) bleibt immer unbekannt
- Beispiel oben: Kennen Datensätze und können die blauen Geraden bestimmen, aber kennen rote unbekannt.

- Wir haben also zwei (oder mehr) Geraden, die den Zusammenhang zwischen erklärender und Zielgrössen beschreiben

- Bloss kennen wir die Gleichung der wahren (roten) Geraden im Allgemeinen nicht

- Wir ziehen also aufgrund eines Datensatzes (blaue Gerade) Rückschlüsse auf den wahren Zusammenhang (rote Gerade), den wir aber nicht kennen

- Das ist allerdings die natürliche Vorgehensweise in der Statistik, wo von Beobachtungen auf die Gesamtheit geschlossen wird

Beispiel

- Durchschnittliche Körpergrösse μ aller 20-Jährigen auf der Erde
- Unmöglich, die Körpergrösse aller 20-Jährigen zu bestimmen
- Das heisst: μ schätzen
- Für ungefähren Wert $\widehat{\mu}$ für μ: wählen Gruppe von 1000 und ermitteln von diesen die Körperlänge y_i für $i = 1, \ldots, 1000$
- Durchschnitt $\overline{y}$
- Eine vernünftige Annahme ist
$$\widehat{\mu} = \overline{y}$$
- Also
$$\mu \approx \widehat{\mu} = \overline{y}$$
- Wählen wir eine andere Gruppe, so wird $\overline{y}$ leicht anders sein
- Aber dies ändert nichts an der Tatsache, dass $\mu \approx \overline{y}$

Vertrauensintervall: Beispiel

- Vertrauensintervall Beispiel `Werbung` mit `Python`:

```
fit.conf_int()
```

```
##                0         1
## const  6.129719  7.935468
## TV     0.042231  0.052843
```

- 95 %-Vertrauensintervall von β_0:

$$[6.130, 7.935]$$

- Für β_1:

$$[0.042, 0.053]$$

- Ohne Werbung: Verkauf zwischen 6130 und 7935 Einheiten

- Für zusätzliche CHF 1000 für TV-Werbung durchschnittlich zwischen 42 und 53 Einheiten mehr verkaufen

Hypothesentest: Statistische Signifikanz von β_1

- Häufigste Hypothesentest: Testen der *Nullhypothese*

$$H_0: \quad \text{Es gibt } \textit{keinen} \text{ Zusammenhang zwischen } X \text{ und } Y$$

- *Alternativhypothese*

$$H_A: \quad \text{Es gibt } \textit{einen} \text{ Zusammenhang zwischen } X \text{ und } Y$$

- Mathematisch:

$$H_0: \quad \beta_1 = 0$$

- Gegen:

$$H_A: \quad \beta_1 \neq 0$$

- $\beta_1 = 0$, dann:

$$Y = \beta_0 + \varepsilon$$

- Y hängt *nicht* von X ab
- Nullhypothese testen: $\widehat{\beta}_1$ genügend weit von 0 weg, damit β_1 nicht 0
- Mit t-Statistik

Beispiel

- *p*-Wert von β_1 im Beispiel `Werbung` berechnen:

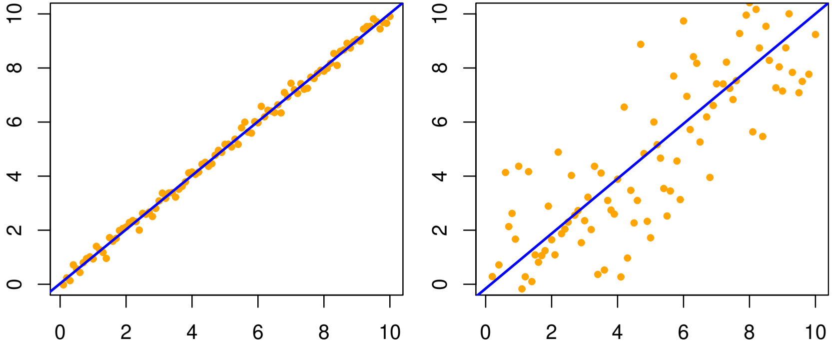
```
fit.pvalues
```

```
## const    1.406300e-35
## TV       1.467390e-42
## dtype: float64
```

- Eintrag `TV` (oder `P(>|t|)` bei `summary`): *p*-Wert 10^{-42}
- Bei weitem kleiner als 0.05
- Nullhypothesen $\beta_1 = 0$ verwerfen: $\beta_1 \neq 0$
- Klarer Hinweise für Zusammenhang zwischen `TV` und `Verkauf`

Abschätzung der Genauigkeit des Modells: R^2

- Nullhypothese verworfen: *In welchem Ausmass passt das Modell zu den Daten?*
- Abbildung:



  - Links: Steigende Gerade passt sehr gut zu Punkten
  - Rechts: Steigende Gerade passt *nicht* gut zu Punkten

- Qualität einer linearen Regression abgeschätzt durch den *residual standard error* (RSE) und die R^2-Statistik

- R^2 wichtiger

- R^2-Statistik: Wert zwischen 0 und 1

- Sie gibt an, welcher Anteil der Variabilität in Y mit Hilfe des Modells durch X erklärt werden

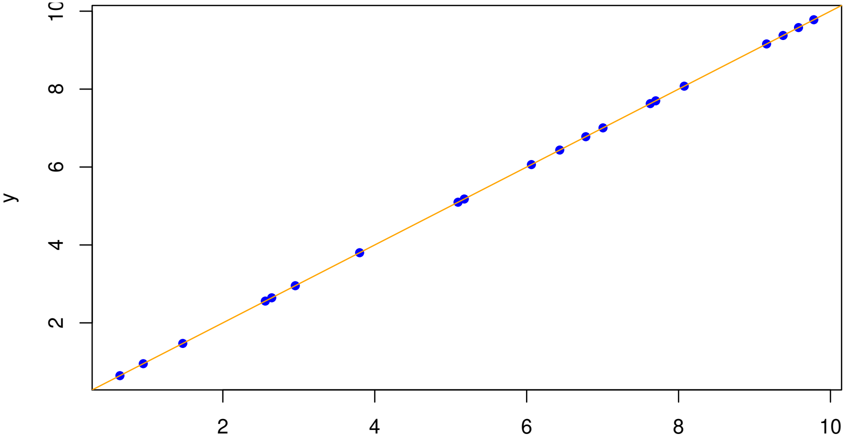
- Wert nahe bei 1: ein grosser Anteil der Variabiliät wird durch die Regression erklärt. Das Modell beschreibt also die Daten sehr gut.

- Wert nahe bei 0: Regression erklärt die Variabilität der Zielvariablen nicht

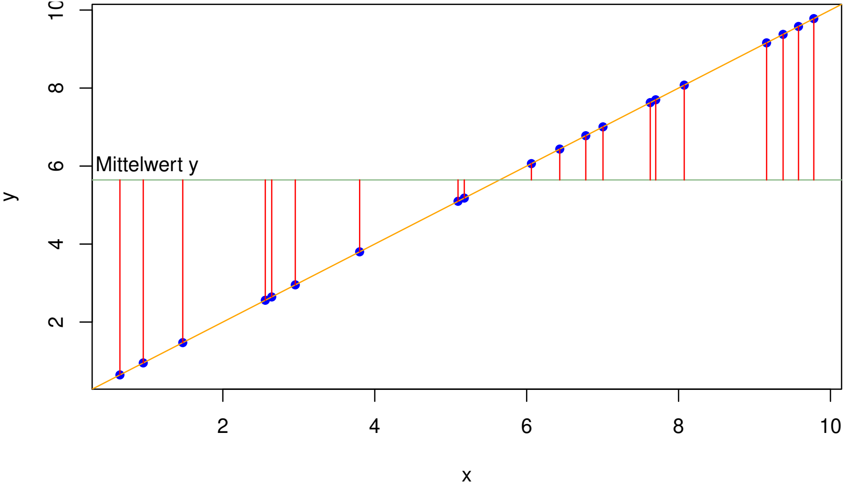
Punkte folgen linearem Modell

- Abbildung:

```r
x <- runif(min = 0, max = 10, n = 20)
y <- x
plot(x, y, col = "blue", pch = 16)
abline(lm(y ~ x), col = "orange")
```



- Abbildung Varianz:



- Varianz: „Mittelwert" der quadrierten Unterschiede der y-Werte der Datenpunkte zu $\overline{y}$

- Output:
  - Korrelation:

```
cor(x, y)
## [1] 1
```

  - R^2:

```
summary(lm(y ~ x))$r.squared
## [1] 1
```

  - Varianz:

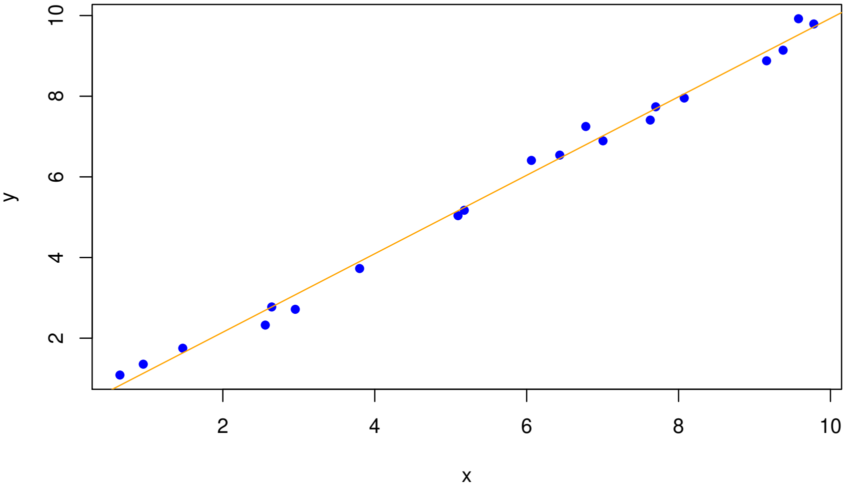
```
var(y)
## [1] 8.998626
```

  - 100% der Varianz von 9 wird durch das Modell erklärt

Punkte folgen mehr oder weniger linearem Modell

- Abbildung:

```r
y <- x + rnorm(n = 20, mean = 0, sd = 0.2)
```



- Output:
  - Korrelation:

```r
cor(x, y)
## [1] 0.9966885
```

  - R^2:

```r
summary(lm(y ~ x))$r.squared
## [1] 0.993388
```

  - Varianz:

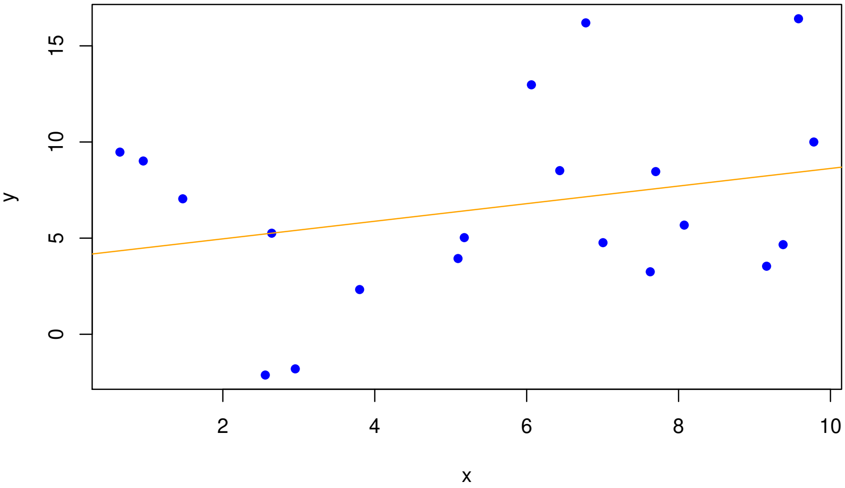
```r
var(y)
## [1] 8.573793
```

  - 99.34% der Varianz von 8.57 wird durch das Modell erklärt

Punkte folgen dem linearen Modell nicht

- Abbildung:

```
y <- x + rnorm(n = 20, mean = 0, sd = 4)
```



- Output:
  - Korrelation:

```r
cor(x, y)
## [1] 0.2769503
```

  - R^2:

```r
summary(lm(y ~ x))$r.squared
## [1] 0.07670148
```

  - Varianz:

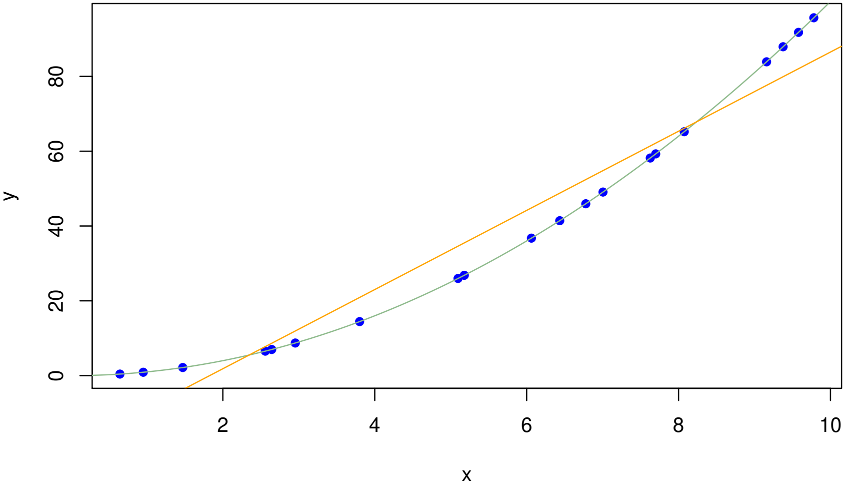
```r
var(y)
## [1] 24.55976
```

  - 7.67% der Varianz von 24.56 wird durch das Modell erklärt

Punkte folgen quadratischem Modell

- Abbildung:

```
y <- x^2
```



- Output:
  - Korrelation:

```
cor(x, y)
## [1] 0.9735588
```

  - R^2:

```
summary(lm(y ~ I(x^2)))$r.squared
## [1] 1
```

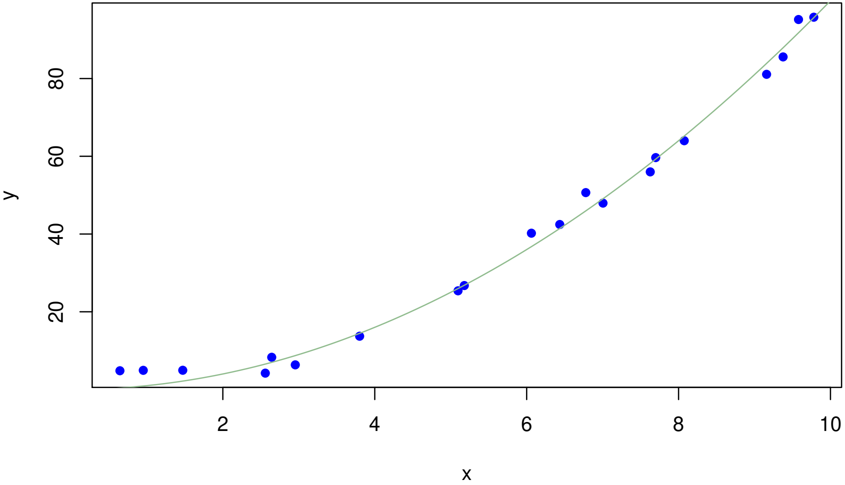
  - Varianz:

```
var(y)
## [1] 1063.22
```

  - 100% der Varianz von 1063.22 wird durch das Modell erklärt

- Punkte folgen Modell:

```
y <- x^2 + rnorm(n = 20, mean = 0, sd = 2)
```



- Output:
  - Korrelation:

```r
cor(x, y)
## [1] 0.9655864
```

  - R^2:

```r
summary(lm(y ~ I(x^2)))$r.squared
## [1] 0.9942619
```

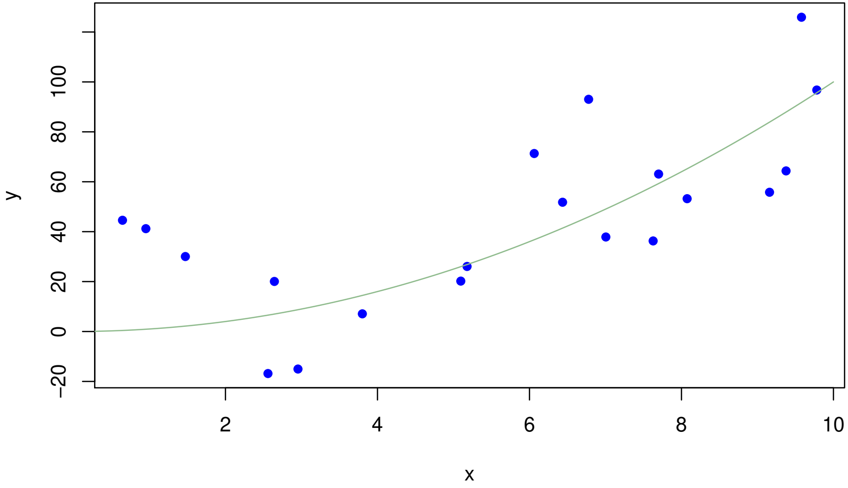
  - Varianz:

```r
var(y)
## [1] 1026.155
```

  - 99.43% der Varianz von 1026.15 wird durch das Modell erklärt

- Punkte folgen Modell:

```
y <- x^2 + rnorm(n = 20, mean = 0, sd = 20)
```



- Output:
    - Korrelation:

```r
cor(x, y)
## [1] 0.6644753
```

    - R^2:

```r
summary(lm(y ~ I(x^2)))$r.squared
## [1] 0.5335559
```

    - Varianz:

```r
var(y)
## [1] 1262.354
```

    - 53.36% der Varianz von 1262.35 wird durch das Modell erklärt

Beispiel

- Im Beispiel der TV-Werbung war der R^2-Wert 0.61

```
fit.rsquared
## 0.611875050850071
```

- Somit werden knapp zwei Drittel der Variabilität in `Verkauf` durch `TV` mit linearer Regression erklärt.