Vorlesung Statistik II

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sommersemester 2026

Prognose

Häufig ist man daran interessiert, zu einem vorgegebenen Wert der Prädiktorvariablen x_0 die Realisierung der Kriteriumsvariablen Y_0 zu schätzen

Mit dem Regressionsmodell

$$Y_0 = \alpha + \beta \cdot x_0 + \varepsilon_0$$

ist wegen $E(\varepsilon_0) = 0$ die Zufallsvariable

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta} \cdot x_0$$

ein vernünftiger Punktschätzer für Y_0

Prognose- und Konfidenzintervall

Über die Eigenschaften von $\hat{\alpha}$ und $\hat{\beta}$ lässt sich die Varianz des Prognosefehlers $Y_0 - \hat{Y}_0$ bestimmen und daraus ein Konfidenzintervall für Y_0 (Prognoseintervall) ableiten

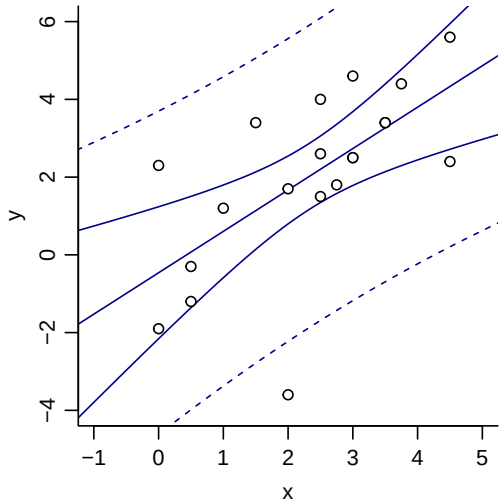
$$\hat{Y}_0 \pm \hat{\sigma} \cdot t_{1-\alpha/2}(n-2) \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Will man nur ein Konfidenzintervall für die "wahre" Regressionsgerade $\alpha + \beta x$, so fällt im Modell der additive Fehlerterm ε_0 weg und – daraus resultierend – die 1 unter der Wurzel

$$\hat{\alpha} + \hat{\beta} x \pm \hat{\sigma} \cdot t_{1-\alpha/2}(n-2) \cdot \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Prognose- und Konfidenzintervall

Prognose- und Konfidenzintervall für die "wahre" Regressionsgerade $\alpha + \beta\, x$ mit einer Irrtumswahrscheinlichkeit von 5%



Modelldiagnose

In jeder Anwendung der Regressionsanalyse sollten die grundlegenden Modellannahmen auf ihre (zumindest approximative) Gültigkeit hin überprüft werden

Diese Überprüfung wird häufig auch grafisch vorgenommen, wozu verschiedene Diagramme verwendet werden

- Streudiagramm
- Residualplot, d. h. die Darstellung der Residuen in Abhängigkeit von den Prädiktorwerten (bzw. den vorhergesagten Werten der Kriteriumsvariablen)
- Normal-Quantil-Plots der Residuen, wobei anstatt der $\hat{\varepsilon}_i$ häufig *studentisierte Residuen* verwendet werden

$$\frac{\hat{\varepsilon}_i}{\sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$

Leverage

Leverage bezeichnet den Anteil des Abweichungsquadrats des Prädiktorwerts x_i vom Mittelwert $\bar{x}$ an der Gesamtsumme der entsprechenden Abweichungsquadrate

$$l_i = \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

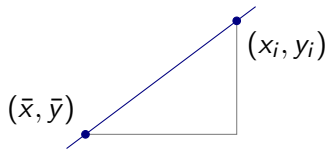
Es ist $\sum_{i=1}^{n} l_i = 1$

Damit gilt für den geschätzten Regressionskoeffizienten $\hat{\beta}$

$$\hat{\beta} = \sum_{i=1}^{n} l_i \cdot \frac{y_i - \bar{y}}{x_i - \bar{x}}$$

Leverage

Der Quotient $(y_i - \bar{y})/(x_i - \bar{x})$ gibt dabei die Steigung der durch die Punkte (x_i, y_i) und $(\bar{x}, \bar{y})$ festgelegten Gerade an



Die durch den Regressionskoeffizienten $\hat{\beta}$ bestimmte Steigung der Regressionsgeraden ergibt sich daher als gewichteter Mittelwert der Steigungen der Geraden, die durch die einzelnen Datenpunkte zusammen mit dem Punkt $(\bar{x}, \bar{y})$ festgelegt sind

Der Gewichtungsfaktor jedes Datenpunktes ist durch den Wert der Leverage l_i gegeben

Cook-Distanz

Die *Cook-Distanz* eines Datenpunktes kennzeichnet seinen Einfluss auf die Regression

Zu deren Bestimmung für den Datenpunkt (x_i, y_i) führt man die Regressionsanalyse einmal mit und einmal ohne diesen Datenpunkt durch

Die Cook-Distanz von (x_i, y_i) wird dann durch das geeignet standardisierte Abstandsquadrat zwischen den beiden durch die Regressionsanalysen mit bzw. ohne den Datenpunkt vorhergesagten Werten berechnet

Datenpunkte mit Werten größer als 0.5 gelten als auffällig einflussreich

Beispiel

Wir wollen wiederum den Zusammenhang zwischen der Prognose des Sachverständigenrates und dem tatsächlich beobachteten Wirtschaftswachstum untersuchen

Zunächst bestimmen wir den Stichprobenumfang n, um mit ca. 95% Power bei $\alpha = 0.05$ die Nullhypothese bzgl. der Steigung der Regressionsgeraden $\beta = 0$ abzulehnen, wenn tatsächlich $\beta = 1$ gilt

```
n <- 16                                  # Anzahl der Jahre
x <- rep(-3:4, each = n/8)
pval <- replicate(2000, {
  y <- 0 + 1*x + rnorm(n, sd = 2)        # y = a + b*x + e, b = 1
  m <- lm(y ~ x)
  coef(summary(m))["x", "Pr(>|t|)"]      # H0: b = 0
})
mean(pval < 0.05)
```

Beispiel

Anschließend betrachten wir die Prognose des Sachverständigenrates (Merkmal X) und das tatsächlich beobachtete Wirtschaftswachstum (Merkmal Y) in den Jahren 1975-1994

Jahr	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
x_i	2.0	4.5	4.5	3.5	3.75	2.75	0.5	0.5	1.0	2.5
y_i	-3.6	5.6	2.4	3.4	4.4	1.8	-0.3	-1.2	1.2	2.6

Jahr	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994
x_i	3.0	3.0	2.0	1.5	2.5	3.0	3.5	2.5	0.0	0.0
y_i	2.5	2.5	1.7	3.4	4.0	4.6	3.4	1.5	-1.9	2.3

Parameterschätzung

Die Schätzung des Regressionskoeffizienten β liefert

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{38.3525}{36.0125} = 1.0650$$

Mit $\bar{x} = 2.325$ und $\bar{y} = 2.015$ erhält man dann für Parameter α folgende Schätzung

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} = 2.015 - 1.0650 \cdot 2.325 = -0.4611$$

Insgesamt hat man also die Regressionsgleichung

$$\hat{y}_i = -0.4611 + 1.0650 \cdot x_i \text{ für alle } i = 1, \ldots, n$$

R-Formeln

Die Funktion `lm()` erlaubt die Schätzung linearer Modelle, wie sie durch lineare Regression, aber etwa auch durch die noch zu besprechende Varianzanalyse, gegeben sind

Zentral für die Anwendung von `lm()` ist die Spezifizierung des unterliegenden linearen Modells

- Basierend auf dem Formel-Operator `~` wird ein Ausdruck der Form

  ```
  y ~ Modell
  ```

  als Argument spezifiziert, bei dem die Variable `y` durch einen linearen Prädiktor vorhergesagt wird, der hier symbolisch durch das angegebene `Modell` gebildet wird (siehe `?formula`)

Anpassen eines linearen Modells

Im vorliegenden Fall wird nur die Variable x als Prädiktor verwendet, sodass das lineare Modell in diesem Fall lautet

```
y ~ 1 + x  # Langform
y ~ x      # Kurzform
```

Man hat also folgende Eingabe

```
dat <- data.frame(
  x = c(2.0, 4.5, 4.5, 3.5, 3.75, 2.75, 0.5, 0.5, 1.0, 2.5,
        3.0, 3.0, 2.0, 1.5, 2.5, 3.0, 3.5, 2.5, 0.0, 0.0),
  y = c(-3.6, 5.6, 2.4, 3.4, 4.4, 1.8, -0.3, -1.2, 1.2, 2.6,
        2.5, 2.5, 1.7, 3.4, 4.0, 4.6, 3.4, 1.5, -1.9, 2.3)
)
lm(y ~ x, dat)
```

Anpassen eines linearen Modells

Als Ausgabe erhält man dann

```
Call:
lm(formula = y ~ x, data = dat)

Coefficients:
(Intercept)            x
    -0.4611       1.0650
```

Mit der Eingabe

```
m <- lm(y ~ x, dat)
```

werden sämtliche Ausgaben der Funktion `lm()` in dem Objekt `m` gespeichert und stehen für weitere Abfragen zur Verfügung

Extraktor-Funktionen

Ausgabe der Koeffizienten

```
coef(m)
```

Ausgabe der vorhergesagten Werte

```
predict(m)
```

Einzeichnen der geschätzten Regressionsgeraden in das Streudiagramm

```
plot(y ~ x, dat); abline(m)
```

Ausgabe der Residuen

```
resid(m)
```

Ausgabe von Konfidenzintervallen

```
confint(m, level = 0.95)
```

Fehlervarianz

Als Schätzung für die Varianz σ^2 hat man dann

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \\
&= \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}\, x_i)^2 \\
&= \frac{1}{18} \left[(-3.6 - (-0.4611) - 1.0650 \cdot 2.0)^2 \right. \\
&\quad \left. + \ldots + (2.3 - (-0.4611) - 1.0650 \cdot 0.0)^2 \right] \\
&= 3.2634
\end{aligned}$$

Standardfehler

Damit gilt dann

$$\hat{\sigma} = 1.8065$$

Für die Standardfehler der Schätzfunktionen $\hat{\alpha}$ und $\hat{\beta}$ erhält man

$$\hat{\sigma}_{\hat{\alpha}} = \hat{\sigma} \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}} = 1.8065 \cdot \sqrt{\frac{144.125}{20 \cdot 36.0125}} = 0.8081$$

und

$$\hat{\sigma}_{\hat{\beta}} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} = \frac{1.8065}{\sqrt{36.0125}} = 0.3010$$

Konfidenzintervalle

Mit $t_{0.975}(18) = 2.1010$ erhält man folgende 0.95-Konfidenzintervalle

Für

$$[\hat{\alpha} - \hat{\sigma}_{\hat{\alpha}} \cdot t_{0.975}(18), \hat{\alpha} + \hat{\sigma}_{\hat{\alpha}} \cdot t_{0.975}(18)]$$

berechnet man

$$[-0.4611 - 0.8081 \cdot 2.1010, -0.4611 + 0.8081 \cdot 2.1010] = [-2.16, 1.24]$$

und für

$$\left[\hat{\beta} - \hat{\sigma}_{\hat{\beta}} \cdot t_{0.975}(18), \hat{\beta} + \hat{\sigma}_{\hat{\beta}} \cdot t_{0.975}(18)\right]$$

ergibt sich

$$[1.0650 - 0.3010 \cdot 2.1010, 1.0650 + 0.3010 \cdot 2.1010] = [0.43, 1.70]$$

Hypothesentests

Unter der Nullhypothese $H_0\colon\ \alpha = 0$ berechnet man für die Teststatistik

$$T_0 = \frac{\hat{\alpha} - 0}{\hat{\sigma}_{\hat{\alpha}}} = \frac{-0.4611}{0.8081} = -0.57$$

und erhält für $\alpha = 0.05$ das kritische Quantil $t_{0.975}(18) = 2.10$

Unter der Nullhypothese $H_0\colon\ \beta = 0$ berechnet man die Teststatistik

$$T_0 = \frac{\hat{\beta} - 0}{\hat{\sigma}_{\hat{\beta}}} = \frac{1.0650}{0.3010} = 3.54$$

die für $\alpha = 0.05$ wiederum mit $t_{0.975}(18) = 2.10$ zu vergleichen ist

summary-Methode

In Fortführung der oben bereits gemachten Eingaben wird durch

```
summary(m)
```

eine Übersicht über die Ergebnisse der linearen Regression ausgegeben, mit folgenden Bestandteilen

- Spezifikation des Modells
- Kennwerte der Residuen
- Parameterschätzung und statistische Tests der Parameter α (`Intercept`) und β (bezeichnet durch die zugehörige Variable `x`) gegen den Wert 0
- Schätzung der Fehler-Standardabweichung σ
- Determinationskoeffizient
- F-Statistik des Overall-Tests

summary-Methode

```
Call:
lm(formula = y ~ x, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-5.269 -0.676  0.082  0.968  2.761

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.461      0.808   -0.57   0.5753
x              1.065      0.301    3.54   0.0024 **

Residual standard error: 1.81 on 18 degrees of freedom
Multiple R-squared:  0.41,    Adjusted R-squared:  0.377
F-statistic: 12.5 on 1 and 18 DF,  p-value: 0.00235
```

Varianzzerlegung

Eine detailliertere Auflistung der Varianzzerlegung, die der Berechnung der F-Statistik zugrunde liegt, erhält man durch die `anova`-Funktion

In Fortführung des Beispiels erhält man mit der Eingabe

```
anova(m)
```

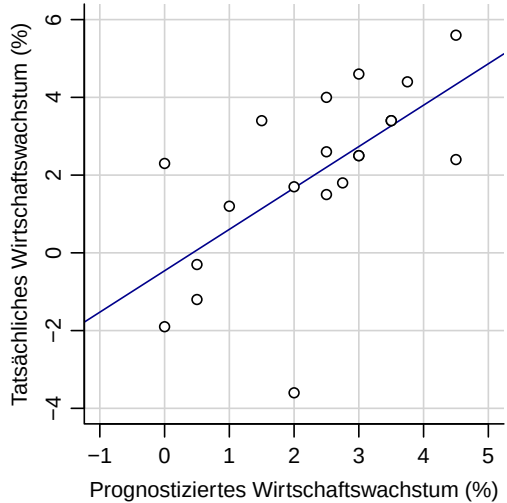
die Ausgabe

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
x          1  40.84   40.84   12.52 0.00235 **
Residuals 18  58.74    3.26
```

Modelldiagnose

Daten und Modellvorhersage



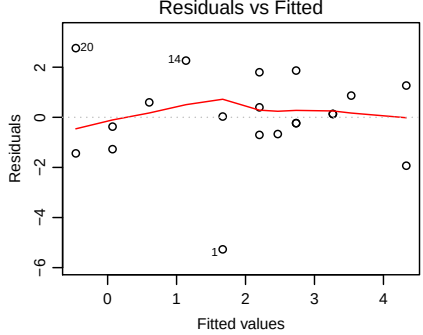
Modelldiagnose

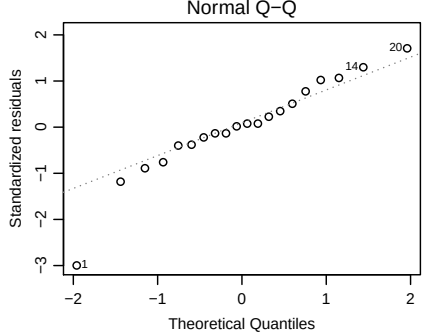
Die plot-Methode für lm-Objekte stellt auch verschiedene Grafiken zur Verfügung, die eine Diagnose der Abweichungen der Daten von der Modellvorhersage sowie von Verletzungen der grundlegenden Annahmen ermöglichen

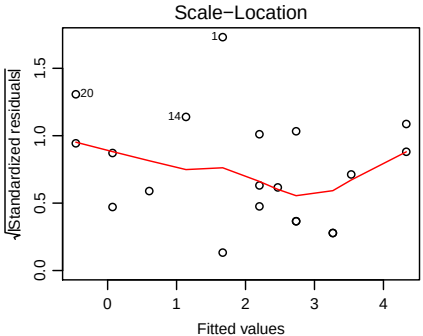
Um die angebotenen vier Grafiken gleichzeitig auf einer Seite darstellen zu können, wird durch die Funktion `par()` und das Argument `mfrow = c(2, 2)` deren Anordnung in einem 2×2-Schema festgelegt

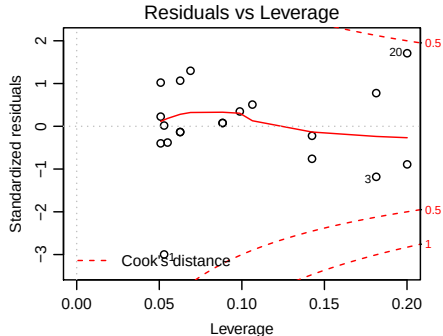
Die Eingabe lautet dann

```
par(mfrow = c(2, 2))
plot(m)
```


Residuals vs Fitted
Residuals
Fitted values
20
14
1


Normal Q−Q
Standardized residuals
Theoretical Quantiles
20
14
1


Scale−Location
√|Standardized residuals|
Fitted values
1
20
14


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5
0.5
1
20
3
1

Multiple lineare Regression

Der Anwendungsbereich der einfachen linearen Regression lässt sich dadurch erweitern, dass neben der Kriteriumsvariablen Y mehrere Prädiktorvariablen $X_1, \ldots, X_p$ betrachtet werden

Die empirischen Beobachtungen bestehen dann für alle Untersuchungseinheiten aus einem Tupel

$$(y_i, x_{i1}, \ldots, x_{ip}) \quad \text{mit} \quad i = 1, \ldots, n$$

Die Kriteriumsvariable Y ist hierbei stets eine metrische Variable, während die (deterministischen oder stochastischen) Prädiktorvariablen $X_1, \ldots, X_p$ metrische oder auch kategoriale Variablen sein können

Modell

In unmittelbarer Verallgemeinerung des Modells der einfachen linearen Regression formuliert man als stochastisches Grundmodell

$$Y_i = \beta_0 + \beta_1 \cdot X_{i1} + \ldots + \beta_p \cdot X_{ip} + \varepsilon_i$$
$$\text{mit } E(\varepsilon_i) = 0, \text{ für alle } i = 1, \ldots, n$$

Entsprechend ergibt sich aus der zusätzlichen Annahme, dass die Zufallsvariablen ε_i, $i = 1, \ldots, n$, stochastisch unabhängig und identisch normalverteilt sind gemäß $\varepsilon_i \sim N(0, \sigma^2)$, für gegebene Werte $x_{i1}, \ldots, x_{ip}$, $i = 1, \ldots, n$, die Folgerung

$$Y_i \sim N(\mu_i, \sigma^2) \quad \text{mit} \quad \mu_i = \beta_0 + \beta_1 \cdot x_{i1} + \ldots + \beta_p \cdot x_{ip}$$

Parameterschätzung

Auch im vorliegenden Fall lässt sich die Parameterschätzung nach dem (gewöhnlichen) Kleinst-Quadrat-Kriterium einsetzen, bei der Schätzfunktionen $\hat{\beta}_0, \ldots, \hat{\beta}_p$ so bestimmt werden, dass die Summe der quadratischen Abweichungen

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot X_{i1} - \ldots - \beta_p \cdot X_{ip})^2$$

minimiert wird

Als Schätzer für die Varianz σ^2 erhält man

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot X_{i1} - \ldots - \beta_p \cdot X_{ip})^2$$

Parameterschätzung

Auch die Verteilungen der Schätzfunktionen, und damit die daraus abgeleiteten Konfidenzintervalle und statistischen Hypothesentests, verallgemeinern sich entsprechend

Die Anzahl der Freiheitsgrade bestimmt sich dabei im allgemeinen Fall durch $df = n - p - 1$, wobei p die Anzahl der Prädiktoren bezeichnet

- Die Anzahl der Freiheitsgrade in der einfachen linearen Regression ($df = n - 2$) ergibt sich dann als Spezialfall $p = 1$

Im Folgenden betrachten wir lediglich den globalen Test (*Overall-F-Test*) zur Klärung der Frage, ob die Prädiktoren überhaupt zur Vorhersage der Kriteriumsvariablen beitragen

Overall-F-Test

Hypothesen

$$H_0: \ \beta_1 = \ldots = \beta_p = 0$$
$$H_1: \ \beta_j \neq 0 \ \text{ für mindestens ein } j \in \{1, \ldots, p\}$$

Teststatistik

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - p - 1}{p}$$

Verteilung der Teststatistik unter H_0

$$F \sim F(p, n - p - 1)$$

Ablehnungsbereich

$$F > F_{1-\alpha}(p, n - p - 1)$$

Multikollinearität

Multikollinearität bezeichnet Abhängigkeiten unter den Prädiktorvariablen, die die Interpretion der Ergebnisse beeinträchtigt

- Sie beschränkt den maximal erreichbaren erklärten Varianzanteil R^2
- Sie erschwert die Bewertung der Bedeutsamkeit eines Prädiktors
  - Durch die Wechselbeziehungen zu anderen Prädiktoren hängt der für einen Prädiktor resultierende Regressionskoeffizient bei Multikollinearität davon ab, welche anderen Prädiktoren im Regressionsmodell enthalten sind
- Sie erhöht die Varianz der Schätzfunktionen und beeinträchtigt dadurch die Schätzung der Regressionskoeffizienten
- Sie kann zu systematischen Verzerrungen der Prüfgrößen (t- bzw. F-Werte) führen

Multiple lineare Regression in R

Für die Durchführung einer multiplen linearen Regression ist das anzugebende Modell entsprechend zu erweitern

Werden die Prädiktorvariablen X_1 und X_2 in R durch `x1` und `x2` im Dataframe `dat` repräsentiert, dann wird das Modell

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \varepsilon$$

in folgenden äquivalenten Funktionsaufrufen spezifiziert

```
lm(y ~ 1 + x1 + x2, dat)
lm(y ~     x1 + x2, dat)
```

Adjustierung des Determinationskoeffizienten

Entscheidend für die Bewertung einer Regressionsanalyse ist ihre Eigenschaft die Ergebnisse einer unabhängigen Stichprobe vorherzusagen

Durch geeignete Verfahren der Validierung ist also die Generalisierbarkeit der Regressionsgleichung zu bestimmen

Es wurde eine Vielzahl von Formeln vorgeschlagen, die eine Abschätzung der Verringerung der Vorhersagegüte beim Übergang von der betrachteten Stichprobe auf die Population bzw. eine unabhängige Stichprobe ermöglichen sollen

Die Adäquatheit und Nützlichkeit dieser Formeln wird in der Literatur durchaus kontrovers diskutiert

Adjustierung des Determinationskoeffizienten

Eine der hierzu vorgeschlagenen Methoden ist die Adjustierung des Determinationskoeffizienten

- Dieser Wert wird in R durch die Funktion `summary.lm()` mit ausgegeben (siehe oben unter `Adjusted R-squared`)

Die von Wherry vorgeschlagene Formel lautet

$$\hat{\rho}^2 = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

wobei $\hat{\rho}$ eine Schätzfunktion für die multiple Korrelation ρ in der Population ist und p die Anzahl der Prädiktoren

Beachte: Im Gegensatz zu R^2 steigt der Wert von $\hat{\rho}^2$ nicht notwendig mit der Hinzunahme weiterer Prädiktoren an, sondern kann dabei sogar abnehmen

Polynome von Prädiktoren

Eine weitere Verallgemeinerung der einfachen linearen Regression bilden Polynome von Prädiktoren

Das Modell bleibt linear in den Koeffizienten, auch wenn die Vorhersage nichtlinear ist

Beispiel

- Durch Hinzufügen eines quadratischen Terms erhält man eine Vorhersage des Zusammenhangs zwischen Prädiktor und Kriterium durch das stochastische Grundmodell

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2 + \varepsilon_i$$
$$\text{mit } E(\varepsilon_i) = 0, \text{ für alle } i = 1, \ldots, n$$

In diesem Fall (und allen analogen Situationen) kann hinsichtlich Parameterschätzung, Bestimmung von Konfidenzintervallen und statistischer Hypothesentests prinzipiell dieselbe Vorgehensweise wie bei der einfachen linearen Regression gewählt werden

Binäre Regression

In den bisher behandelten Regressionsmodellen wurde stets vorausgesetzt, dass die Kriteriumsvariable eine metrische Variable ist

Häufig tritt aber auch der Fall auf, dass die vorherzusagende Variable binär ist

Beispiele

- Präferenzen im Rahmen eines 2AFC-Paradigmas (z. B. Entscheidung für empfindungsmäßig intensiveren Reiz in der Psychophysik)
- Klassifikation von Personen als krank oder gesund

Im Folgenden betrachten wir den einfachsten Fall einer Prädiktorvariablen X_i und einer binären Kriteriumsvariablen $Y_i \in \{0, 1\}$. Für die Bernoulli-verteilte Zufallsvariable Y_i sei

$$\pi_i = P(Y_i = 1) \quad \text{und} \quad 1 - \pi_i = P(Y_i = 0)$$

Linearer Prädiktor

Die Beschreibung der Abhängigkeit der Wahrscheinlichkeit π_i von dem Wert x_i der Prädiktorvariablen durch das Regressionsmodell

$$Y_i = \alpha + \beta \cdot x_i + \varepsilon_i$$

mit $E(\varepsilon_i) = 0$ impliziert

$$E(Y_i) = \pi_i = \alpha + \beta \cdot x_i$$

Problematisch an dieser Regressionsgleichung ist, dass π_i als Wahrscheinlichkeit die Restriktion $\pi_i \in [0, 1]$ erfüllen muss

Will man eine (von den Parametern abhängige) Einschränkung der Wertebereiche der Prädiktorvariablen vermeiden, ist das lineare Modell geeignet zu transformieren

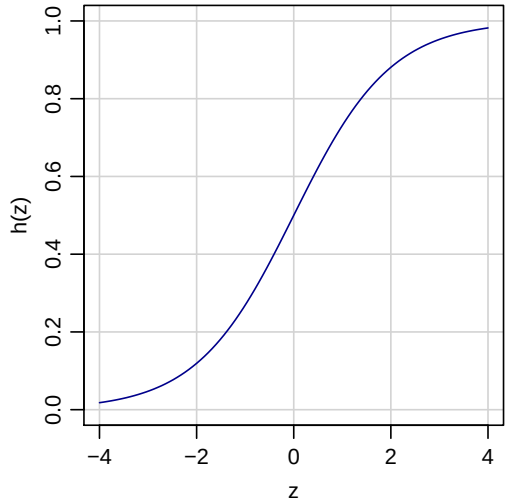
Logistische Funktion

Eine geeignete Transformation bildet die *logistische Funktion*

$$h(z) = \frac{\exp(z)}{\exp(z)+1} = \frac{1}{1+\exp(-z)}$$

Damit erhält man das *logistische Regressionsmodell*

$$\pi_i = h(\alpha + \beta \cdot x_i)$$
$$= \frac{\exp(\alpha + \beta \cdot x_i)}{\exp(\alpha + \beta \cdot x_i)+1} = \frac{1}{1+\exp[-(\alpha + \beta \cdot x_i)]}$$



Der Erwartungswert von Y_i hängt nun lediglich mittelbar, über eine Transformation durch die Funktion h, von der Prädiktorvariablen ab

Linkfunktion

Eine einfache Umformung zeigt, dass das logistische Regressionsmodell auch darstellbar ist in der Form

$$\log \frac{\pi_i}{1-\pi_i} = \alpha + \beta \cdot x_i$$

Dabei werden die Quotienten

$$\frac{\pi_i}{1-\pi_i} \quad \text{bzw.} \quad \log \frac{\pi_i}{1-\pi_i}$$

als "Odds" (Wettchance) bzw. "Logits" (logarithmierte Wettchance) bezeichnet

Allgemein verknüpft die Linkfunktion g den Erwartungswert von Y mit dem linearen Prädiktor

$$g(E(Y)) = \alpha + \beta \cdot x$$

Beispiel: Psychometrische Funktion

Im Rahmen der Konstanzmethode werden Schätzungen der Wahrscheinlichkeit $p(x_i, s)$ erhoben, einen Vergleichsreiz x_i als intensiver als eine Standardreiz s zu beurteilen

Zur Beschreibung des Zusammenhangs wurde als *psychometrische Funktion* die durch zwei Parameter c und d charakterisierte *logistische Funktion* vorgeschlagen

$$p(x_i, s) = \frac{1}{1 + \exp\left(-\frac{x_i - c}{d}\right)}$$

Nach Logit-Transformation hat man dann

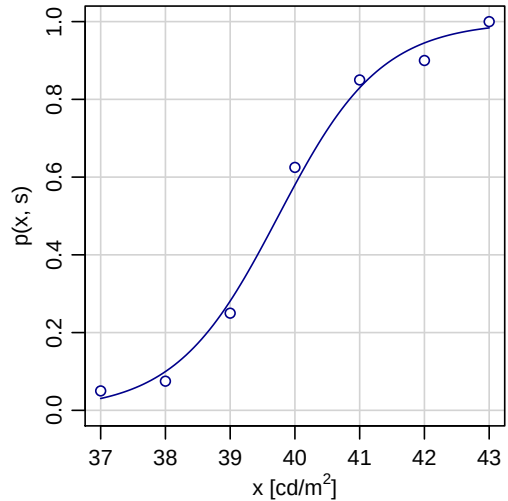
$$\log \frac{p(x_i, s)}{1 - p(x_i, s)} = -\frac{c}{d} + \frac{1}{d} x_i$$

Beispiel: Psychometrische Funktion

Ergebnisse zur Diskrimination der Helligkeit achromatischer Farben

- Standardreiz $s = 40$ cd/m^2
- Anzahl der Darbietungen $n = 40$ pro Vergleichsreiz
- Anzahl der Fälle y, in denen x_i als intensiver als s beurteilt wird

	Vergleichsreize [cd/m^2]						
x	37	38	39	40	41	42	43
y	2	3	10	25	34	36	40
$\hat{p}(x, s)$	.05	.075	.25	.625	.85	.90	1.00



Beispiel: Psychometrische Funktion

Eine Schätzung der Parameter α und β im Regressionsmodell

$$\log \frac{p(x_i, s)}{1 - p(x_i, s)} = \alpha + \beta\, x_i$$

liefert dann die Schätzungen

$$\hat{d} = \frac{1}{\hat{\beta}} \quad \text{und} \quad \hat{c} = -\frac{\hat{\alpha}}{\hat{\beta}}$$

für die Parameter der logistischen psychometrischen Funktion
(c Punkt der subjektiven Gleichheit, $1/d$ Diskriminationsfähigkeit)

Generalisierte lineare Modelle

Das logistische Regressionsmodell ist ein Spezialfall sogenannter *generalisierter linearer Modelle*

- Diese Modelle enthalten eine Transformation zwischen dem linearen Term und dem Erwartungswert der Kriteriumsvariablen (Linkfunktion)

Die Funktion `glm()`

- Mit `glm()` wird die entsprechende Verallgemeinerung der Funktion `lm()` in R bereitgestellt
- Die Verteilung der Zielvariable ist als Argument (bezeichnet durch `family`) der Funktion `glm()` zu spezifizieren
- Im vorliegenden Fall einer binären Kriteriumsvariablen ist `family = binomial(link = logit)` anzugeben

Beispiel: Psychometrische Funktion

Mit der Eingabe

```
dat <- data.frame(
  x = c(37, 38, 39, 40, 41, 42, 43),
  y = c(2, 3, 10, 25, 34, 36, 40)
)
glm(cbind(y, 40 - y) ~ x, family = binomial, dat)
```

erhält man die Ausgabe

```
Call:  glm(formula = cbind(y, 40 - y) ~ x, family = binomial, data = dat)

Coefficients:
(Intercept)            x
    -50.150        1.262
Degrees of Freedom: 6 Total (i.e. Null);  5 Residual
Null Deviance:      191.8
Residual Deviance: 3.967        AIC: 28.06
```

Beispiel: Psychometrische Funktion

Ergebnisse zur Diskrimination der Helligkeit achromatischer Farben

- Damit ergeben sich die Schätzungen

$$\hat{d} = \frac{1}{1.262} = 0.793$$

und

$$\hat{c} = -\frac{-50.150}{1.262} = 39.74$$

für die Parameter der logistischen psychometrischen Funktion