

Lineare Regression

Peter Büchel

HSLU I

ASTAT: Block 11

Lineare Regression

Peter Büchel

HSLU I

ASTAT: Block 11

Lineare Regression

- Fortsetzung von Block 3: Jetzt mit Hypothesentest
- Lineare Regression ist 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:

```
Werbung <- read.csv("../Data/Werbung.csv")[, -1]
head(Werbung, 3)
```

##		TV	Radio	Zeitung	Verkauf
##	1	230.1	37.8	69.2	22.1
##	2	44.5	39.3	45.1	10.4
##	3	17.2	45.9	69.3	9.3

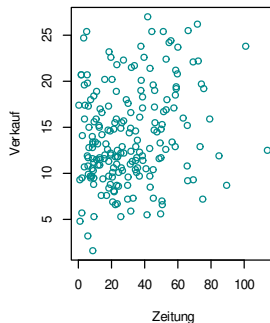
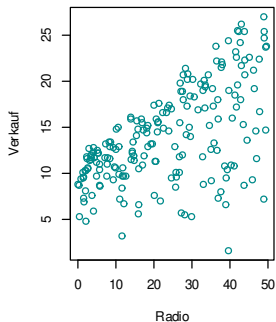
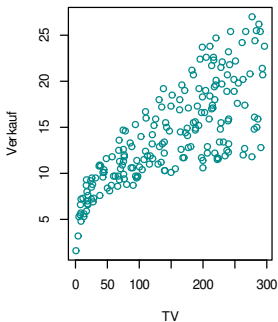
- Daten in Streudiagrammen in Abbildung dargestellt:

```
TV <- Werbung[, 1]  
Radio <- Werbung[, 2]  
Zeitung <- Werbung[, 3]  
Verkauf <- Werbung[, 4]
```

```
plot(Verkauf ~ TV, col = "darkcyan", xlab = "TV", ylab = "Verkauf")
```

```
plot(Verkauf ~ Radio, col = "darkcyan", xlab = "Radio", ylab = "Verkauf")
```

```
plot(Verkauf ~ Zeitung, col = "darkcyan", xlab = "Zeitung", ylab = "Verkauf")
```



- 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*

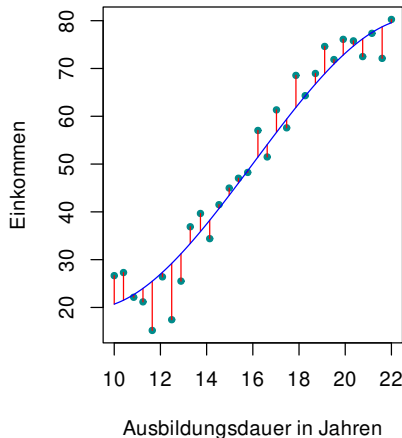
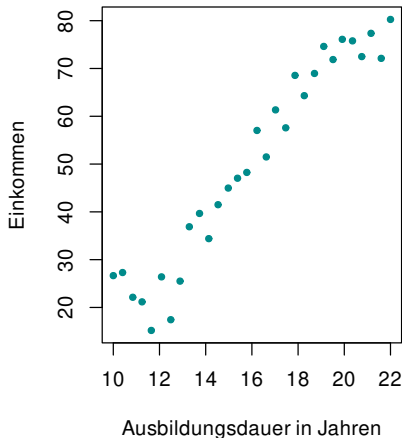
- Allgemein: Quantitative Zielgrösse Y und p verschiedene Prädiktoren X_1, X_2, \dots, X_p
- Annahme: Es besteht *irgendein* Zusammenhang zwischen Y und X_1, X_2, \dots, X_p
- Allgemeine Form:

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

- f irgendeine feste, aber *unbekannte* Funktion von X_1, X_2, \dots, X_p
- Grösse ε : *Zufälliger Fehlerterm* unabhängig von X_1, X_2, \dots, X_p mit Mittelwert 0
- Bedeutung Fehlerterm $\varepsilon \rightarrow$ 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 Prädiktoren und 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 $\hat{\cdot}$ gekennzeichnet
- \hat{Y} : Schätzung der unbekannten Grösse Y
- \hat{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, \dots, X_p einfach verfügbar, aber die Zielgrösse nicht
- In so einem Fall: Y schätzen durch

$$\hat{Y} = \hat{f}(X_1, X_2, \dots, X_p)$$

- Fehlerterm im Mittel 0

Beispiel

- Prädiktoren X_1, X_2, \dots, 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, \dots, 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 \hat{Y} als Vorhersage von Y hängt von zwei Größen ab:
 - ▶ *Reduzibler Fehler*
 - ▶ *Irreduzibler Fehler*
- Allgemein: \hat{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

$$\hat{Y} = f(X_1, X_2, \dots, X_p)$$

- Vorhersage Y enthält immer noch Fehler
- Liegt am Fehlerterm ε : Hängt nicht von X_1, X_2, \dots, X_p ab
- Variabilität 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 → 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 assoziiert?*
 - ▶ Natürlich alle, denkt man zuerst
 - ▶ Aber oft sind es einige wenige Variablen, die auf Y einen substantiellen Einfluss haben
 - ▶ Sehr viele Inputvariablen: *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 Prädiktoren haben einen positiven Zusammenhang mit der Outputvariable
 - ▶ Vergrößerung der Inputvariable hat in diesem Fall 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, \dots, X_p :

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

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

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

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

Beispiele

- Beispiel **Werbung**: Lineares Modell:

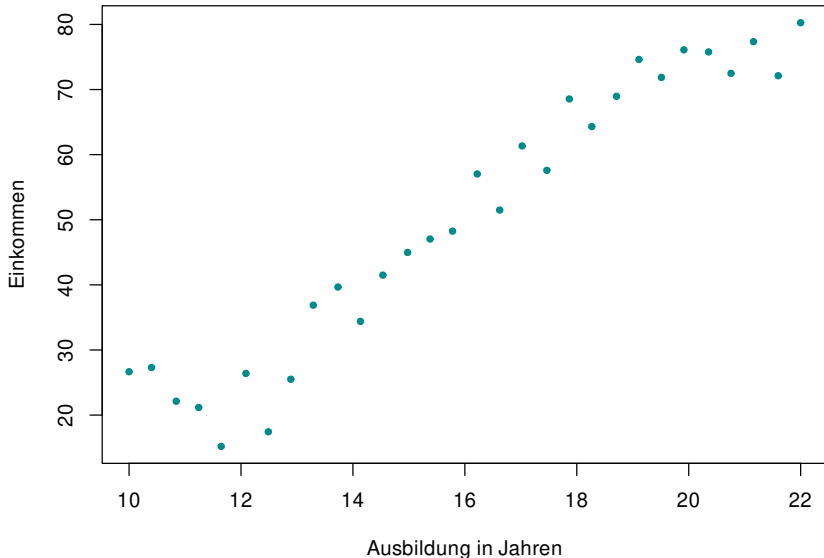
$$\text{Verkauf} \approx \beta_0 + \beta_1 \cdot \text{TV} + \beta_2 \cdot \text{Radio} + \beta_3 \cdot \text{Zeitung}$$

- Beispiel **Einkommen**: Lineares Modell:

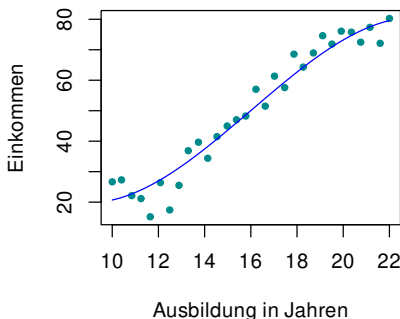
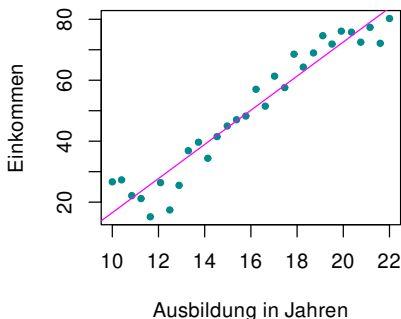
$$\text{Einkommen} \approx \beta_0 + \beta_1 \cdot \text{Ausbildung}$$

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):

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

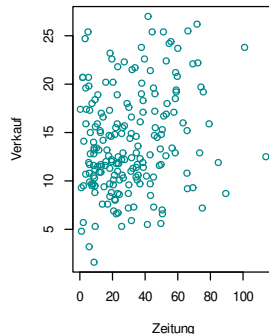
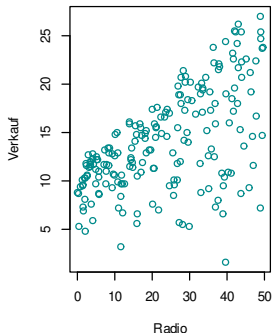
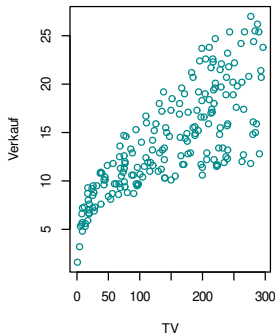
- Viele weitere Modelle denkbar
- Aber welches ist nun das „richtige“?
- Dies lässt sich in dieser Absolutheit nicht entscheiden
- Funktion f i. A. unbekannt: Liegt an uns „bestes“ Modell zu wählen
- Statistik: Bei Entscheidungsfindung behilflich
- Welches Modell ist in unserem Beispiel das „bessere“?
- 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?

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

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

- 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 $\hat{\beta}_0$ und $\hat{\beta}_1$ für die Modellkoeffizienten
- Sind diese Koeffizienten bekannt, so können zukünftige Verkäufe auf der Basis eines bestimmten Werbebudgets für TV vorhersagen
- Berechnung mittels:

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

wobei \hat{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 benutzen \rightarrow Koeffizienten schätzen
- Gehen von n Beobachtungspaaren aus:

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

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

- Das heisst, dass

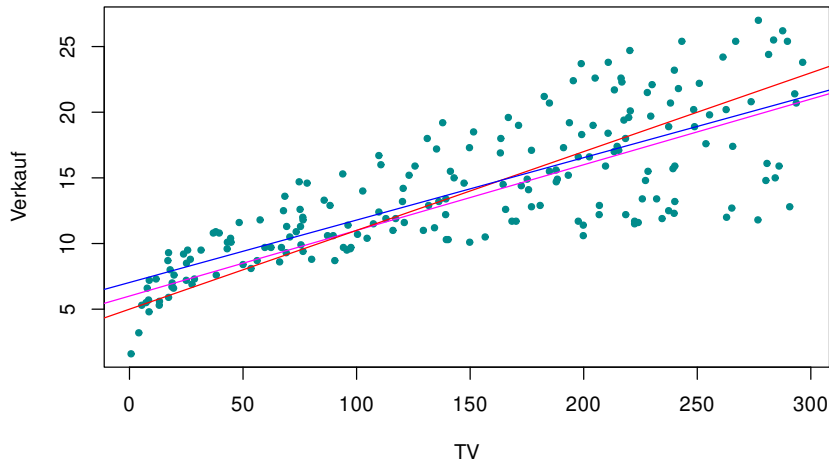
$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$$

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?

Methode der kleinsten Quadrate

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

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

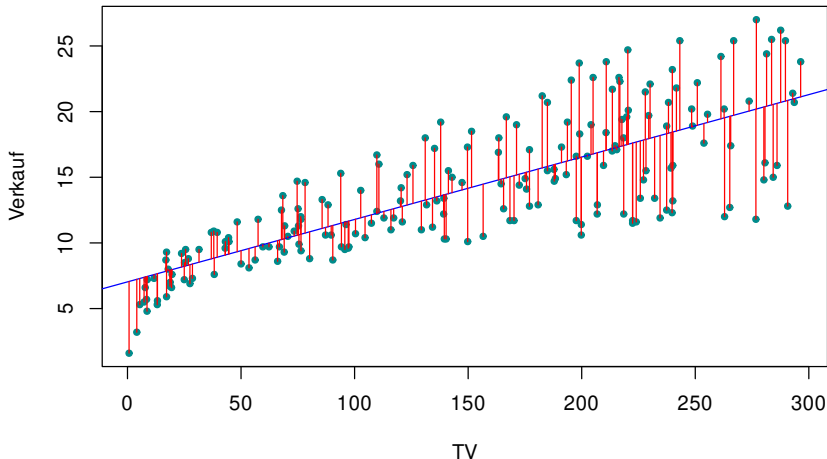
- i -tes *Residuum*:

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

- 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

$$\text{RSS} = r_1^2 + r_2^2 + \dots + r_n^2$$

- Oder äquivalent:

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

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

Für die, die es interessiert

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

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

mit

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

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

Beispiel

- Beispiel **Werbung**: $\hat{\beta}_0$ und $\hat{\beta}_1$ und die Regressionsgerade bestimmen:

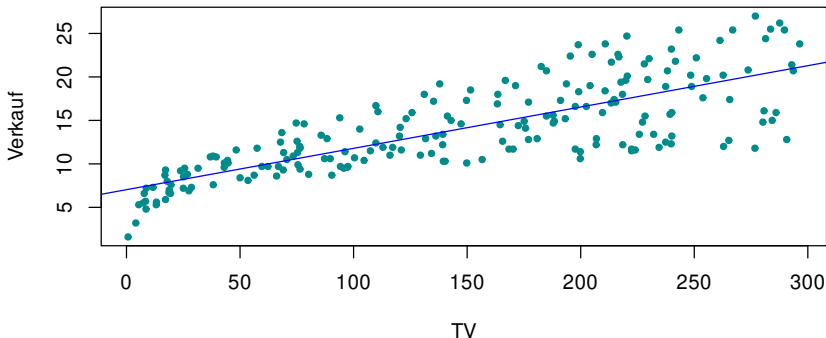
```
lm(Verkauf ~ TV)
##
## Call:
## lm(formula = Verkauf ~ TV)
##
## Coefficients:
## (Intercept)          TV
##      7.03259      0.04754
```

- Wert unter **Intercept**: $\hat{\beta}_0 \rightarrow y$ -Achsenabschnitt
- Wert unter **TV**: $\hat{\beta}_1 \rightarrow$ Steigung der Geraden
- Lineares Modell:

$$Y \approx 7.03 + 0.0475X$$

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

```
plot(TV, Verkauf, col = "darkcyan", xlab = "TV", ylab = "Verkauf",  
     pch = 20)  
  
abline(lm(Verkauf ~ TV), col = "blue")
```



Vertrauensintervall: Beispiel

- Vertrauensintervall Beispiel Werbung mit R:

```
confint(lm(Verkauf ~ TV), level = 0.95)
##                2.5 %      97.5 %
## (Intercept) 6.12971927 7.93546783
## TV          0.04223072 0.05284256
```

- 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

- Standardfehler: Hypothesentest für die Regressionsparameter durchführen
- Häufigste Hypothesentest: Testen der *Nullhypothese*

H_0 : Es gibt *keinen* Zusammenhang zwischen X und Y

- *Alternativhypothese*

H_A : Es gibt *einen* Zusammenhang zwischen X und Y

- Mathematisch:

$$H_0 : \beta_1 = 0$$

- Gegen:

$$H_A : \beta_1 \neq 0$$

- $\beta_1 = 0$, dann:

$$Y = \beta_0 + \varepsilon$$

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

Beispiel

- p -Wert von β_1 im Beispiel Werbung berechnen:

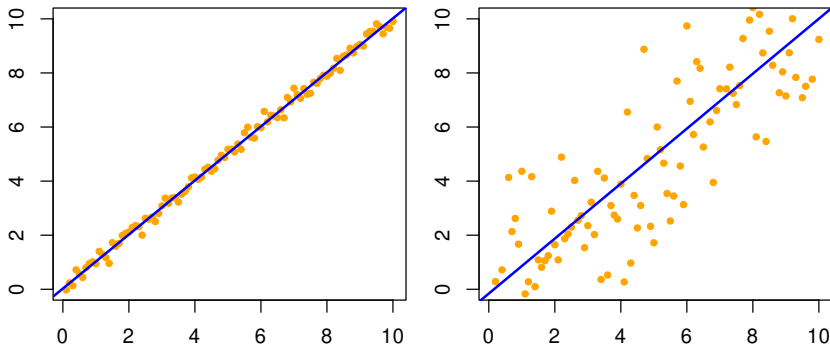
```
summary(lm(Verkauf ~ TV))

##
## Call:
## lm(formula = Verkauf ~ TV)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.032594   0.457843   15.36  <2e-16 ***
## TV           0.047537   0.002691   17.67  <2e-16 ***
## ---
## Signif. codes:
##  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF, p-value: < 2.2e-16
```

- Eintrag **Coefficients** unter **Pr(>|t|)**: p -Wert $2 \cdot 10^{-16}$
- 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:



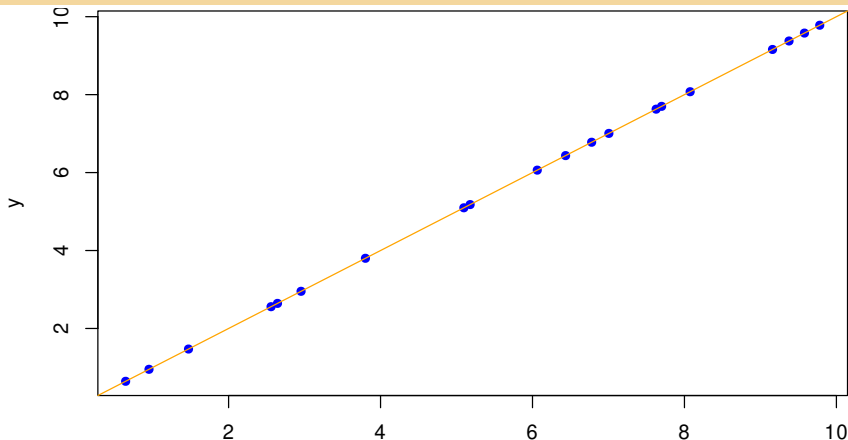
- ▶ Links: Steigende Gerade passt sehr gut zu Punkten
- ▶ Rechts: Steigende Gerade passt *nicht* gut zu Punkten

- 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 Variabilität wird durch die Regression erklärt. Das Modell beschreibt also die Daten sehr gut.
- Wert nahe bei 0: Regression erklärt die Variabilität der Zielvariablen nicht

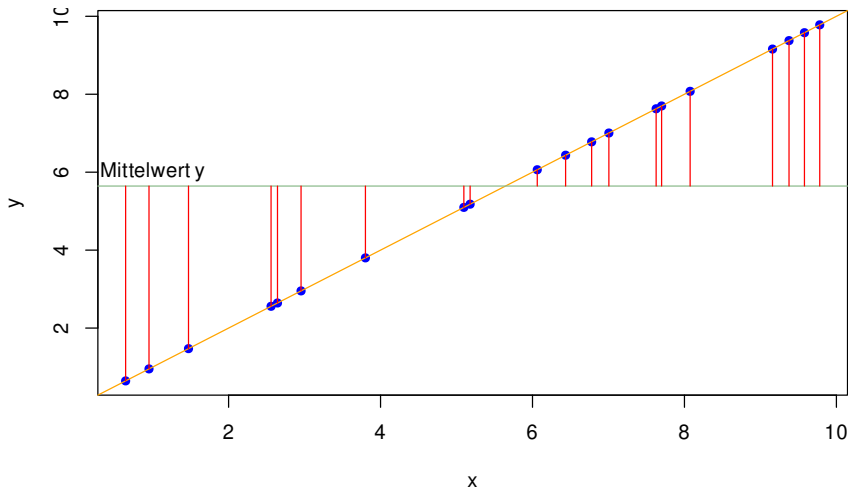
Punkte folgen linearem Modell

- Abbildung:

```
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 \bar{y}

- Output:

- ▶ Korrelation:

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

- ▶ R^2 :

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

- ▶ Varianz:

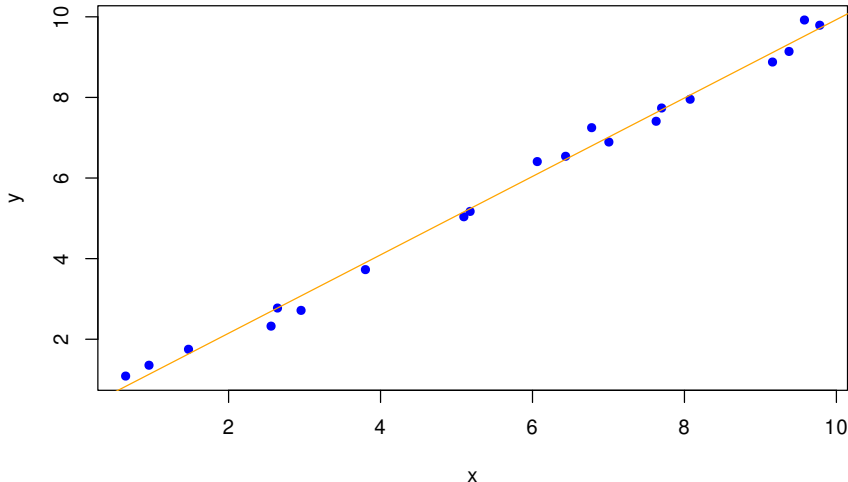
```
var(y)
## [1] 8.998626
```

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

Punkte folgen mehr oder weniger linearem Modell

- Abbildung:

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



- Output:

- ▶ Korrelation:

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

- ▶ R^2 :

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

- ▶ Varianz:

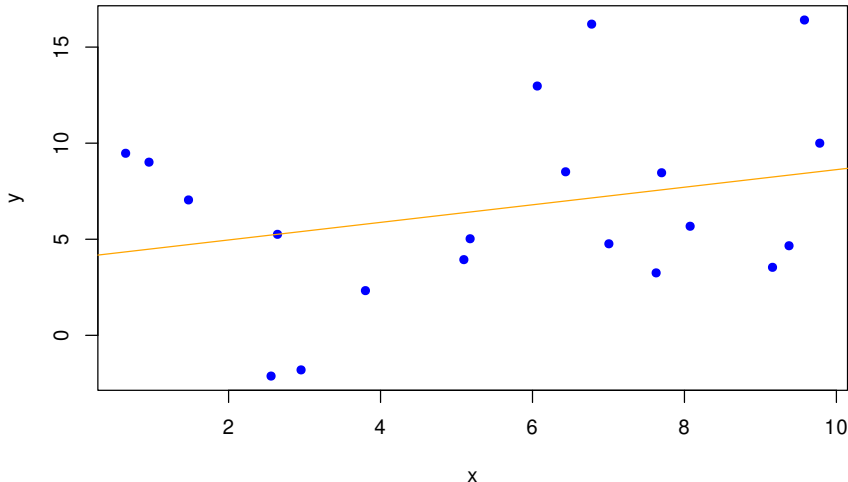
```
var(y)
## [1] 8.573793
```

- ▶ 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:

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

- ▶ R^2 :

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

- ▶ Varianz:

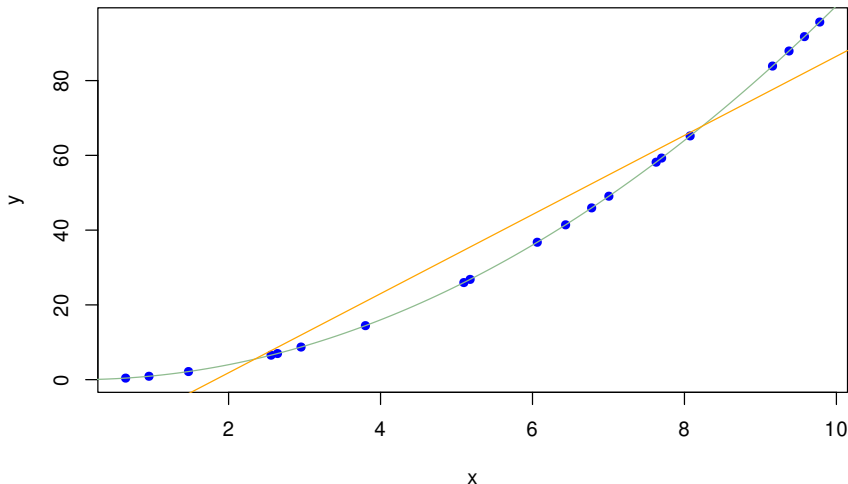
```
var(y)
## [1] 24.55976
```

- ▶ 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
```

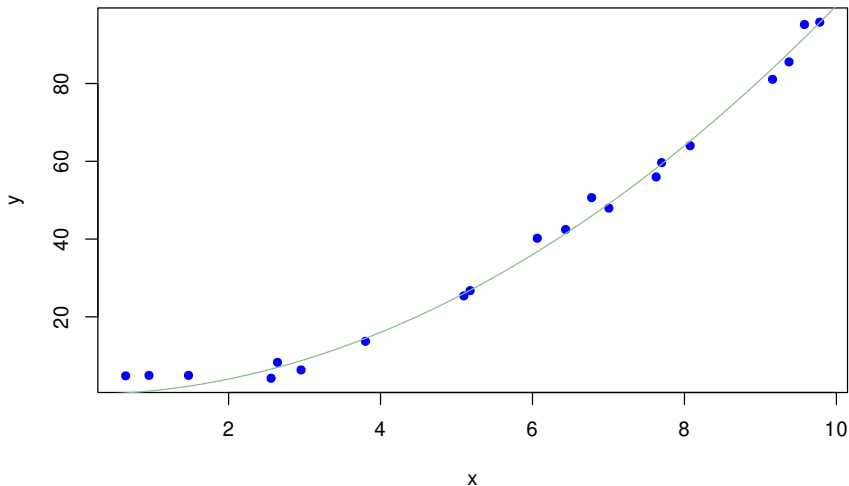
- ▶ Varianz:

```
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:

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

- ▶ R^2 :

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

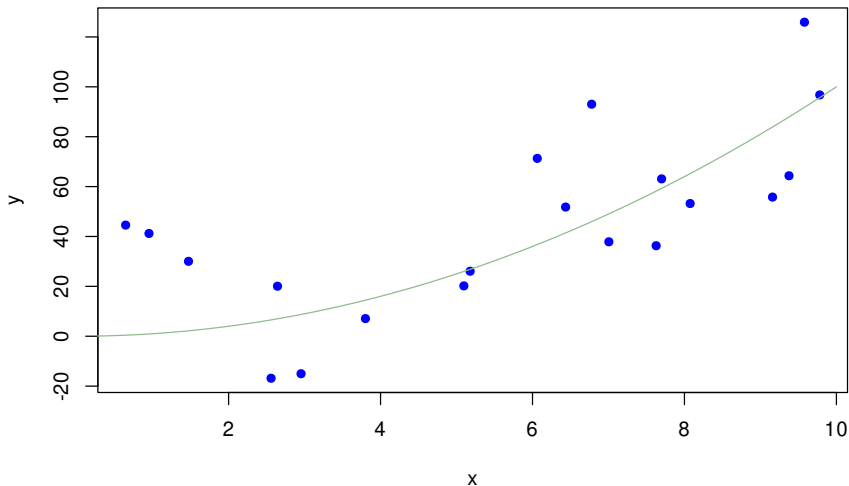
- ▶ Varianz:

```
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:

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

- ▶ R^2 :

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

- ▶ Varianz:

```
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

```
summary(lm(Verkauf ~ TV))$r.squared  
## [1] 0.6118751
```

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