Blockpraktikum zur Statistik mit R

17. September 2014

Johannes Blank

Institut für Mathematische Statistik
Universität Münster
WS 14/15

Beispiel: Ausgangsfrage

- Wie wirkt sich die eingesetzte Menge des Düngers auf den Ernteertrag aus?
- Wie stark ist der Zusammenhang zwischen der eingesetzten Menge eines Düngemittels und der Erntemenge?

Ziel:

- **Quantifizierung des Zusammenhanges** zweier Merkmalsausprägungen

Gliederung

1. Lineare Zusammenhänge erkennen
   - Korrelationskoeffizienten
   - Korrelation vs. Kausalität
2. Lineare Regressionsmodell
   - Methode der kleinsten Quadrate
   - Residuenanalyse und Bestimmtheitsmaß
3. Multiple lineare Regression
   - Die Modellierung
   - Der KQ-Schätzer
4. Logit Modelle
   - Modellierung binärer Regression
   - Beispiel und Parameterschätzung

Gliederung

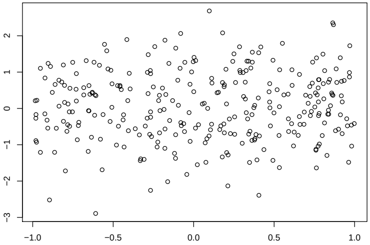
Lineares Modell

Modellannahme: Zwischen zwei Merkmalen X und Y besteht ein *zufällig gestörter* funktionaler Zusammenhang der Form

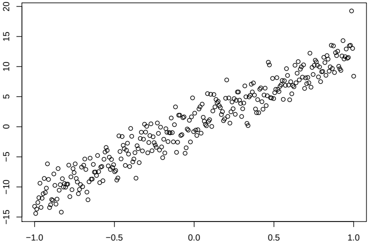
$$Y \;=\; f(X) \;+\; \epsilon$$

- Y ist im Wesentlichen eine *deterministische lineare* Funktion von X
- Der *zufällige Fehler* ϵ geht additiv in das Modell ein
- Gilt $Y_i = \alpha + \beta X_i + \epsilon_i \rightsquigarrow$ (einfaches) Lineares Regressionsmodell
- Gilt $Y_i = \alpha + \beta_1 X_i^1 + \ldots + \beta_p X_i^p + \epsilon_i \rightsquigarrow$ Polynomiale Regression
- Gilt $Y_i = \alpha + \beta_1 X_{i,1} + \ldots + \beta_p X_{i,p} + \epsilon_i \rightsquigarrow$ Multiple lineare Regression
- Streudiagramm der Daten: Gibt Hinweis auf einen Zusammenhang

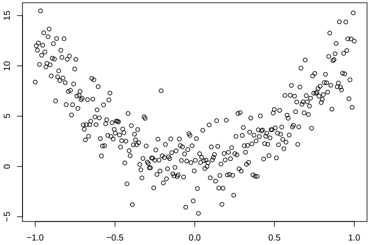
Kein Zusammenhang



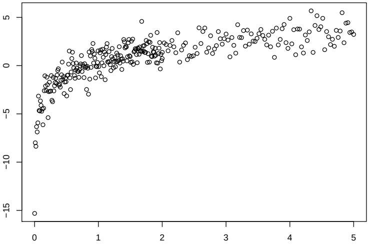
Linearer Zusammenhang



Quadratischer Zusammenhang



logistischer Zusammenhang



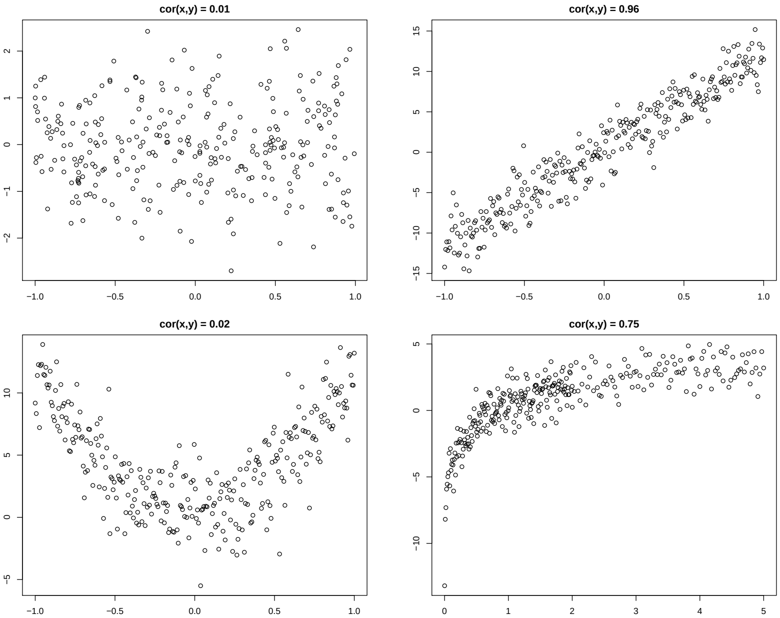
Bravais-Pearson-Korrelationskoeffizient

Bei linearem Zusammenhang ist der *empirische Korrelationskoeffizient* (auch *Bravais-Pearson-Korrelationskoeffizient*) ein Maß für die Stärke des Zusammenhangs:

$$r \ := \ r_{xy} \ := \ \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

für Beobachtungen $(x_1, y_1), \cdots, (x_n, y_n)$.

- $r > 0$ deutet auf *gleichsinnigen* linearen Zusammenhang (d.h. die Werte liegen um eine Gerade mit positiver Steigung)
- $r < 0$: *gegensinniger* linearen Zusammenhang
- $r = 0$: unkorreliert, kein linearer Zusammenhang


cor(x,y) = 0.01
cor(x,y) = 0.96
cor(x,y) = 0.02
cor(x,y) = 0.75

Eigenschaften von r

- r ist das empirische Gegenstück zum theoretischen Korrelationskoeffizienten (Ersetzung der (Ko)Varianzen durch empirische Gegenstücke)
- r nimmt Werte in $[-1, 1]$ an
- $|r| = 1$: Punkte liegen genau auf einer Geraden
- Berechnung in `R` mit `cor(x,y)`

Grobes Einteilungsraster von Korrelationen:

„schwache Korrelation“	$\lvert r \rvert < 0.5$
„mittlere Korrelation“	$0.5 \leq \lvert r \rvert < 0.8$
„starke Korrelation“	$0.8 \leq \lvert r \rvert$

Spearman'scher Korrelationskoeffizient

Bei Vermutung auf monotonen Zusammenhang $\rightsquigarrow$ bilde *Spearman'schen Korrelationskoeffizient* (robuster als Bravais-Pearson)

$$r_{\mathrm{SP}} = \frac{\sum_{i=1}^{n}(\mathrm{rg}(x_i) - \overline{\mathrm{rg}}_x)(\mathrm{rg}(y_i) - \overline{\mathrm{rg}}_y)}{\sqrt{\sum_{i=1}^{n}(\mathrm{rg}(x_i) - \overline{\mathrm{rg}}_x)^2 \sum_{i=1}^{n}(\mathrm{rg}(y_i) - \overline{\mathrm{rg}}_y)^2}} \quad (= r_{\mathrm{rg}(x)\,\mathrm{rg}(y)})$$

wobei $\mathrm{rg}(x_i)$ = Rang von x_i, $\mathrm{rg}(x) = (\mathrm{rg}(x_1), \ldots, \mathrm{rg}(x_n))$ der Vektor der Ränge von x und $\overline{\mathrm{rg}}_x = n^{-1} \sum_{i=1}^{n} \mathrm{rg}(x_i) = (n+1)/2$.

- $r_{\mathrm{SP}} > 0$: gleichsinniger monotoner Zusammenhang
  also: „x groß $\Leftrightarrow$ y groß“ sowie: „x klein $\Leftrightarrow$ y klein“
- $r_{\mathrm{SP}} < 0$: gegensinniger monotoner Zusammenhang
- $r_{\mathrm{SP}} = 0$: kein monotoner Zusammenhang
- $|r_{\mathrm{SP}}| = 1$: Die Punkte $(\mathrm{rg}(x_i), \mathrm{rg}(y_i))$, $i = 1, \ldots, n$ liegen auf einer Geraden

Spearman'scher Korrelationskoeff.: Bsp & R-Befehle

Für $x = (7, 5, 9, 6)$ bzw. $y = (7, 9, 6, 7)$ gilt

- $\mathrm{rg}(x_1) = 3$ (7 ist drittkleinster Wert in x)
- $\mathrm{rg}(x_2) = 1$ (5 ist kleinster Wert in x)
- $\mathrm{rg}(y_1) = \mathrm{rg}(y_4) = \frac{2+3}{2} = 2.5$
- Also: $\mathrm{rg}(x) = (3, 1, 4, 2)$ bzw. $\mathrm{rg}(y) = (2.5, 4, 1, 2.5)$
- Somit: $r_{\mathrm{SP}} = -0.95$

Die zugehörigen R-Befehle lauten

- `rank(x)` für die Rang-Statistik von x
- `cor(x,y,method="spearman")` für r_{SP}

Korrelation vs. Kausalität

- Korrelation erfasst die Stärke von Zusammenhängen, *keine* Wirkungen
- Kausalzusammenhänge lassen sich nur aus inhaltlichen Überlegungen begründen
- Dies erfordert tiefere Kenntnisse aus dem entspr. Forschungsgebiet
- Bsp: X beeinflusst zwar Y, aber nicht direkt, sondern über ein direktes Merkmal Z
- Nichtberücksichtigung von Z führt dann zu falschen Schlüssen

Scheinkorrelation

Bei 5 Kindern wurde der Wortschatz X und die Körpergröße Y gemessen

x_i	37	30	20	28	35
y_i	130	112	108	114	136

- Korrelation ist hoch: `cor(x,y) = 0.863`
- Sachlogisch lässt sich eine Beeinflussung nicht begründen
- Merkmal Alter (Z) muss hier berücksichtigt werden

z_i	12	7	6	7	13

- Es gilt: `cor(x,z) = 0.867` und `cor(y,z) = 0.995`

Scheinkorrelation

Bei 5 Kindern wurde der Wortschatz X und die Körpergröße Y gemessen

x_i	37	30	20	28	35
y_i	130	112	108	114	136

- Korrelation ist hoch: `cor(x,y) = 0.863`
- Sachlogisch lässt sich eine Beeinflussung nicht begründen
- Merkmal Alter (Z) muss hier berücksichtigt werden

z_i	12	7	6	7	13

- Es gilt: `cor(x,z) = 0.867` und `cor(y,z) = 0.995`

Verdeckte Korrelation

Nichtbeachtung eines Merkmals kann Korrelation verschleiern oder das Vorzeichen eines Zusammenhanges ändern:
z.B. Einfluss auf statistische Objekte mit unbeachteten Merkmal Neu ist genau entgegengesetzt bei den Ausprägung A und Ausprägung B. Die Nichtberücksichtgung des Merkmals liefert stattdessen eine Korrelation von 0.

Die Aufgaben 9.1 - 9.3 des Aufgabenblattes können jetzt bearbeitet werden.

Gliederung

(homoskedastische) lineare Regression

Einfaches lineares Modell:

$$y_i = \alpha + \beta x_i + \epsilon_i \quad i = 1, \dots, n$$

wobei

- $y_1, \dots, y_n$ beobachtbare Werte einer metrischen Zielvariablen (z.B. Ernteertrag)
- $x_1, \dots, x_n$ gegebene *Kontrollparameter* (z.B. Menge des Düngemittels)
- α und β unbekannte, zu schätzende Parameter
- $\epsilon_1, \dots, \epsilon_n$ (unbeobachtbare) Zufallsvariablen mit:
  - $\epsilon_1, \dots, \epsilon_n$ sind unabhängig
  - $\mathbb{E}(\epsilon_i) = 0$ für alle i
  - $\mathrm{Var}(\epsilon_i) = \sigma^2$ für alle i (*homoskedastisches* Verhalten)

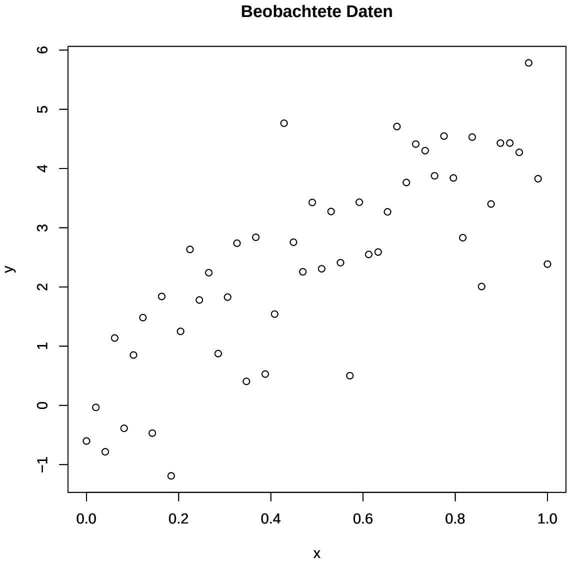
Methode der kleinsten Quadrate

- **Ziel:** Schätze anhand der Daten x und y die Parameter α und β so, dass die Daten „gut an das Modell angepasst sind"
- Methode der kleinsten Quadrate: Suche $\widehat{\alpha}$, $\widehat{\beta}$ mit

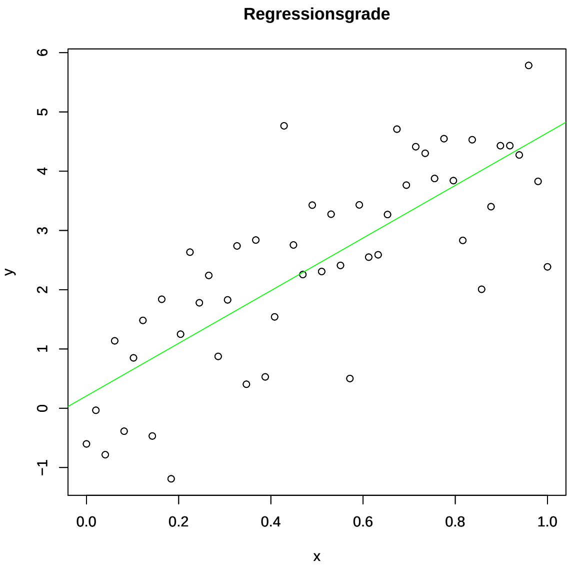
$$\sum_{i=1}^{n}(y_i - \widehat{\alpha} - \widehat{\beta}x_i)^2 = \min_{(\alpha,\beta)} \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

$\rightsquigarrow$ (summierte) quadratische Abstand zwischen den beobachteten Daten und der Geraden $x \mapsto \widehat{\alpha} + \widehat{\beta}x$ ist minimal

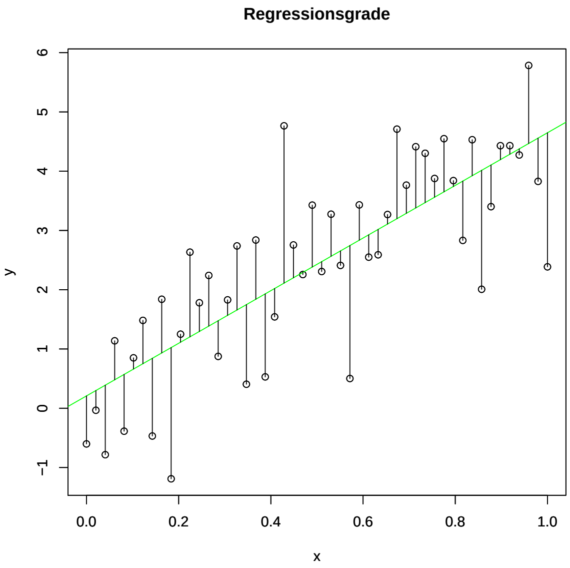
Regressionsgerade



Regressionsgerade



Regressionsgerade



Kleinste-Quadrate-Schätzer

- Kurvendiskussion der Funktion $(\alpha, \beta) \mapsto \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ liefert die *Kleinste-Quadrate-Schätzer (KQS)* $\widehat{\alpha}$ und $\widehat{\beta}$:

$$\widehat{\alpha} = \overline{y} - \hat{\beta} \cdot \overline{x} \qquad \text{und} \qquad \widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- Lineare Regression in `R`: Aufruf von `lm(y~x)`
- Wenn man sachlogisch $\alpha = 0$ annimmt: Aufruf `lm(y~x+0)`
- `lm()` gibt eine spezielle Liste (`lm`-Objekt) zurück, die man besser immer zwischenspeichert
- Durch den anschließenden Aufruf von `summary(lm)` erhält man eine detaillierte Ausgabe der Modellanpassung
- Einfügen der Regressionsgerade in das Streudiagramm: `abline(lm)`

Residuen

- Abweichungen $\widehat{\epsilon}_i$ zwischen Beobachtungen y_i und den durch das Modell vorhergesagten y-Werten $\widehat{y}_i$ nennt man *Residuen*
- Bei linearer Regression mit KQS: $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} \cdot x_i$, also

$$\widehat{\epsilon}_i \; = \; y_i - \widehat{y}_i \; = \; y_i - \widehat{\alpha} - \widehat{\beta} \cdot x_i$$

- **Ziel:** „Güte des Modells“ anhand der Residuen beurteilen
- Mögliches Gütemaß: Welcher Anteil der Streuung der y_i lässt sich durch die Regression erklären? (dieser Wert sollte möglichst groß sein)
- Gesamtstreuung SQT *(sum of squares total)* ist gegeben durch

$$\mathrm{SQT} \; = \; \sum_{i=1}^{n} (y_i - \overline{y})^2$$

Residuen

- Abweichungen $\widehat{\epsilon}_i$ zwischen Beobachtungen y_i und den durch das Modell vorhergesagten y-Werten $\widehat{y}_i$ nennt man *Residuen*
- Bei linearer Regression mit KQS: $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} \cdot x_i$, also

$$\widehat{\epsilon}_i = y_i - \widehat{y}_i = y_i - \widehat{\alpha} - \widehat{\beta} \cdot x_i$$

- **Ziel:** „Güte des Modells" anhand der Residuen beurteilen
- Mögliches Gütemaß: Welcher Anteil der Streuung der y_i lässt sich durch die Regression erklären? (dieser Wert sollte möglichst groß sein)
- Gesamtstreuung SQT *(sum of squares total)* ist gegeben durch

$$\mathrm{SQT} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

Streuungszerlegung

- SQT lässt sich wie folgt zerlegen:

$$\begin{aligned} \mathrm{SQT} &= \mathrm{SQE} + \mathrm{SQR} \\ \sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \end{aligned}$$

- $\mathrm{SQE} \mathrel{\widehat{=}}$ durch das Modell erklärte Abweichungen (*sum of squares explained*)
- $\mathrm{SQR} \mathrel{\widehat{=}}$ Streuung der Residuen (*sum of squares residuals*)

⇝ Der Anteil der Streuung der y_i, der durch die Regression erklärt wird, ist durch $\mathrm{SQE}\,/\,\mathrm{SQT}$ gegeben

- Diesen Wert nennt man das *Bestimmtheitsmaß* R^2

Bestimmtheitsmaß R^2

Für R^2 gilt

$$R^2 = \frac{\mathrm{SQE}}{\mathrm{SQT}} = 1 - \frac{\mathrm{SQR}}{\mathrm{SQT}}$$

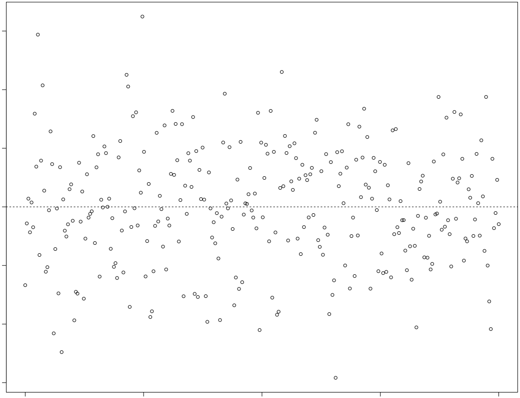
- $R^2 \in [0, 1]$
- Zusammenhang mit ▶ Bravais-Pearson-Korrelationskoeffizienten r:

$$R^2 = r_{x,y}^2 \qquad (\textit{längere Rechnung})$$

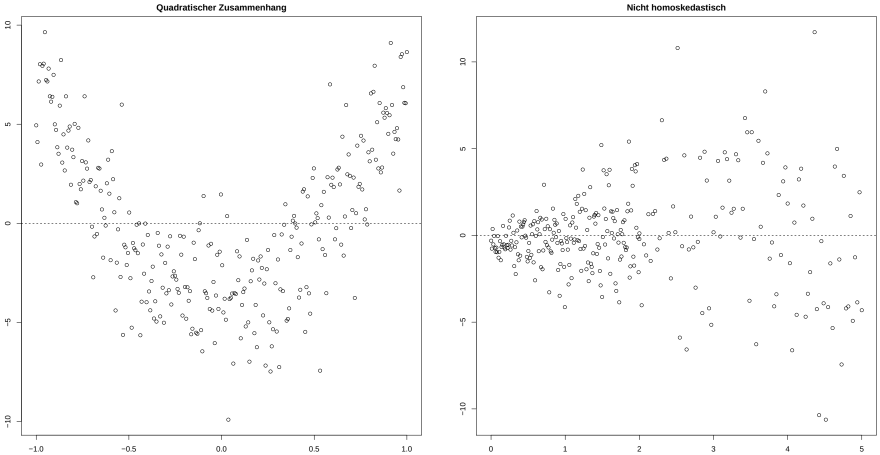
- $R^2 = 1$ heißt: Daten x, y liegen auf einer Geraden
- $R^2 = 0$: Erklärte Streuung ist 0 (Regressionsmodell ist ungeeignet)

Residualplots

- Anhand der Residuen können die Modellannahmen überprüft werden
- Residuen sollten mit ähnlicher Schwankungsbreite (homoskedastisch) um den Nullpunkt ($\mathbb{E}(\epsilon_i) = 0$) streuen



Residualplots: Beispiele



Struktur von `lm()`

- Betrachte Körpergröße und Alter von 5 Kindern (siehe Folie 2) und erstelle mit `model<-lm(groesse~alter)` ein lineares Modell
- `model` ist eine Liste, deren Struktur man sich mit `str(model)` anzeigen lassen kann
- In `model$coefficients` stehen die KQS
- In `model$residuals` die Residuen
- Mit `summary(model)` erhät man eine detaillierte Information über das Modell

Ausgabe von `summary(model)`

```
Call:
lm(formula = y ~ z)

Residuals:
      1       2       3       4       5
-1.2857 -0.4762 -0.7143  1.5238  0.9524

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  86.1429     1.9967   43.14 2.74e-05 ***
z             3.7619     0.2112   17.81 0.000386 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.369 on 3 degrees of freedom
Multiple R-squared:  0.9906, Adjusted R-squared: 0.9875
F-statistic: 317.3 on 1 and 3 DF,  p-value: 0.0003857
```

- KQS: $\widehat{\alpha} = 86.1429$, $\widehat{\beta} = 3.7619$ (in der Spalte `Estimate`)
- $R^2 = 0.9906$ (`Multiple R-squared`)
- Es werden Tests durchgeführt, ob die Parameter α und β signifikant in das Modell eingehen: Bei vielen Sternchen ist das der Fall, bei keinem Sternchen ist das Modell ungeeignet

Die Aufgaben 9.4. bis 9.6. des Aufgabenblattes können jetzt bearbeitet werden.

Gliederung

Multiples Regressionsmodell

Allgemeine Modellannahme:

$$Y_i \;=\; \theta_1 \cdot X_{i,1} \;+\; \ldots \;+\; \theta_p \cdot X_{i,p} \;+\; \epsilon_i, \qquad i = 1, \ldots, n$$

bzw. in Matrixnotation

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \underbrace{\begin{pmatrix} X_{1,1} & \cdots & X_{1,p} \\ \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,p} \end{pmatrix}}_{=\mathbf{X}} \cdot \underbrace{\begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}}_{=\theta} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

mit:

- *Designmatrix* $\mathbf{X}$ hat vollen Rang ($\Rightarrow$ *rang*$(\mathbf{X}) = p \leq n$)
- Fehler $\epsilon_1, \ldots, \epsilon_n$ unabh. mit Erwartungswert 0, Varianz σ^2

Spezialfälle des Modells

- Der Fall $p = 2$, $X_{i,1} = 1$ für alle i ist das einfache Lineare Regressionsmodell
- Den Fall $X_{i,k} = X_i^k$ nennt man *polynomiales Regressionsmodell*

$\rightsquigarrow$ Hier hängt Y_i polynomial von *einer* Variablen X_i ab, d.h.

$$Y_i = P(X_i) + \epsilon_i$$

für ein Polynom P mit Grad p

KQ-Schätzer für θ

- Gesucht: „guter“ Schätzer für θ
- Auch hier erhält man durch die Methode der kleinsten Quadrate einen KQS $\widehat{\theta} = \widehat{\theta}(y)$ für θ

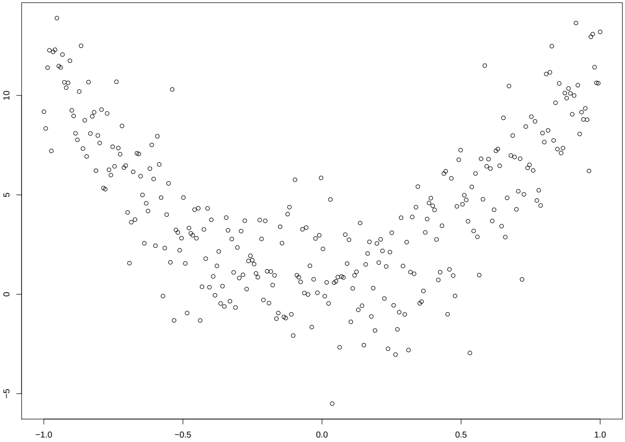
$$\widehat{\theta} \;=\; (\mathbf{X}^t \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^t \cdot y$$

  für eine Beobachtung $y = (y_1, \ldots, y_n)$

- Berechnung erfolgt numerisch
- In R mit `lm(y~x1+x2+...+xp)`

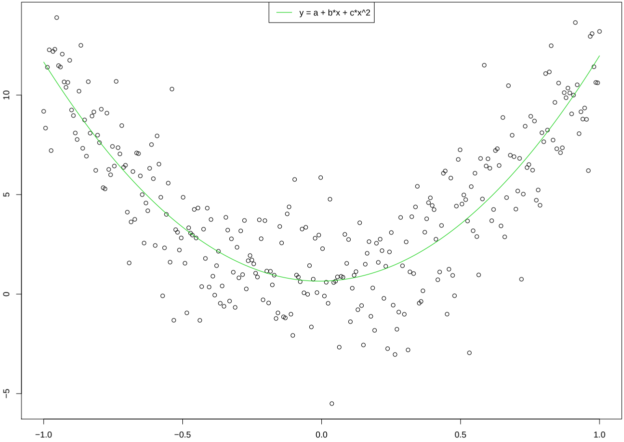
Beispiel: Polynomielle Regression

Betrachte Beispiel von Folie 6



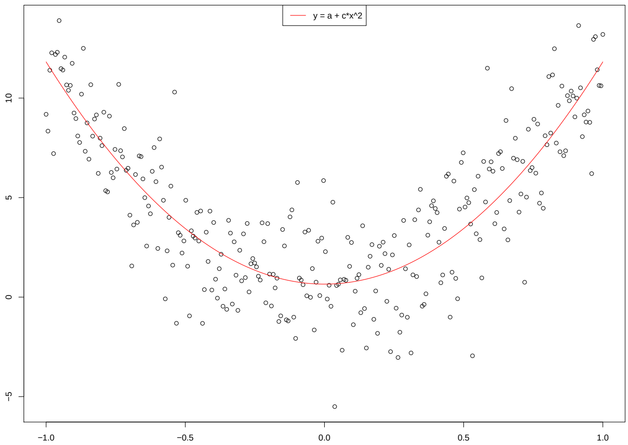
Beispiel: Polynomielle Regression

Betrachte Beispiel von Folie 6



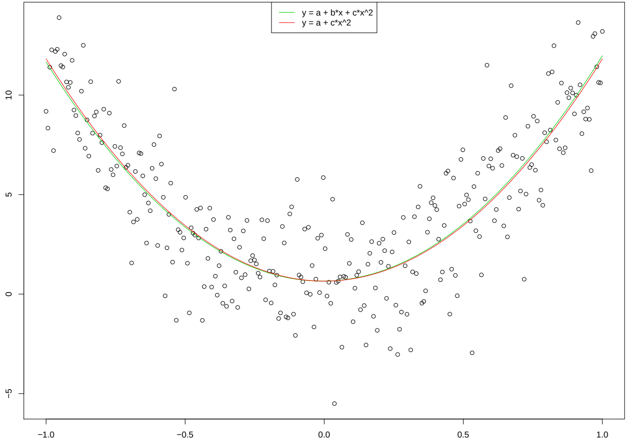
Beispiel: Polynomielle Regression

Betrachte Beispiel von Folie 6



Beispiel: Polynomielle Regression

Betrachte Beispiel von Folie 6



Beispiel: Polynomielle Regression, Modellanpassung

- `summary` Ausgabe für das Modell $Y_i = a + b \cdot X_i + c \cdot X_i^2 + \epsilon_i$:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.6527     0.1848   3.532 0.000478 ***
b              0.1586     0.2127   0.746 0.456398
c             11.1651     0.4105  27.200  < 2e-16 ***
```

$\Rightarrow$ Parameter b geht nicht signifikant in das Modell ein

- `summary` Ausgabe für das Modell $Y_i = a + c \cdot X_i^2 + \epsilon_i$:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.6527     0.1847   3.535 0.000473 ***
c             11.1651     0.4102  27.220  < 2e-16 ***
```

$\Rightarrow$ Modell mit weniger Parametern ist hier vorzuziehen

Modellanpassung - Der Befehl `step`

Frage: Welches Modell beschreibt die Situation am besten?

Bsp.: Auf der vorherigen Folie geht b nicht signifikant in in das Modell ein. Das Modell ohne den linearen Term beschreibt die Situation besser. (Die p-*Werte* sind kleiner.)

- Man nehme das signifikanteste Modell.
- Hat man viele Einflussgrößen, so ist es aufwendig jedes einzelne Modell zu überprüfen.
- In `R` erledigt dies die Funktion `step`
- `step` entfernt oder fügt zu einem bestehenden Modell Einflussfaktoren hinzu und bewertet die Güte des neuen Modells.
- Als Güte Maß dient der **AIC**-Wert (Akaike's Information Criterion). Je niedriger der Wert, desto besser ist die Modellanpassung.

Beispiel - step

- `step(lm(y ~ b + c))`

```
Start:  AIC=457.75
y ~ b + c

       Df Sum of Sq    RSS    AIC
- b     1       2.5 1354.9 456.31
<none>              1352.4 457.75
- c     1    3368.8 4721.2 830.81

Step:  AIC=456.31
y ~ c

       Df Sum of Sq    RSS    AIC
<none>              1354.9 456.31
- c     1    3368.8 4723.7 828.97

Call:
lm(formula = y ~ c)

Coefficients:
(Intercept)            c
     0.6527      11.1651
```

- Alternativ: `step(model, .~.+groesse1+groesse2)`

`add1` und `update`

- `add1` überprüft wie sich die Güte eines Modells bei Hinzunahme von Einflussfaktoren verändert.
- `update` fügt Grössen zu einem Modell hinzu

Beispiel

- `add1(model, .~.+groesse1+groesse2)`
- `update(model, .~.+groesses2)`

Die Aufgaben 9.7. bis 9.8. des Aufgabenblattes
können jetzt bearbeitet werden.

Gliederung

Binäre Regression

- Lineare Modelle eignen sich gut, wenn die Zielvariable stetig ist (und zumindestens approximativ normalverteilt)
- In vielen Anwendungen ist Zielvariable nicht stetig, sondern binär
- Bsp: Antwort auf eine Ja/Nein Frage
- **Ziel einer binären Regression:** Modellierung und Schätzung des Effekts der Variablen $X_{i,1}, \ldots, X_{i,p}$ auf die (bedingte) Wahrscheinlichkeit

$$\pi_i \; = \; \mathbb{P}(Y_i = 1 \mid X_{i,1}, \ldots, X_{i,p})$$

Modellierungsprobleme

Der intuitive Ansatz

$$\pi_i \;=\; \theta_0 + \theta_1 \cdot X_{i,1} \;+\; \ldots \;+\; \theta_p \cdot X_{i,p} \;+\; \epsilon_i, \quad i = 1, \ldots, n$$

ist ungeeignet:

- Y_i ist binär, was für die Regressorwerte der rechten Seite eine zu starke Einschränkung darstellt

Modellierung

- **Lösungsansatz:** Verknüpfe den *linearen Prädiktor* $\eta_i = \theta_0 + \theta_1 \cdot X_{i,1} + \ldots + \theta_p \cdot X_{i,p}$ mit einer monoton wachsenden Funktion $h : \mathbb{R} \to [0, 1]$, also

$$\pi_i \;=\; h(\theta_0 + \theta_1 \cdot X_{i,1} \;+\; \ldots \;+\; \theta_p \cdot X_{i,p})$$

  und untersuche das Modell

$$h^{-1}(\pi_i) \;=\; \theta_0 + \theta_1 \cdot X_{i,1} \;+\; \ldots \;+\; \theta_p \cdot X_{i,p}$$

- Als *Responsefunktion* h wählt man häufig eine Verteilungsfunktion
- Im Logit-Modell: $h(t) = \frac{\exp(t)}{1+\exp(t)}$ (logistische Verteilungsfunktion)
- Umkehrfunktion $h^{-1}(t) = \log(\frac{t}{1-t})$ heißt auch *Logit-Funktion*

Modelle für die Chancen

- Für die *Chancen (odds)*

$$\frac{\pi_i}{1-\pi_i} = \frac{\mathbb{P}(Y_i = 1 \mid X_{i,1}, \ldots, X_{i,p})}{\mathbb{P}(Y_i = 0 \mid X_{i,1}, \ldots, X_{i,p})}$$

erhält man (im Logit Modell!) das multiplikative Modell

$$\frac{\mathbb{P}(Y_i = 1 \mid X_{i,1}, \ldots, X_{i,p})}{\mathbb{P}(Y_i = 0 \mid X_{i,1}, \ldots, X_{i,p})} = \exp(\theta_0) \cdot \exp(\theta_1 X_{i,1}) \cdot \ldots \cdot \exp(\theta_p X_{i,p})$$

$\Rightarrow$ Für die logarithmierten Chancen *(log-odds)* erhält man ein lineares Modell

Interpretation

Bei Erhöhung der Kovariablen $X_{i,k}$ um 1 ändert sich das Verhältnis der Chancen um $\exp(\theta_k)$

$$\Rightarrow \exp(\theta_k) \cdot \Big(\exp(\theta_0) \cdot \exp(\theta_1 X_{i,1}) \cdot \ldots \cdot \exp(\theta_p X_{i,p}) \Big)$$

Also:

- $\theta_k > 0$: Chance wird größer
- $\theta_k < 0$: Chance wird kleiner
- $\theta_k = 0$: Chance bleibt gleich

Beispiel: Tod durch Herzversagen

- Zielvariable: Patient stirbt an Herzversagen (0/1 codiert)
- Kovariablen: Alter, Geschlecht, Cholesterinspiegel, ...
- Ist $\theta_1 = 0.08$: Chance an Herzversagen zu sterben erhöht sich bei einem 10 Jahre älterem Patienten um den Faktor $\exp(10 \cdot 0.08) \approx 2.2$
- Vorsicht: Bei unterschiedlichen Wahlen von h können die geschätzten Werte für θ stark abweichen, die Verhältnisse θ_2/θ_1 usw. bleiben jedoch nahezu konstant

$\Rightarrow$ Es kann festgestellt werden, welche Variable den größten Einfluss hat

Schätzung der Parameter

- θ wird *nicht* durch die Methode der kleinsten Quadrate geschätzt
- Mit einem Maximum-Likelihood-Ansatz lässt sich (numerisch) ein guter Schätzer für θ bestimmen. $(Y_1, \ldots, Y_n)$ ist bedingt unter den Kovariablen $\bigotimes_{i=1}^{n} B(1, \pi_i)$-verteilt.
- In `R` erhält man diese mit
  `glm(Y ~ x1 + ... + xp, family = binomial(link="logit"))`