Einführung in die induktive Statistik

Friedrich Leisch

Institut für Statistik
Ludwig-Maximilians-Universität München

SS 2009, Lineare Regression

Fragen

- Welche Unsicherheitsfaktoren beeinflussen die Schätzung einer Regressionsgeraden?
- Wenn wir die Körpergröße aus der Unterarmlänge prognostizieren wollen, wie könnten wir vorgehen? Was haben wir bei der Konstruktion von Konfidenzintervallen zu beachten? Wie könnten Konfidenzbereiche aussehen?
- Welche statischen Tests könnten bei der linearen Regression von Interesse sein?

Übersicht

- Wiederholung aus Deskriptive Statistik: Lineare Einfachregression
- Das stochastische Modell der einfachen Regression
- Tests für Parameter
- Korrelation der Parameter
- Multiple Regression

Lineare Einfachregression

Lineare Einfachregression

Modell:

$$y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, \ldots, n$$

Kleinste-Quadrate-Schätzer:

$$\text{SQR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \Big(y_i - (\hat{\alpha} + \hat{\beta} x_i)\Big)^2 \quad \rightarrow \quad \min$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \qquad \hat{\beta} = \frac{s_{XY}}{s_X^2}$$

Residuen

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta} x_i)$$

Es gilt:

$$\begin{aligned}
\sum_{i=1}^{n} \hat{\epsilon}_i &= \sum_{i=1}^{n} \Big(y_i - (\hat{\alpha} + \hat{\beta} x_i)\Big) \\
&= \sum_{i=1}^{n} (y_i - (\bar{y} - \hat{\beta}\bar{x} + \hat{\beta} x_i) \\
&= \sum_{i=1}^{n} (y_i - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta} x_i) \\
&= \sum_{i=1}^{n} (y_i - \bar{y}) + \hat{\beta} \sum_{i=1}^{n} (\bar{x} - x_i) \\
&= 0 + \hat{\beta} 0 = 0
\end{aligned}$$

Streuungszerlegung

Frage: Wie gut paßt die Regressionsgerade zu den Daten?

Maß für die Variabilität der abhängigen Variablen Y ist die Varianz:

$$\tilde{s}_Y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Bei Regression betrachtet man üblicherweise die Quadratsumme

$$\text{SQT} = n\tilde{s}_Y^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

(SQT = „**S**um of s**Q**uares **T**otal“)

Streuungszerlegung

$$\text{SQT} = \text{SQE} + \text{SQR}$$

mit

- **S**um of s**Q**uares **T**otal
$$\text{SQT} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
- **S**um of s**Q**uares **E**xplained
$$\text{SQE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$
- **S**um of s**Q**uares **R**esidual
$$\text{SQR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Erklärte Varianz

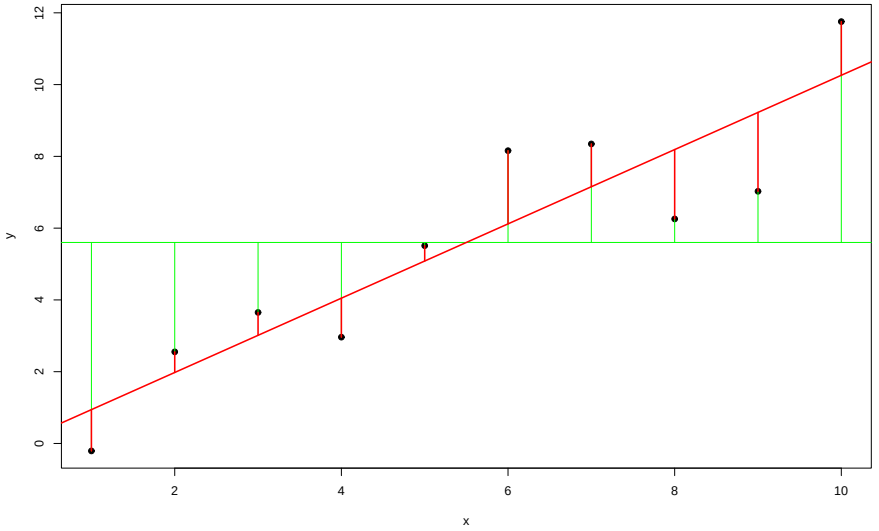
Bestimmtheitsmaß:

$$R^2 = \frac{\text{SQE}}{\text{SQT}} = 1 - \frac{\text{SQR}}{\text{SQT}} = r^2_{XY} \quad \in [0,1]$$

$R^2 \approx 0$: Varianz der Residuen identisch zur Varianz von Y, Regressionsgerade horizontal, X hat keinen linearen (!) Einfluß auf Y

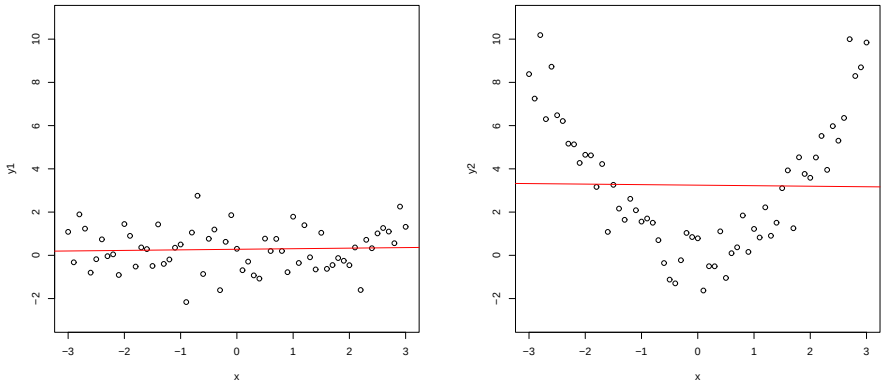
$R^2 \approx 1$: Varianz der Residuen fast 0, Daten liegen fast perfekt auf einer Geraden

Erklärte Varianz



Erklärte Varianz: $R^2 \approx 0$

Regressionsgerade horizontal:



Verbesserungsmöglichkeiten

Die deskriptive Anpassung einer Ausgleichgeraden an bivariate Daten kann auf verschiedene Arten verbessert werden: Wünschenswert wäre

- Test, ob die Varianzerklärung signifikant von Null verschieden ist,
- Tests, ob α und β von Null verschieden sind,
- mehr als eine erklärende Variable zu verwenden, und
- kategorische erklärende Variablen zu verwenden.

Stochastisches Regressionsmodell

Modell bleibt gleich:

$$y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, \ldots, n$$

Aber wir modellieren nun (zumindest) y_i und ϵ als Zufallsvariablen:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

Im einfachsten Fall wird X_i als deterministisch angesehen („geplante Experimente“), falls die beobachteten Paare (x_i, y_i) jedoch aus einer Stichprobe stammen, ist auch X_i eine Zufallsvariable.

Annahmen des Modells

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

- Die abhängige Variable Y ist metrisch skaliert.
- Die Regressionsfunktion ist linear.
- Die Fehler ϵ sind unabhängig von X.
- Fehler sind unabhängig identisch verteilt (Homoskedastizität) mit

$$\mathbb{E}\epsilon_i = 0, \qquad \mathrm{Var}(\epsilon_i) = \sigma^2$$

Eigenschaften von Y

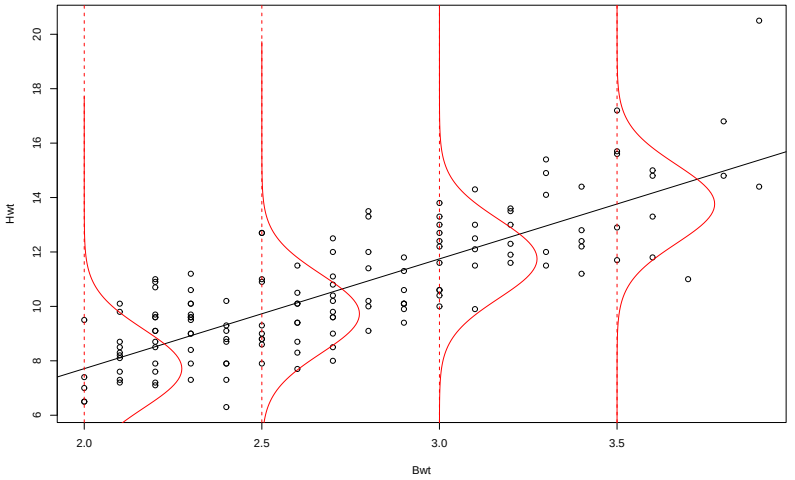
Aus den Modellannahmen folgt direkt:

$$\begin{aligned}
\mathbb{E}(Y_i|X_i = x_i) &= \mathbb{E}(\alpha + \beta X_i + \epsilon_i|X_i = x_i) \\
&= \alpha + \beta\mathbb{E}(X_i|X_i = x_i) + \mathbb{E}(\epsilon_i) \\
&= \alpha + \beta x_i \\
\mathrm{Var}(Y_i|X_i = x_i) &= \mathrm{Var}(\alpha + \beta X_i + \epsilon_i|X_i = x_i) \\
&= \beta^2\mathrm{Var}(X_i|X_i = x_i) + \mathrm{Var}(\epsilon_i) \\
&= \beta^2 0 + \sigma^2 = \sigma^2
\end{aligned}$$

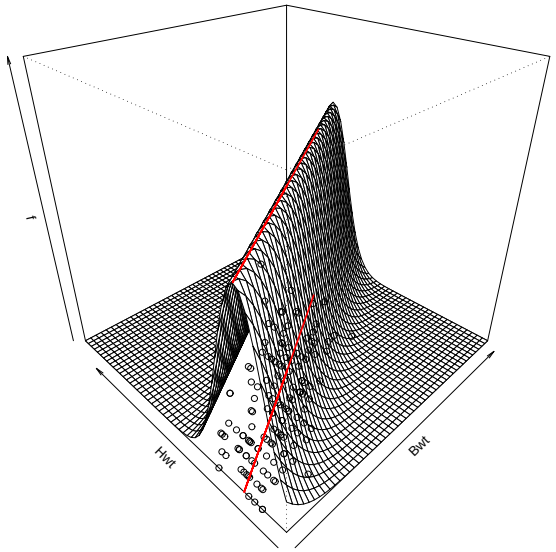
Falls die Fehler normalverteilt sind ($\epsilon_i \sim N(0, \sigma^2)$) gilt weiters:

$$Y_i \sim N(\alpha + \beta X_i, \sigma^2)$$

Bsp: Herzgewicht von Katzen



Bsp: Herzgewicht von Katzen



Schätzen der Parameter

Da bei der Normalverteilung Kleinstquadrat-Schätzung und Maximum-Likelihood-Schätzung identisch sind, ändern sich die Schätzer nicht:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \qquad \hat{\beta} = \frac{s_{XY}}{s_X^2} = \frac{r_{XY} s_X s_Y}{s_X^2} = r_{XY}\frac{s_Y}{s_X}$$

Als Schätzer für die unbekannte Fehlervarianz σ^2 verwenden wir die Varianz der Residuen:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2$$

(Nenner $n-2$ wegen 2 davor geschätzten Parametern α und β).

Eigenschaften der KQ-Schätzer

Verteilung der geschätzten Regressionskoeffizienten:

$$\hat{\alpha} \sim N(\alpha, \sigma^2_{\hat{\alpha}}) \quad \text{mit} \quad Var(\hat{\alpha}) = \sigma^2_{\hat{\alpha}} = \sigma^2 \frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta} \sim N(\beta, \sigma^2_{\hat{\beta}}) \quad \text{mit} \quad Var(\hat{\beta}) = \sigma^2_{\hat{\beta}} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Schätzer $\hat{\sigma}^2_{\hat{\alpha}}$ und $\hat{\sigma}^2_{\hat{\beta}}$ ergeben sich mit $\hat{\sigma}^2$ statt σ^2.

$\hat{\alpha}$, $\hat{\beta}$ und $\hat{\sigma}$ sind erwartungstreue Schätzer,
und konsistent falls $\sum_{i=1}^n (x_i - \bar{x})^2 \to \infty$ für $n \to \infty$.

Verteilung der standardisierten Schätzfunktionen:

$$\frac{\hat{\alpha} - \alpha}{\hat{\sigma}_{\hat{\alpha}}} \sim t(n-2) \qquad \frac{\hat{\beta} - \beta}{\hat{\sigma}_{\hat{\beta}}} \sim t(n-2)$$

Eigenschaften der KQ-Schätzer

- $(1-\gamma)$-Konfidenzintervalle für α und β:

  für α: $\left[\hat{\alpha} - \hat{\sigma}_{\hat{\alpha}} t_{1-\frac{\gamma}{2}}(n-2),\ \hat{\alpha} + \hat{\sigma}_{\hat{\alpha}} t_{1-\frac{\gamma}{2}}(n-2)\right]$

  für β: $\left[\hat{\beta} - \hat{\sigma}_{\hat{\beta}} t_{1-\frac{\gamma}{2}}(n-2),\ \hat{\beta} + \hat{\sigma}_{\hat{\beta}} t_{1-\frac{\gamma}{2}}(n-2)\right]$

- Testen von Hypothesen: Teststatistiken

$$T_{\alpha_0} = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}_{\hat{\alpha}}} \quad \text{und} \quad T_{\beta_0} = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_{\hat{\beta}}}$$

Prognose

Regressionsgerade:

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

Konfidenzintervall für $\hat{Y}_0$:

$$\left[\hat{Y}_0 \pm t_{1-\frac{\gamma}{2}}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}} \right]$$

Beobachtete Werte:

$$Y_0 = \hat{Y}_0 + \epsilon_0 = \hat{\alpha} + \hat{\beta} x_0 + \epsilon_0$$

Konfidenzintervall für Y_0:

$$\left[\hat{Y}_0 \pm t_{1-\frac{\gamma}{2}}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}} \right]$$

Erklärte Varianz: signifikant?

Zur Beantwortung der Frage, ob des Modell signifikant zur Erklärung der Daten beiträgt, kann man testen, ob

1. $\sigma^2 = \text{Var}(\epsilon)$ kleiner als $\sigma_y^2 = \text{Var}(y)$ ist.
2. R^2 von Null verschieden ist.
3. die Korrelation von X und Y von Null verschieden ist.
4. $\hat{\beta}$ von Null verschieden ist.

Im Fall der linearen Einfachregression sind alle 4 Tests de facto identisch, für mehr als eine erklärende Variable sind

- 1. und 2. identisch
- 3. und 4. verschieden (siehe später)

Erklärte Varianz: signifikant?

Die Teststatistik für Korrelation von X und Y ist:

$$T = \frac{r_{XY}}{\sqrt{1 - r_{XY}^2}}\sqrt{n-2} \quad \sim \quad t(n-2)$$

In der Regressionsanalyse ist es üblicher, das Quadrat dieser Statistik zu betrachten:

$$F = T^2 = \frac{r_{XY}^2}{1 - r_{XY}^2}(n-2) = \frac{R^2}{1-R^2}(n-2) \quad \sim \quad F(1, n-2)$$

Dieser F-Test läßt sich leichter für mehr als eine erklärende Variable verallgemeinern (Varianzanalyse, VO Lineare Modelle).

Ablehnung der Nullhypothese „kein Zusammenhang zwischen X und Y“ zum Signifikanzniveau γ für $F > F_{1-\gamma}(1, n-2)$.

Erklärte Varianz: signifikant?

Wegen

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

gilt weiters

$$F = (n-2)\frac{R^2}{1-R^2} = (n-2)\frac{\text{SQE/SQT}}{\text{SQR/SQT}} = \frac{\text{SQE}}{\frac{1}{n-2}\text{SQR}} = \frac{\text{SQE}}{\hat{\sigma}^2}$$

Die F-Statistik wird also groß, wenn

- SQE groß ist (steile Gerade)
- SQR klein ist (kleine Residuen)

Ohne BW: $F = T_{\beta_0}^2$ für $\beta_0 = 0$ in der linearen Einfachregression.

Bsp: Katzen

Welche Koeffizienten sind notwendig?

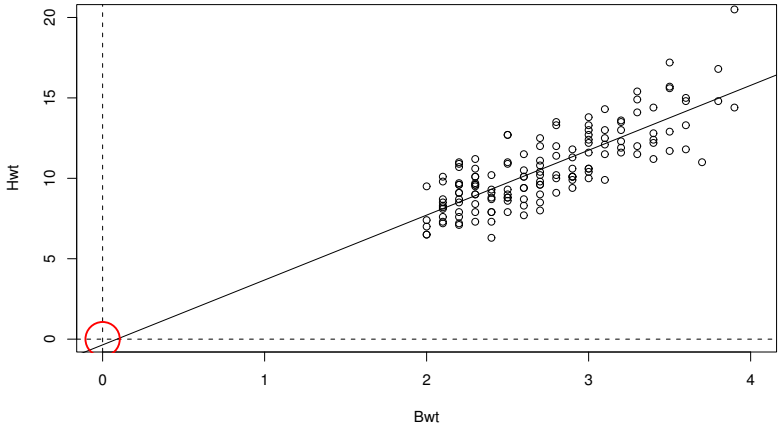
Für die Hypothesen $\alpha = 0$ und $\beta = 0$ kann man t-Tests formulieren:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−0.3567	0.6923	−0.52	0.6072
Bwt	4.0341	0.2503	16.12	0.0000

$$\hat{\sigma} = 1.452,\ R^2 = 0.64,\ F = 259.8,\ p < 10^{-15}$$

Interpretation: Pro kg Körpergewicht steigt das Herzgewicht im Schnitt um 4.03g, die Konstante ist nicht notwendig.

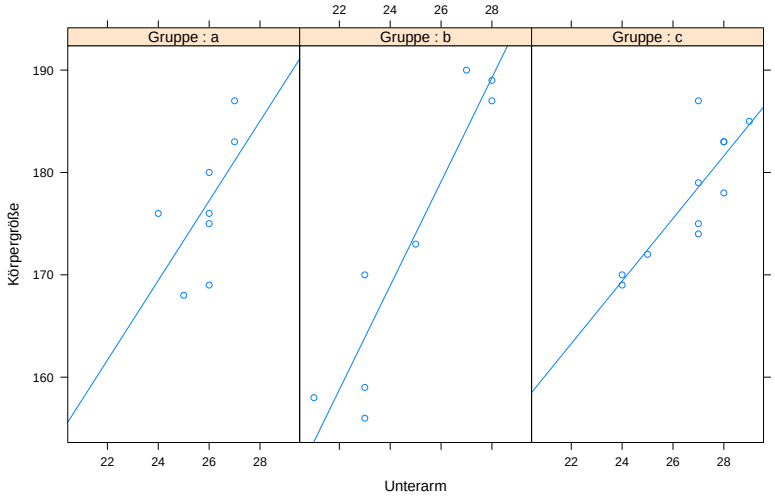
Bsp: Katzen



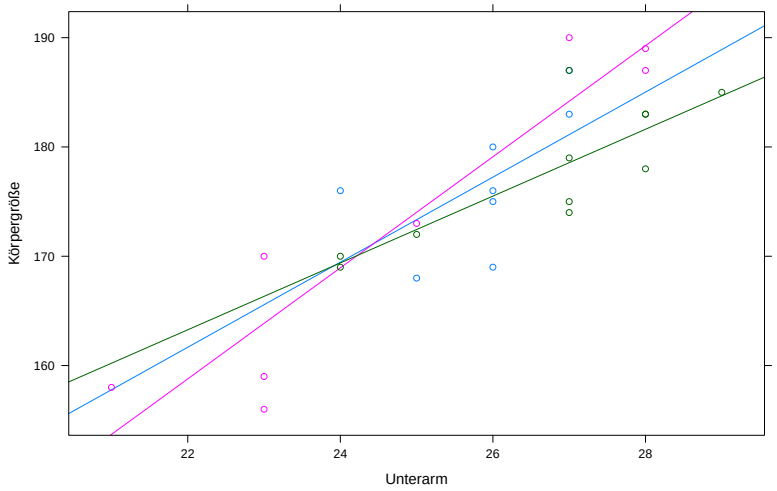
Bsp: Unterarm und Körpergröße

```
> arm <- read.csv("arm09.csv")
> summary(arm)
 Gruppe Geschlecht  Körpergröße       Unterarm
 a: 8   m:17       Min.   :156.0   Min.   :21.00
 b: 8   w:10       1st Qu.:170.0   1st Qu.:24.50
 c:11               Median :176.0   Median :26.00
                    Mean   :176.0   Mean   :25.89
                    3rd Qu.:183.0   3rd Qu.:27.00
                    Max.   :190.0   Max.   :29.00
```

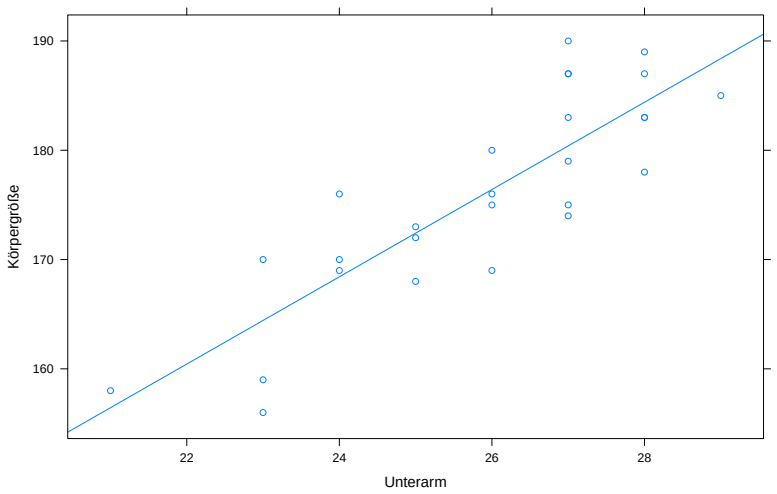

Bsp: Unterarm und Körpergröße



Bsp: Unterarm und Körpergröße



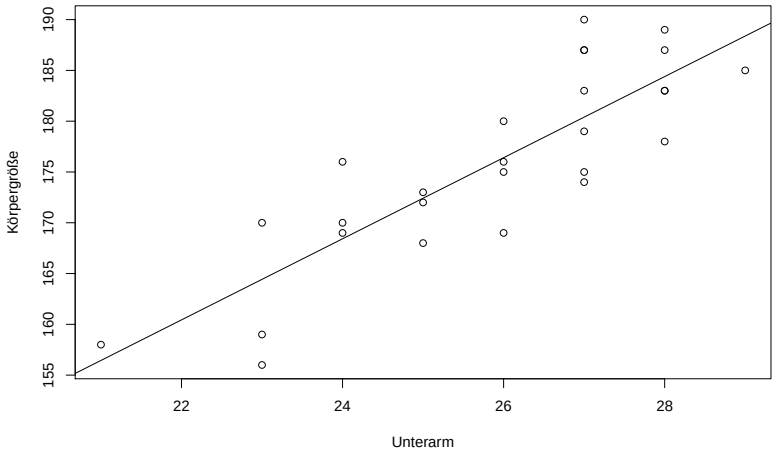
Bsp: Unterarm und Körpergröße



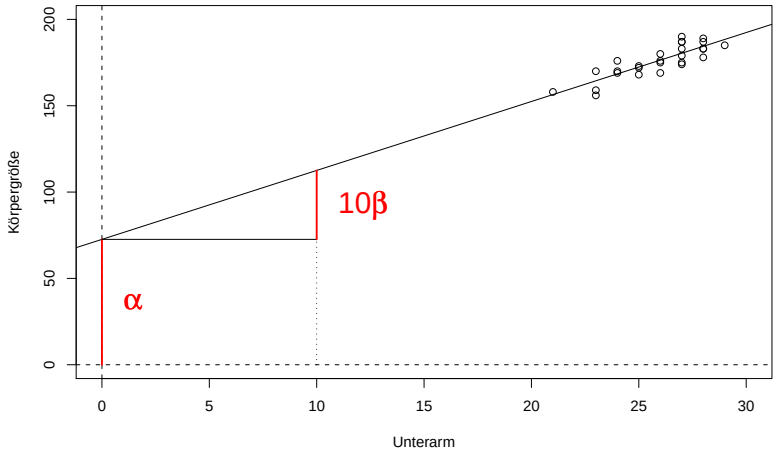
Bsp: Unterarm und Körpergröße

```
> lm1 <- lm(Körpergröße ~ Unterarm, data = arm)
> summary(lm1)
Call:
lm(formula = Körpergröße ~ Unterarm, data = arm)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4293 -3.8990 -0.4066  3.1010  9.6010

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.6035    12.8269   5.660 6.83e-06 ***
Unterarm      3.9924     0.4941   8.081 1.96e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.006 on 25 degrees of freedom
Multiple R-squared: 0.7231,        Adjusted R-squared: 0.7121
F-statistic:  65.3 on 1 and 25 DF,  p-value: 1.955e-08
```


Bsp: Unterarm und Körpergröße



Bsp: Unterarm und Körpergröße



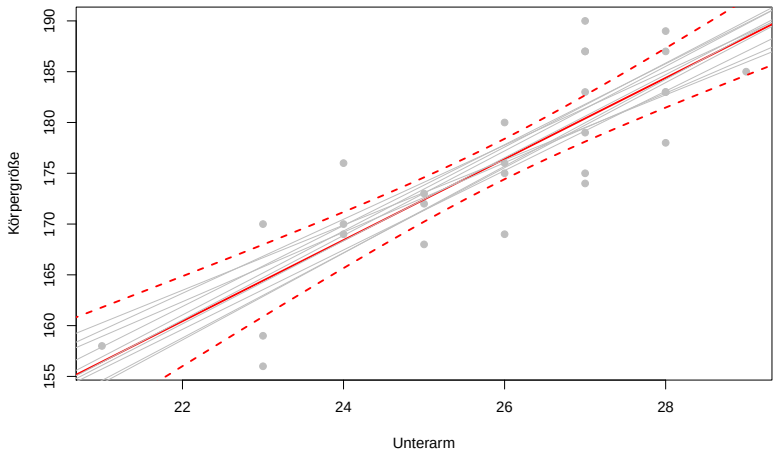
Bsp: Unterarm und Körpergröße

Parameter haben eine bivariate Normalverteilung mit Varianz-Kovarianzmatrix

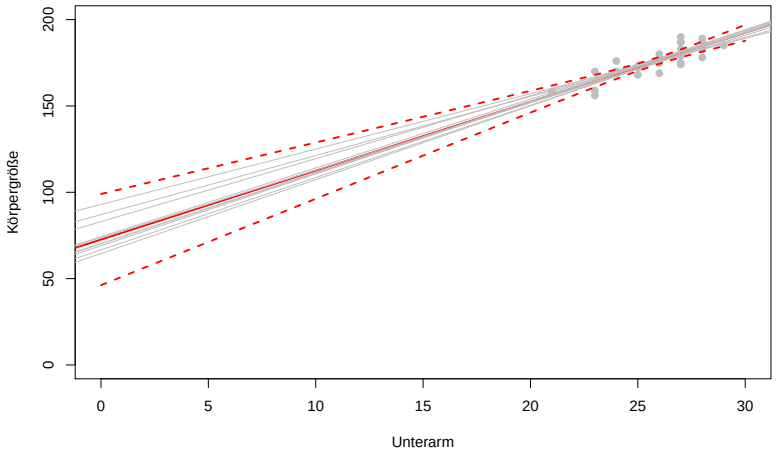
```
            (Intercept)   Unterarm
(Intercept)   164.52963 -6.3193698
Unterarm       -6.31937  0.2440958
```

und Korrelationskoeffizienten $\rho = -0.9972$.

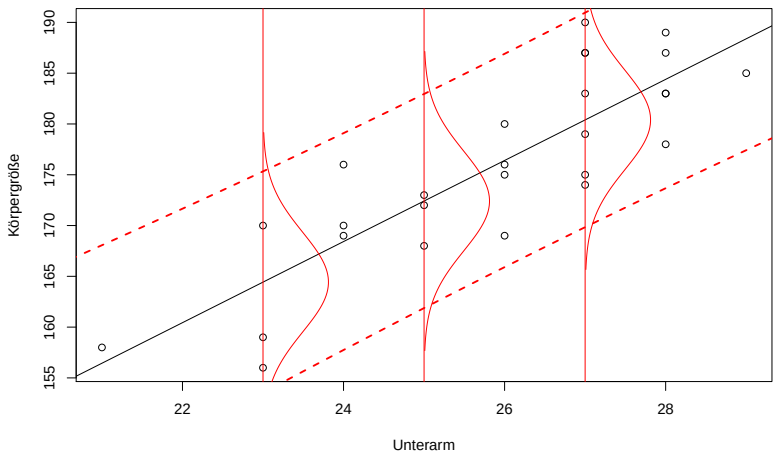
Bsp: Unterarm und Körpergröße



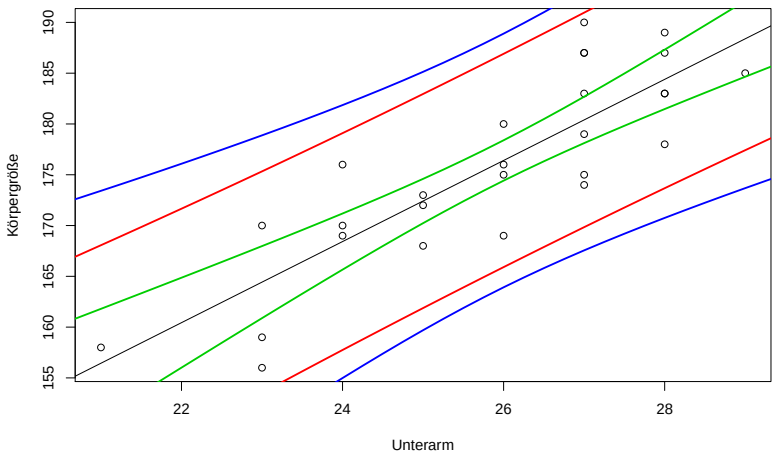
Bsp: Unterarm und Körpergröße



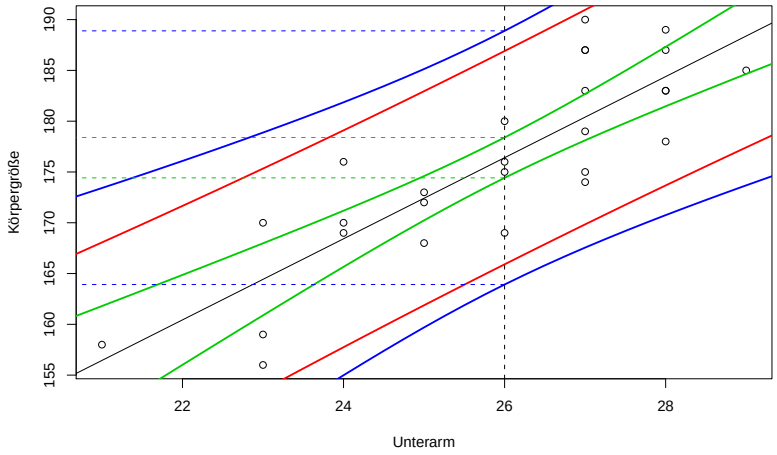
Bsp: Unterarm und Körpergröße



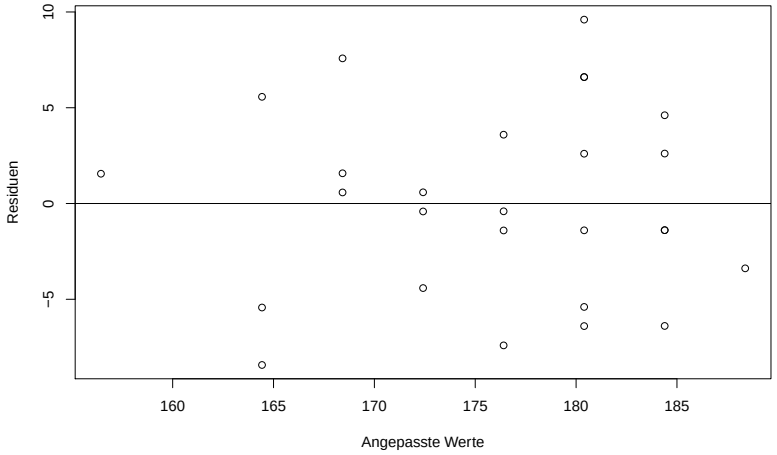
Bsp: Unterarm und Körpergröße



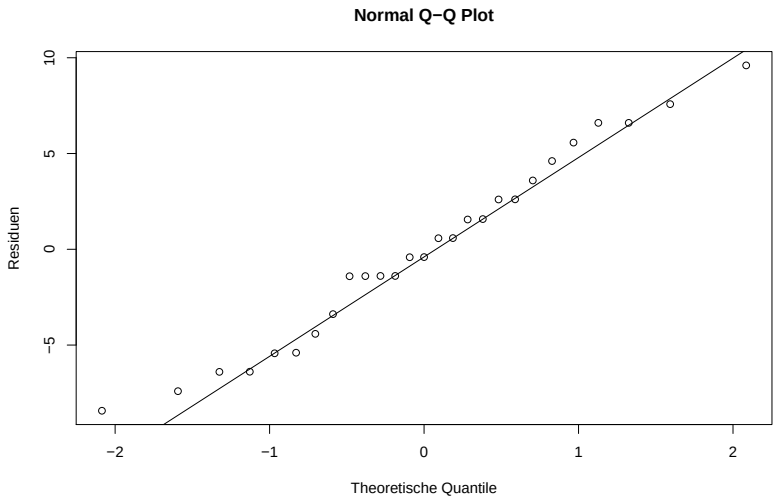
Bsp: Unterarm und Körpergröße



Bsp: Unterarm und Körpergröße



Bsp: Unterarm und Körpergröße

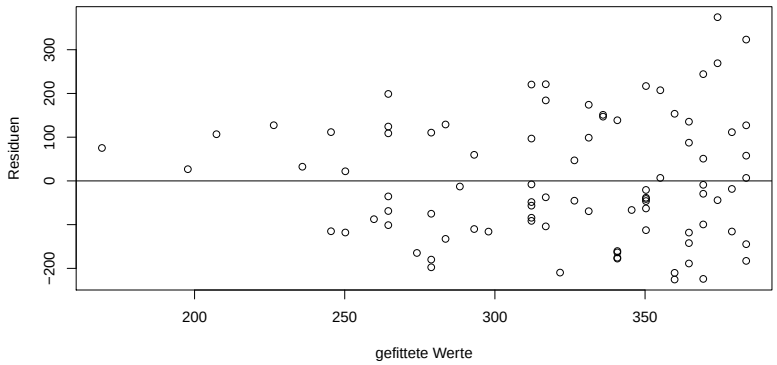


Bsp: Mietspiegel

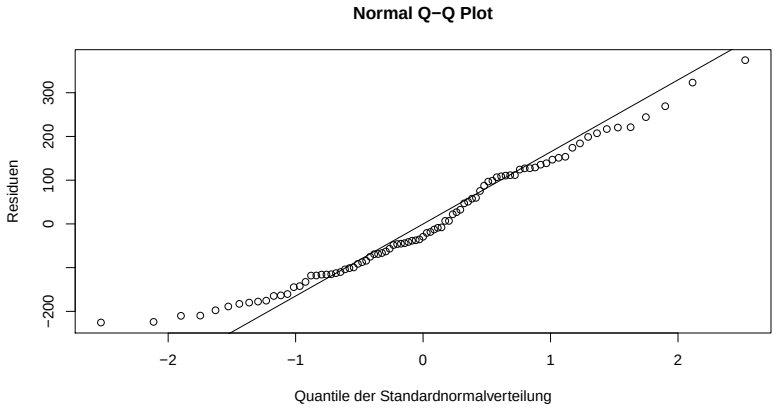
Prognose Nettomiete aus Wohnfläche:

```
Call:
lm(formula = nm ~ wfl, data = miete, plot = FALSE)

Residuals:
    Min      1Q  Median      3Q     Max
-225.47 -111.28  -29.36  110.93  374.41

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   73.822     77.271   0.955  0.34210
wfl            4.767      1.458   3.270  0.00155 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 140 on 85 degrees of freedom
Multiple R-squared: 0.1118,        Adjusted R-squared: 0.1013
F-statistic: 10.69 on 1 and 85 DF,  p-value: 0.001553
```


Bsp: Mietspiegel



Bsp: Mietspiegel

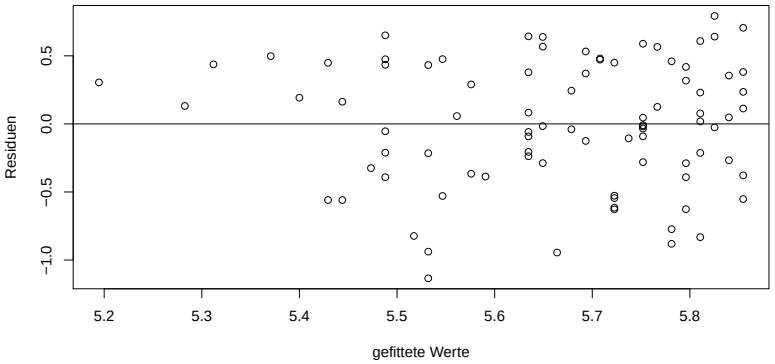


Bsp: Mietspiegel

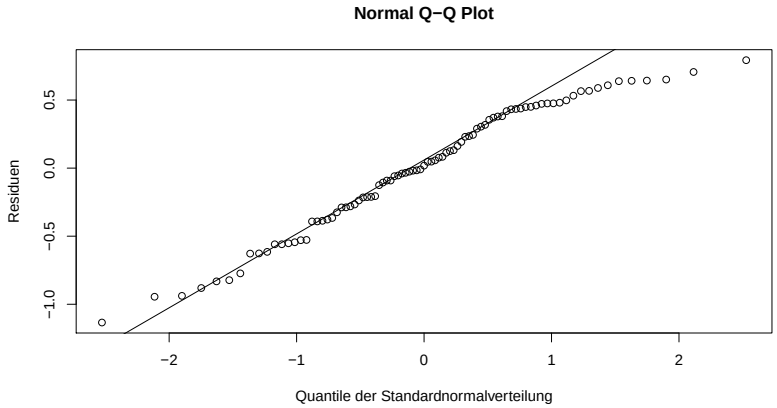
Prognose Logarithmus der Nettomiete aus Wohnfläche:

```
Call:
lm(formula = log(nm) ~ wfl, data = miete, plot = FALSE)

Residuals:
     Min       1Q   Median       3Q      Max
-1.13406 -0.30656  0.01837  0.42550  0.79286

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.901325   0.255908  19.153  < 2e-16 ***
wfl         0.014666   0.004828   3.038  0.00316 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4635 on 85 degrees of freedom
Multiple R-squared: 0.09794,        Adjusted R-squared: 0.08733
F-statistic: 9.229 on 1 and 85 DF,  p-value: 0.003164
```


Bsp: Mietspiegel



Bsp: Mietspiegel



Multiple Regression

Multiple lineare Regression

Ziel: Erweiterung der linearen Einfachregression für mehrere Kovariablen $X_1, \ldots, X_p$

Daten: $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$

Zielvariable Y**:** metrisch bzw. stetig

Kovariablen: metrisch oder kategorial

Metrische Kovariable x kann auch Transformation $x = f(z)$ einer ursprünglichen erklärenden Variablen z sein, z.B. $x = z^2$, $x = ln\, z$, usw..

Bsp: Katzen

Naive Lösung: 2 getrennte Regressionsmodelle

Weiblich:

	Estimate	Std. Error	t	p
Constant	2.9813	1.4855	2.01	0.0508
Bwt	2.6364	0.6254	4.22	0.0001

$$R^2 = 0.28,\ F = 17.77,\ p = 0.0001186$$

Männlich:

	Estimate	Std. Error	t	p
Constant	−1.1841	0.9983	−1.19	0.2385
Bwt	4.3127	0.3399	12.69	0.0000

$$R^2 = 0.62,\ F = 161,\ p < 10^{-15}$$

Bsp: Katzen

„Einfacher“ ist die Einführung von Hilfsvariablen

$m_n = 1$ falls Katze n männlich, sonst 0
$w_n = 1$ falls Katze n weiblich, sonst 0

und die Formulierung des Modells

$$\text{Hwt}_n = \beta_1 m_n + \beta_2 w_n + (\beta_3 m_n + \beta_4 w_n) * \text{Bwt}_n + \epsilon_n$$

das beide Einzel-Regressionsmodelle vereint:

	Estimate	Std. Error	t	p
Constant F	2.9813	1.8428	1.62	0.1080
Constant M	−1.1841	0.9245	−1.28	0.2024
Bwt F	2.6364	0.7759	3.40	0.0009
Bwt M	4.3127	0.3148	13.70	0.0000

Schätzer ident zu Einzelmodellen, Varianz, t und p verschieden.

Bsp: Katzen

Da $m_n = 1 - w_n$, genügt es, eine der beiden Variablen explizit ins Modell aufzunehmen:

$$\text{Hwt}_n = \beta_1 + \tilde{\beta}_2 m_n + (\beta_3 + \tilde{\beta}_4 m_n) * \text{Bwt}_n + \epsilon_n$$

mit $\tilde{\beta}_2 = \beta_2 - \beta 1$ und $\tilde{\beta}_4 = \beta_4 - \tilde{\beta}_3$.

Unterschiede zwischen den Gruppen kann man testen, indem man auf Unterschiede zwischen den Parametern der Gruppen testet:

	Estimate	Std. Error	t	p
Const F	2.9813	1.8428	1.62	0.1080
Bwt F	2.6364	0.7759	3.40	0.0009
Const M - Const F	−4.1654	2.0618	−2.02	0.0453
Bwt M - Bwt F	1.6763	0.8373	2.00	0.0472

Dummy-Kodierung

Kategoriale erklärende Variable mit k Kategorien $1, \ldots, k$ durch $k-1$ Dummy-Variablen $x^{(1)}, \ldots, x^{(k-1)}$ kodiert; mit k als Referenzkategorie.

$$x^{(j)} = \begin{cases} 1, & \text{falls Kategorie } j \text{ vorliegt} \\ 0, & \text{sonst,} \end{cases}$$

wobei $j = 1, \ldots, k-1$.
$x^{(1)} = \ldots = x^{(k-1)} = 0 \quad \Leftrightarrow$ Referenzkategorie k liegt vor.

Je nach Software kann auch Kategorie 1 die Referenzkategorie sein (z.B. in R).

Koeffizienten für die Dummy-Variablen geben dann jeweils Differenz zur Referenzkategorie an.

Standardmodell

Es gilt

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i\,, \quad i = 1, \ldots, n\,.$$

Dabei sind

$Y_1, \ldots, Y_n$	beobachtbare metrische Zufallsvariablen,
$x_{1j}, \ldots, x_{nj}$	deterministische Werte der Variablen X_j oder Realisierungen von Zufallsvariablen X_j,
$\epsilon_1, \ldots, \epsilon_n$	unbeobachtbare Zufallsvariablen, die unabhängig und identisch verteilt sind mit $E(\epsilon_i) = 0$ und $Var(\epsilon_i) = \sigma^2$.

Normalverteilungsannahme:

$$\epsilon_i \sim N(0, \sigma^2) \Leftrightarrow Y_i \mid x_{i1}, \ldots, x_{ip} \sim N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \sigma^2)$$

Matrixnotation

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Y Beobachtungsvektor der Zielvariablen, X Designmatrix

$Y = X\beta + \epsilon$, $E(\epsilon) = 0$; Annahme: Rang von $X = p + 1$

Schätzen & Testen

Schätzer $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)'$ nach dem KQ-Prinzip

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta 1 x_{i1} - \ldots - \beta_p x_{ip})^2 = (Y - X\beta)'(Y - X\beta) \to \min_{\beta}$$

Lösung: KQ-Schätzer

$$\hat{\beta} = (X'X)^{-1} X'Y$$

Analog zur Einfachregression ist der KQ-Schätzer auch der ML-Schätzer: $Y|X$ ist univariat normalverteilt, BW identisch.

Schätzen & Testen

Gefittete Werte:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}$$

Residuen:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i \ , \quad i = 1, \ldots, n.$$

Schätzer für die Varianz σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Schätzen & Testen

Erwartungstreue:

$$E(\hat{\beta}_j) = \beta_j, \ \ j = 0, \ldots, p; \ \ E(\hat{\sigma}^2) = \sigma^2$$

Varianz der Koeffizienten:

$$\sigma_j^2 := Var(\hat{\beta}_j) = \sigma^2 v_j; \ \ v_j \ \ j\text{-tes Diagonalelement von } (X'X)^{-1}$$

Geschätzte Varianz der Koeffizienten:

$$\hat{\sigma}_j^2 = \hat{\sigma}^2 v_j$$

(BW Varianz in VO Lineare Modelle)

Schätzen & Testen

Verteilung der standardisierten Schätzfunktionen:

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim t(n-p-1) \ , \quad j = 0, \ldots, p$$

$(1-\alpha)$-Konfidenzintervalle für β_j:

$$\left[\hat{\beta}_j - \hat{\sigma}_j t_{1-\frac{\alpha}{2}}(n-p-1), \ \hat{\beta}_j + \hat{\sigma}_j t_{1-\frac{\alpha}{2}}(n-p-1)\right]$$

Schätzen & Testen

Teststatistiken:

$$T_j = \frac{\hat{\beta}_j - \beta_{0j}}{\hat{\sigma}_j} \ , \quad j = 0, \ldots, p$$

Hypothesen und Ablehnbereiche:

Hypothesen			Ablehnbereich
$H_0 : \beta_j = \beta_{0j}$	vs.	$H_1 : \beta_j \neq \beta_{0j}$	$\lvert T_j \rvert > t_{1-\frac{\alpha}{2}}(n-p-1)$
$H_0 : \beta_j \geq \beta_{0j}$	vs.	$H_1 : \beta_j < \beta_{0j}$	$T_j < -t_{1-\alpha}(n-p-1)$
$H_0 : \beta_j \leq \beta_{0j}$	vs.	$H_1 : \beta_j > \beta_{0j}$	$T_j > t_{1-\alpha}(n-p-1)$

Schätzen & Testen

Overall–F–Test:

- Hypothesen:

$$H_0: \ \beta_1 = \ldots = \beta_p = 0$$
$$H_1: \ \beta_j \neq 0 \quad \text{für mindestens ein } j$$

- Teststatistik:

$$F = \frac{R^2}{1-R^2}\frac{n-p-1}{p} = \frac{SQE/p}{SQR/(n-p-1)}$$

- Ablehnungsbereich:

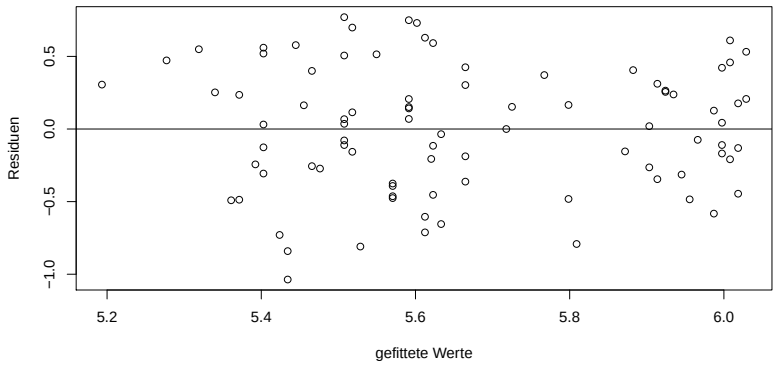
$$F > F_{1-\alpha}(p, n-p-1)$$

Bsp: Mietspiegel

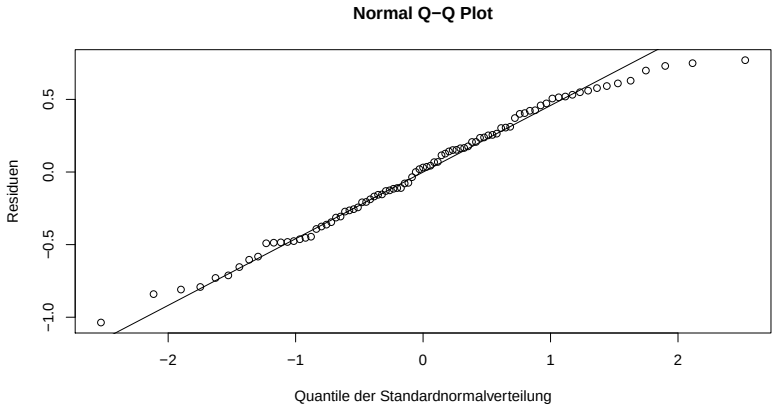
Prognose Log-Nettomiete aus Fläche, Lage und Badaustattung:

```
Call:
lm(formula = log(nm) ~ wfl + wohngut + badextra, data = miete,
    plot = FALSE)

Residuals:
     Min       1Q   Median       3Q      Max
-1.03636 -0.31017  0.03115  0.30874  0.77012

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.983865   0.241857  20.607  < 2e-16 ***
wfl          0.010474   0.004693   2.232 0.028317 *
wohngut      0.364225   0.100320   3.631 0.000488 ***
badextra    -0.237815   0.442659  -0.537 0.592538
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4357 on 83 degrees of freedom
Multiple R-squared: 0.2216,        Adjusted R-squared: 0.1935
F-statistic: 7.877 on 3 and 83 DF,  p-value: 0.0001096
```


Bsp: Mietspiegel



Bsp: Mietspiegel

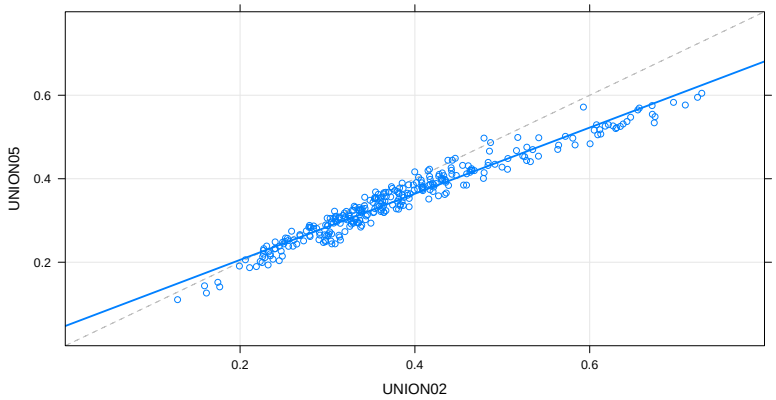


Beispiel: Wahlen

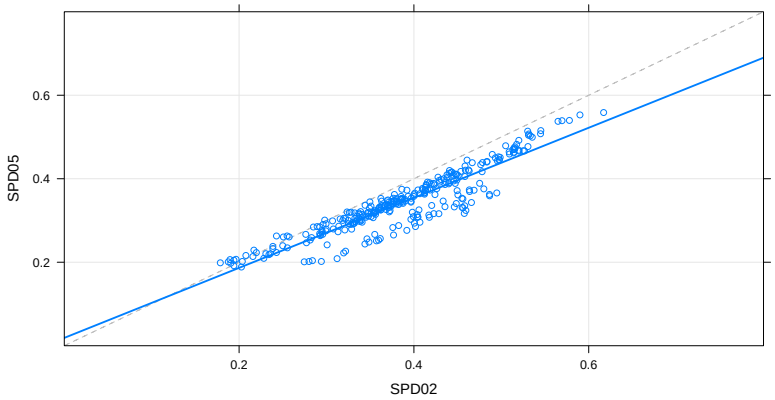
Bsp: Bundestagswahlen

- Daten von Bundestagswahlen 2002 & 2005: Prozent Zweitstimmen der im Parlament vertretenen Parteien in jedem der 299 Wahlkreise.
- Beobachtung: Obwohl absolute Prozentwerte in jedem Wahlkreis sehr unterschiedlich sind, sind relative Veränderungen (Gewinne, Verluste) meist recht ähnlich.
- Regressionsmodell dient als Basis für Hochrechnungen und Wählerstromanalysen

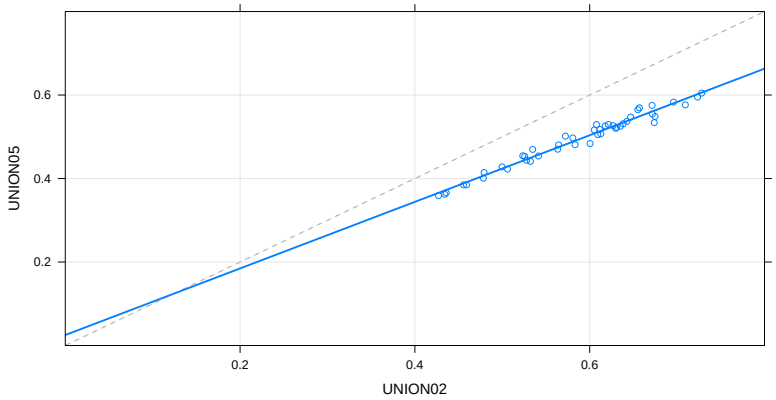
Ganz Deutschland: Union



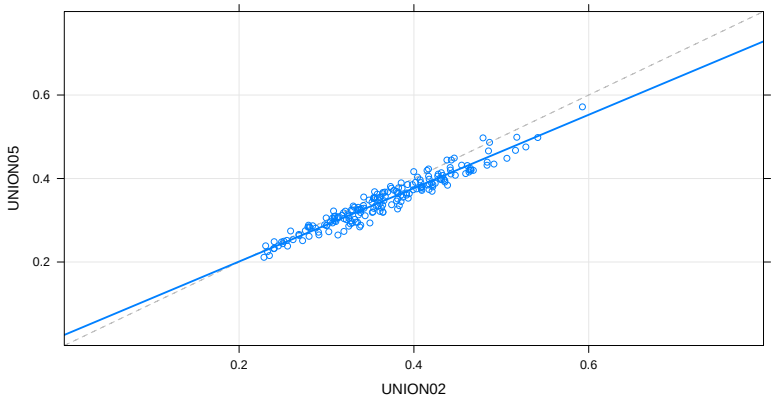
Ganz Deutschland: SPD



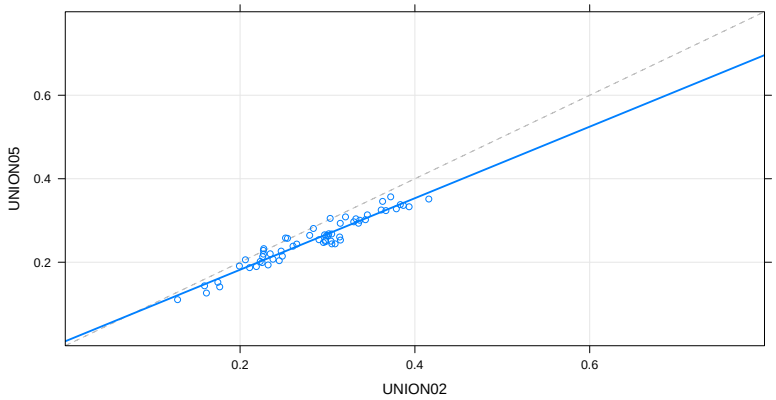
Bayern: Union



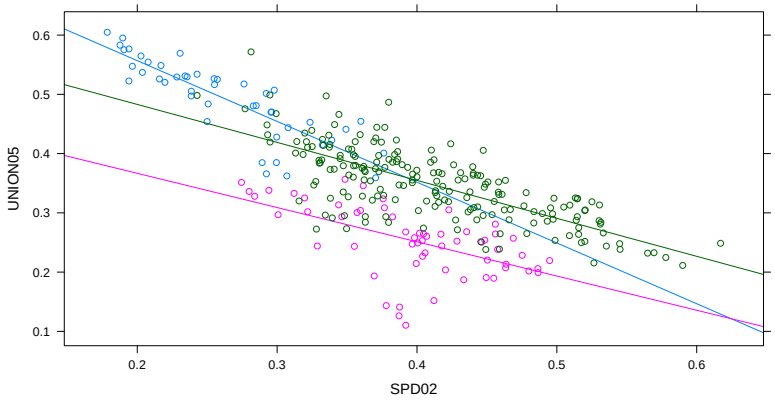
Westen ohne Bayern: Union



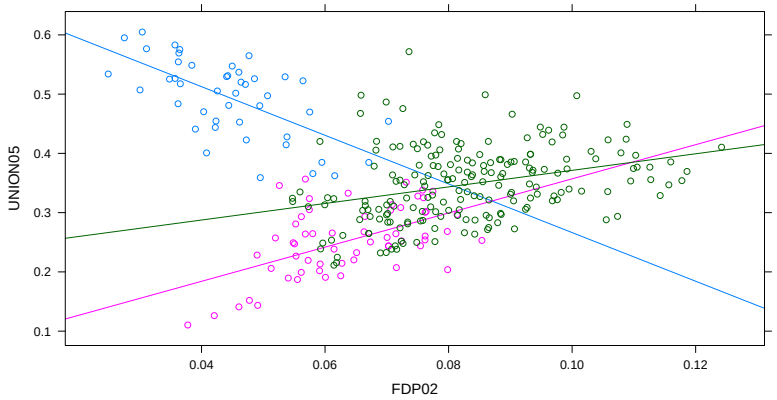
Osten: Union



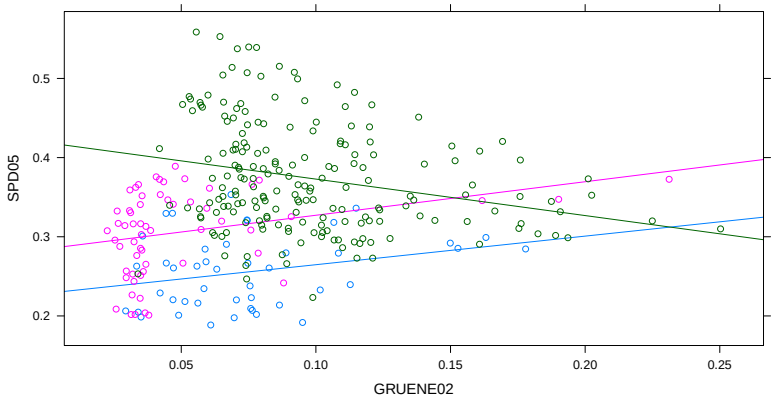
Union 05 vs. SPD 02



Union 05 vs. FDP 02



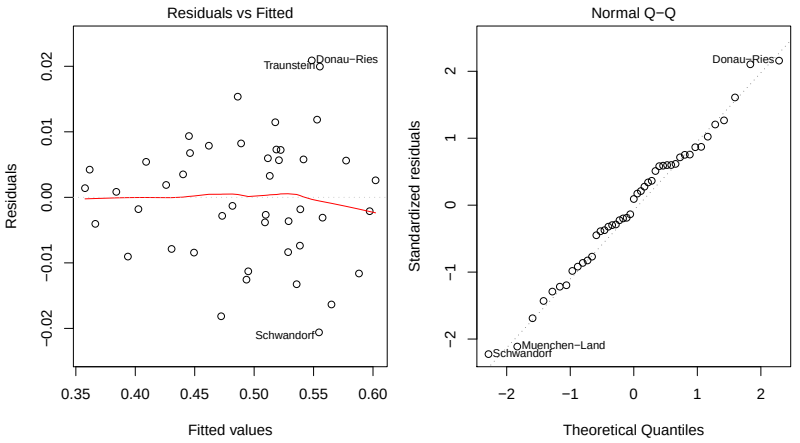
SPD 05 vs. Grüne 02



Prognose Union in Bayern

```
Call:
lm(formula = UNION05 ~ UNION02 + SPD02 + FDP02 + GRUENE02 + LINKE02 -
    1, data = BAYERN)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0206029 -0.0073709  0.0008394  0.0059677  0.0208974

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
UNION02   0.801043   0.010346  77.423  < 2e-16 ***
SPD02     0.006984   0.036099   0.193  0.84758
FDP02     0.869598   0.263899   3.295  0.00207 **
GRUENE02 -0.160878   0.134729  -1.194  0.23948
LINKE02  -0.838218   1.643826  -0.510  0.61291
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009956 on 40 degrees of freedom
Multiple R-squared: 0.9996,        Adjusted R-squared: 0.9996
F-statistic: 2.247e+04 on 5 and 40 DF,  p-value: < 2.2e-16
```

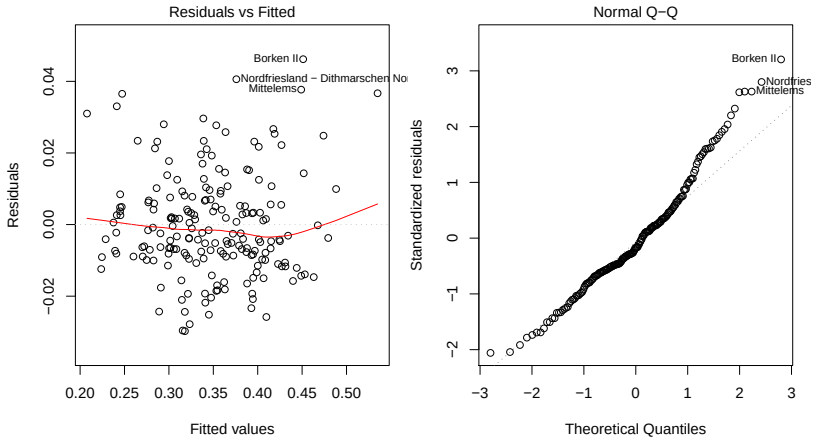

Prognose Union in Bayern



Prognose Union im Westen oB

```
Call:
lm(formula = UNION05 ~ UNION02 + SPD02 + FDP02 + GRUENE02 + LINKE02 -
    1, data = WOB)

Residuals:
      Min        1Q    Median        3Q       Max
-0.029766 -0.008920 -0.002461  0.006935  0.046190

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
UNION02   0.84070    0.01402  59.975  < 2e-16 ***
SPD02     0.05152    0.01153   4.470 1.35e-05 ***
FDP02     0.37827    0.07292   5.188 5.43e-07 ***
GRUENE02  0.04358    0.04466   0.976 0.330382
LINKE02  -1.43033    0.38666  -3.699 0.000283 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01459 on 190 degrees of freedom
Multiple R-squared: 0.9983,        Adjusted R-squared: 0.9983
F-statistic: 2.282e+04 on 5 and 190 DF,  p-value: < 2.2e-16
```

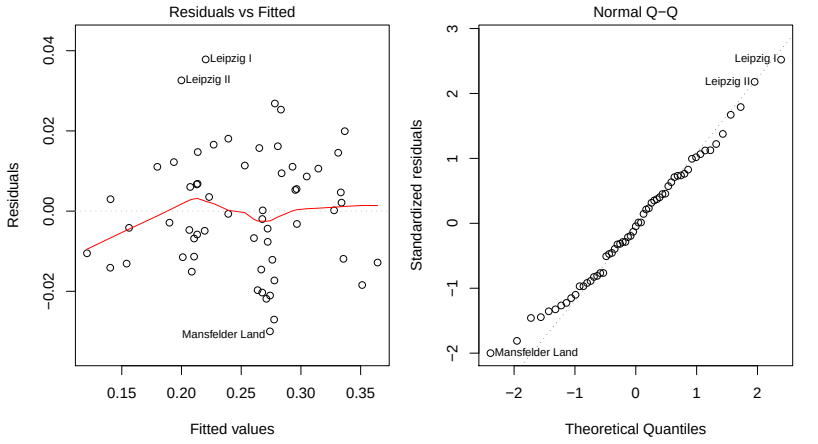

Prognose Union im Westen oB



Prognose Union im Osten

```
Call:
lm(formula = UNION05 ~ UNION02 + SPD02 + FDP02 + GRUENE02 + LINKE02 -
    1, data = OST)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0299906 -0.0116946 -0.0006884  0.0108240  0.0378346

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
UNION02   0.9368426  0.0414105  22.623  < 2e-16 ***
SPD02     0.0843117  0.0276189   3.053  0.00352 **
FDP02    -0.5193158  0.2339320  -2.220  0.03064 *
GRUENE02 -0.0007849  0.0557606  -0.014  0.98882
LINKE02  -0.0739955  0.0555572  -1.332  0.18849
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01536 on 54 degrees of freedom
Multiple R-squared: 0.9968,        Adjusted R-squared: 0.9965
F-statistic:  3340 on 5 and 54 DF,  p-value: < 2.2e-16
```

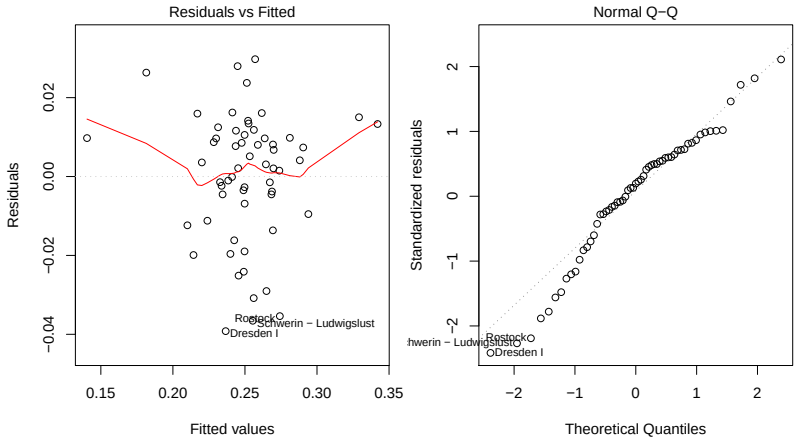

Prognose Union im Osten



Prognose Linke im Osten

```
Call:
lm(formula = LINKE05 ~ UNION02 + SPD02 + FDP02 + GRUENE02 + LINKE02 -
    1, data = OST)

Residuals:
      Min        1Q    Median        3Q       Max
-0.039197 -0.008196  0.003086  0.010183  0.029752

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
UNION02   0.006738   0.045149   0.149   0.8819
SPD02     0.243788   0.030112   8.096 6.91e-11 ***
FDP02     0.453942   0.255051   1.780   0.0807 .
GRUENE02 -0.327579   0.060795  -5.388 1.60e-06 ***
LINKE02   0.835727   0.060573  13.797  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01675 on 54 degrees of freedom
Multiple R-squared: 0.996,        Adjusted R-squared: 0.9956
F-statistic:  2687 on 5 and 54 DF,  p-value: < 2.2e-16
```


Prognose Linke im Osten



In der Realität

Klarerweise sind die hier vorgestellten Modelle für „echte“ Hochrechnungen oder Wählerstromanalysen zu einfach. Es sollte auch berücksichtigt werden:

- Wahlberechigte
- Nichtwähler
- andere Parteien
- Zusammenhänge zwischen den Parteien (Modelle simultan schätzen)
- Wählerströme strikt positiv
- . . .

Wegen des extrem hohen R^2 liefern aber auch diese einfachen Modelle schon recht gute erste Erkenntnisse.