Statistik

Vorlesung 8 (Lineare Modelle)

K.Gerald van den Boogaart
http://www.math-inf.uni-greifswald.de/statistik

■echo=FALSE,term=FALSE,print=FALSE■ options(width=60)
R2 <- function(mod) y <- predict(mod)+resid(mod) gvar <-
var(predict(mod)+resid(mod)) 1-var(resid(mod))/gvar

Lineare Modelle

Def: Statistische Modelle der Form:

$$Y_i = c_0 + c_1 f_1(X_i) + c_2 f_2(X_i) + \ldots c_p f_p(X_i) + \epsilon_i$$

mit $\epsilon_i \sim N(0, \sigma^2)$
(wobei X_i eine Reihe von Einflußgrößen darstellt)

heißen **lineare Modelle.**

Multiple Regression

Wiederholung Lineare Regression

Wir hatten eine Modellvorstellung:

$$Y_i = a + bX_i + \epsilon_i$$

Wiederholung Lineare Regression

Wir hatten eine Modellvorstellung:

$$Y_i = a + bX_i + \epsilon_i$$

wobei a, b unbekannte Parameter,

Wiederholung Lineare Regression

Wir hatten eine Modellvorstellung:

$$Y_i = a + bX_i + \epsilon_i$$

wobei a, b unbekannte Parameter,
ϵ_i (normalverteilter) stochastisch unabhängige Fehler mit konstanter Varianz σ^2

Wiederholung Lineare Regression

Wir hatten eine Modellvorstellung:

$$Y_i = a + bX_i + \epsilon_i$$

wobei a, b unbekannte Parameter,
ϵ_i (normalverteilter) stochastisch unabhängige Fehler mit konstanter Varianz σ^2

und X_i reelle Einflußgrößen sind.

Lineare Regression als Lineares Modell

Lineare Regression:

$$Y_i = a + bX_i + \epsilon_i$$

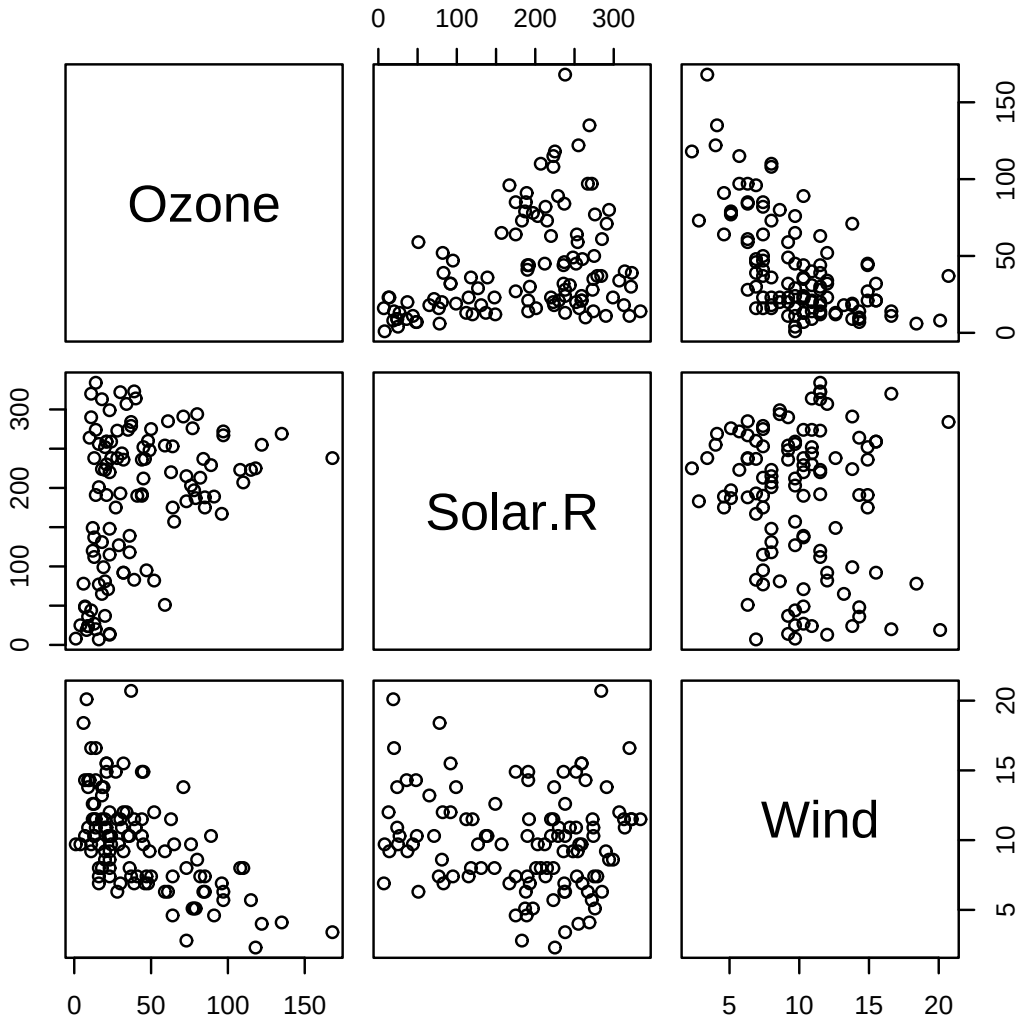
$\epsilon \sim N(0, \sigma^2)$,
als Lineares Modell:

$$Y_i = a + bf_1(X_i) + \epsilon_i$$

$\epsilon \sim N(0, \sigma^2)$,

$$f_1(X) = X$$

Beispiel: Ozon (städtisch)



Beobachtungen

- Sonnenstrahlung erhöht das Ozon
- Wind reduziert das Ozon

Einstrahlung: Parameter

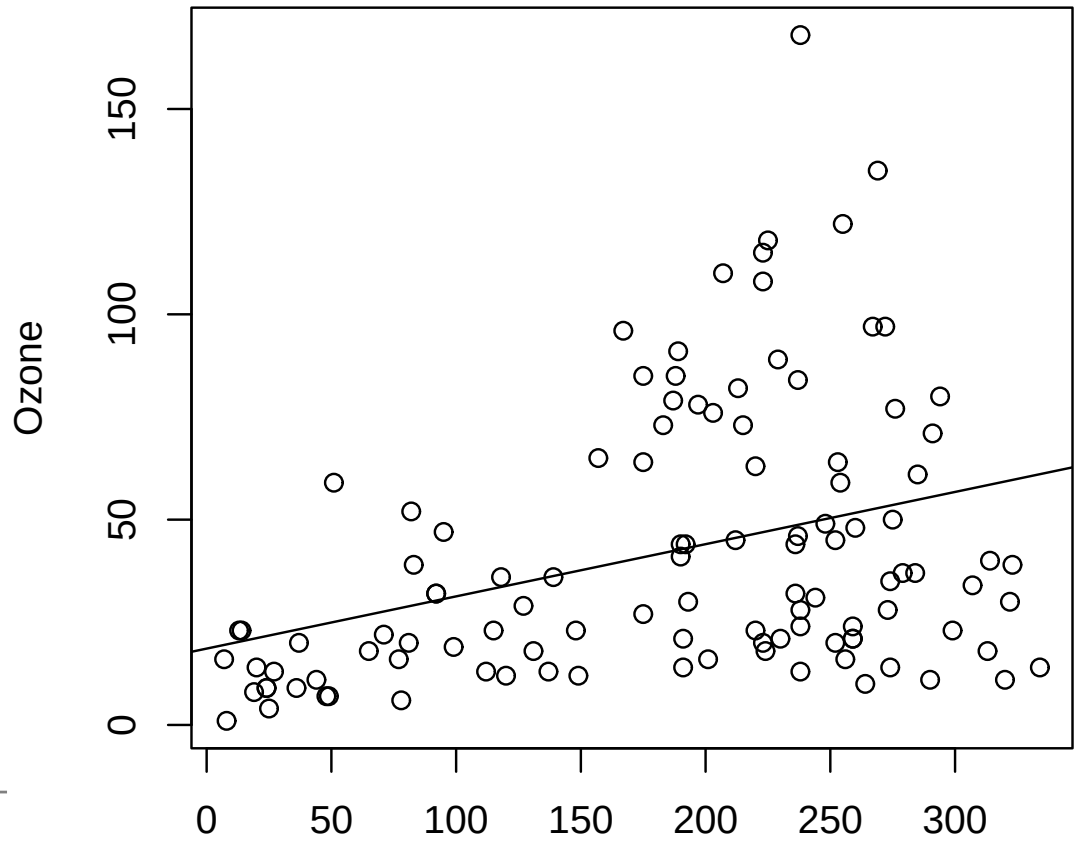
```
> mod <- lm(Ozone ~ Solar.R, data = ozon)
> mod

Call:
lm(formula = Ozone ~ Solar.R, data = ozon)

Coefficients:
(Intercept)      Solar.R
    18.5987       0.1272
```

Einstrahlung: Varianzanalysetabelle

```
> anova(mod)

Analysis of Variance Table

Response: Ozone
           Df Sum Sq Mean Sq F value    Pr(>F)
Solar.R     1  14780   14780  15.053 0.0001793 ***
Residuals 109 107022     982
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.â

> R2(mod)

[1] 0.1213419
```

Einstrahlung: Regressionsgerade

```
> plot(Ozone ~ Solar.R, data = ozon)
> abline(mod)
```



Wind: Parameter

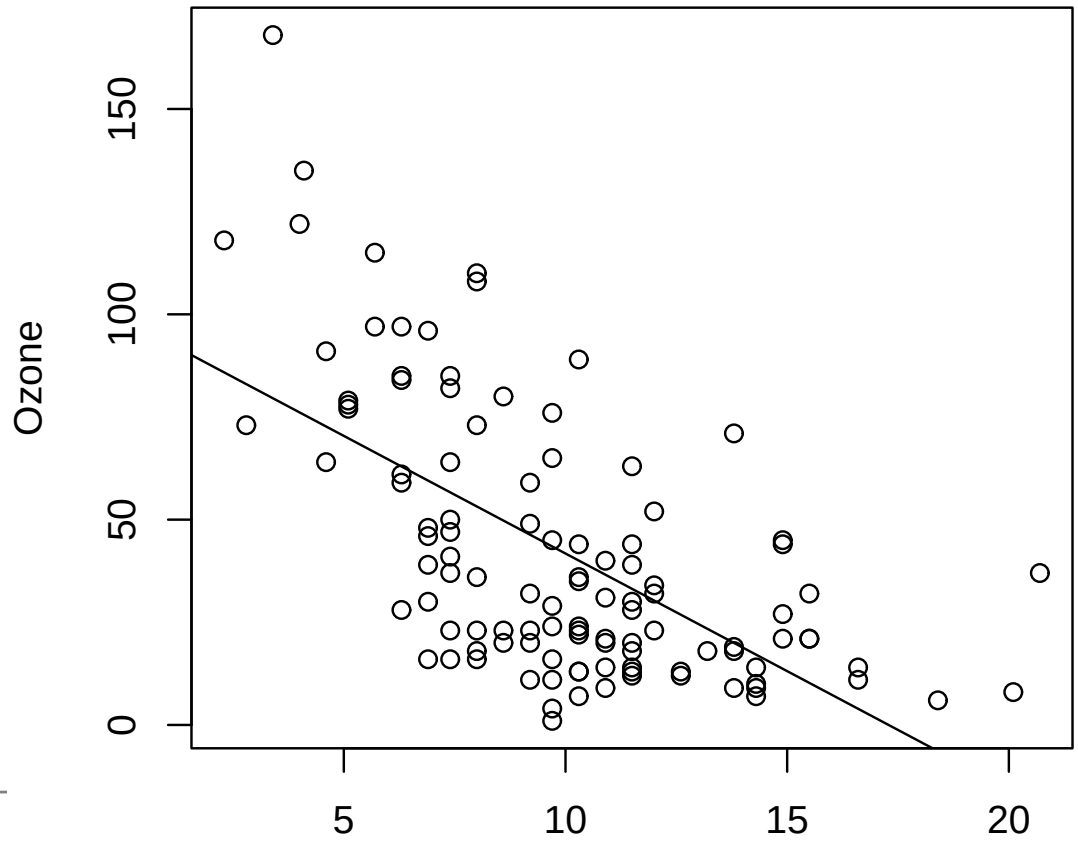
```
> mod <- lm(Ozone ~ Wind, data = ozon)
> mod

Call:
lm(formula = Ozone ~ Wind, data = ozon)

Coefficients:
(Intercept)         Wind
     99.041       -5.729
```

Wind: Tests

```
> anova(mod)

Analysis of Variance Table

Response: Ozone
           Df Sum Sq Mean Sq F value   Pr(>F)
Wind        1  45694   45694  65.442 9.09e-13 ***
Residuals 109  76108     698
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.â

> R2(mod)

[1] 0.3751521
```

Wind: Gerade

```
> plot(Ozone ~ Wind, data = ozon)
> abline(mod)
```



Konzeptionelles Modell

- Ozon entsteht im Umfeld der Verschmutzung durch Sonneneinstrahlung.
- Wind verbläst das enstandene Ozon.
- Problem: Wind und wenig Sonne gehen oft Hand in Hand.
- Beide Einflüsse sind also wichtig.

Multiples Regressionsmodell

Modell 1 (lineare Regression):

$$\text{“}\texttt{Ozone}\text{”}_i = a + b\text{“}\texttt{Solar.R}\text{”}_i + \epsilon_i$$

Modell 2 (lineare Regression):

$$\text{“}\texttt{Ozone}\text{”}_i = a + b\text{“}\texttt{Wind}\text{”}_i + \epsilon_i$$

Kombiniertes Modell (multiple Regression):

$$\text{“}\texttt{Ozone}\text{”}_i = a + b\text{“}\texttt{Wind}\text{”}_i + c\text{“}\texttt{Solar.R}\text{”}_i + \epsilon_i$$

Multiple Regression

- Man hat einen Parameter mehr.
- Die Bestimmung der Parameter wird komplizierter, aber die gleiche Idee der minmalen quadratischen Fehler funktioniert wieder.
- Man kann das natürlich auch mit mehrern Verschiednen Einflüssen machen.
- Daher der Name: Multiple Regression.

Multiple Regression als Lineares Modell

Lineare Regression:

$$Y_i = a + bf_1\left(\begin{pmatrix} \texttt{“Wind”}_i \\ \texttt{“Solar.R”}_i \end{pmatrix}\right) + cf_2\left(\begin{pmatrix} \texttt{“Wind”}_i \\ \texttt{“Solar.R”}_i \end{pmatrix}\right) + \epsilon_i$$

$\epsilon \sim N(0, \sigma^2)$,
mit

$$f_1\left(\begin{pmatrix} \texttt{“Wind”} \\ \texttt{“Solar.R”} \end{pmatrix}\right) = \texttt{“Wind”}$$

$$f_2\left(\begin{pmatrix} \texttt{“Wind”} \\ \texttt{“Solar.R”} \end{pmatrix}\right) = \texttt{“Solar.R”}$$

Parameter

```
> mod <- lm(Ozone ~ Wind + Solar.R, data = ozon)
> mod

Call:
lm(formula = Ozone ~ Wind + Solar.R, data = ozon)

Coefficients:
(Intercept)         Wind      Solar.R
    77.2460      -5.4018       0.1004
```

$$\text{“Ozone”}_i = 77.25 - 5.40 \cdot \text{“Wind”}_i + 0.10 \cdot \text{“Solar.R”} + \epsilon_i$$

Notation für Modelle

Die meisten Programme verwenden eine Notation zur Beschreibung es genauen Modells. Hier z.B.

`Ozone~Wind+Solar.R`

Mit der Bedeutung:

- `Ozone` ist der Regressant
- `Wind` ist Regressor
- `Solar.R` ist Regressor
- Die beiden Einflüsse sollen addiert werden.

$$\text{“}\texttt{Ozone}\text{”} = a + b \cdot \text{“}\texttt{Wind}\text{”}_i + c \cdot \text{“}\texttt{Solar.R}\text{”}_i + \epsilon_i$$

Varianzanalysetabelle

```
> anova(mod)

Analysis of Variance Table

Response: Ozone
           Df Sum Sq Mean Sq F value    Pr(>F)
Wind        1  45694   45694  73.599 7.719e-14 ***
Solar.R     1   9055    9055  14.585 0.0002240 ***
Residuals 108  67053     621
---
Signif. codes:  0 âÃŸ***âÃŹ 0.001 âÃŸ**âÃŹ 0.01 âÃŸ*âÃŹ 0.05 âÃŸ.â
```

Die Anova-Tests

Diese ANOVA Tabelle enthält zwei Tests:

Test Wind

$$H_0 : Y_i = a + \epsilon_i$$
$$vs.$$
$$H_1 : Y_i = a + b \cdot \text{“}\texttt{Wind}\text{”}_i + \epsilon_i$$

Test Solar.R

$$vs.$$
$$H_2 : Y_i = a + b \cdot \text{“}\texttt{Wind}\text{”}_i + c \cdot \text{“}\texttt{Solar.R}\text{”}_i + \epsilon_i$$

In jeder Zeile der Tabelle wird ein Test durchgeführt.

Interpretation der Tests

- Ein signifikanter Test weißt (vorläufig) den Einfluss des Regressors nach.

Interpretation der Tests

- Ein signifikanter Test weißt (vorläufig) den Einfluss des Regressors nach.
- Nach dem Prinzip:
  “So einfach wie mögliche, so kompliziert wie nötig”.

Interpretation der Tests

- Ein signifikanter Test weißt (vorläufig) den Einfluss des Regressors nach.
- Nach dem Prinzip:
  “So einfach wie mögliche, so kompliziert wie nötig”.
- sollten Regressoren mit nicht signifikantem Einflussen entfernt werden,. . .

Interpretation der Tests

- Ein signifikanter Test weißt (vorläufig) den Einfluss des Regressors nach.
- Nach dem Prinzip:
  “So einfach wie mögliche, so kompliziert wie nötig”.
- sollten Regressoren mit nicht signifikantem Einflussen entfernt werden,. . .
- . . . , es sei denn es gibt einen inhaltlichen Grund warum dieser Grund vorhanden sein muß.

Vorläufig, weil . . .

- Wir das Erfülltsein der Voraussetzungen noch nicht geprüft haben.
- Wir noch “Scheineffekte” diskutieren und ausschließen müssen.

Prinzip der sequenziellen Tests

- Für die mehreren Tests innerhalb einer Varianzanalysetabelle wird keine Bonferronie-Korrektur angewendet, da ja das Modell nur dann verwendet wird, wenn alle Tests ablehnen.

Varianzanalysetabelle

```
> anova(mod)

Analysis of Variance Table

Response: Ozone
           Df Sum Sq Mean Sq F value    Pr(>F)
Wind        1  45694   45694  73.599 7.719e-14 ***
Solar.R     1   9055    9055  14.585 0.0002240 ***
Residuals 108  67053     621
---
Signif. codes:  0 âÃŸ***âÃŹ 0.001 âÃŸ**âÃŹ 0.01 âÃŸ*âÃŹ 0.05 âÃŸ.â
```

Bei Effekte sind also vorläufig nachgewiesen.

Voraussetzungen

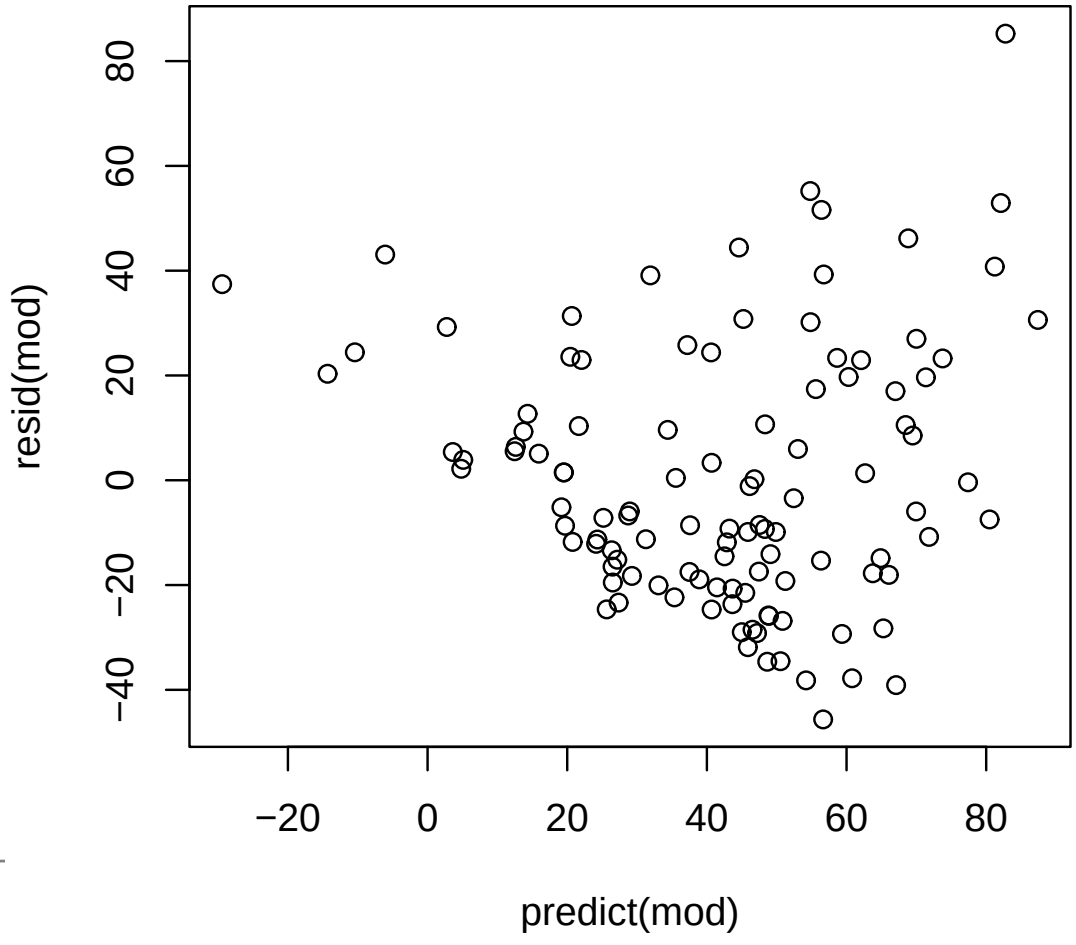
$$Y_i = \text{Formel} + \epsilon_i$$

Die ϵ_i müßen die folgden Eigenschaften haben

- stochastisch unabhängig
- immer die gleiche Varianz
- ungefähr normalverteilt

Residuals vs. Predicted

```
> plot(predict(mod), resid(mod))
```



Heteroskedastizität

- Bei positiv reellen Skalen sieht man oft diese Abhängigkeit der Streuung vom Mittelwert.
- Eine Analyse der Daten auf einer Logrithmischen Skale behebt das Problem oft.
- Für die positive reelle Skala gibt es noch eine Reihe inhaltlicher Argumente für die Verwendung einer logarithmischen Transformation.
- Wir ändern also das Modell.

Parameter

```
> modL <- lm(log(Ozone) ~ log(Wind) + log(Solar.R), data = ozon)
> modL

Call:
lm(formula = log(Ozone) ~ log(Wind) + log(Solar.R), data = ozon)

Coefficients:
 (Intercept)     log(Wind)  log(Solar.R)
      3.6939       -1.1277         0.4484
```

$$\ln(\text{“Ozone”}_i) = 3.7 - 1.13 \cdot \ln(\text{“Wind”}_i) + 0.44 \cdot \ln(\text{“Solar.R”}) + \epsilon_i$$

Varianzanalysetabelle

```
> anova(modL)

Analysis of Variance Table

Response: log(Ozone)
               Df Sum Sq Mean Sq F value    Pr(>F)
log(Wind)       1 28.971  28.971  83.522 4.212e-15 ***
log(Solar.R)    1 16.038  16.038  46.236 6.024e-10 ***
Residuals     108 37.461   0.347
---
Signif. codes:  0 âÃŸ***âÃŹ 0.001 âÃŸ**âÃŹ 0.01 âÃŸ*âÃŹ 0.05 âÃŸ.â
```

Vergleich der Erklärungskraft

```
> R2(mod)

[1] 0.4494936

> R2(modL)

[1] 0.5457572
```

Kriterien für die Modellwahl

- Modell mit nichtsignifikaten Einflüssen (die nicht inhaltlich notwendig sind), gelten als zu kompliziert und werden daher nicht in die Auswahl mit einbezogen.

Kriterien für die Modellwahl

- Modell mit nichtsignifikaten Einflüssen (die nicht inhaltlich notwendig sind), gelten als zu kompliziert und werden daher nicht in die Auswahl mit einbezogen.
- Modelle mit fehlenden Voraussetzungen können nicht direkt eingesetzt werden.

Kriterien für die Modellwahl

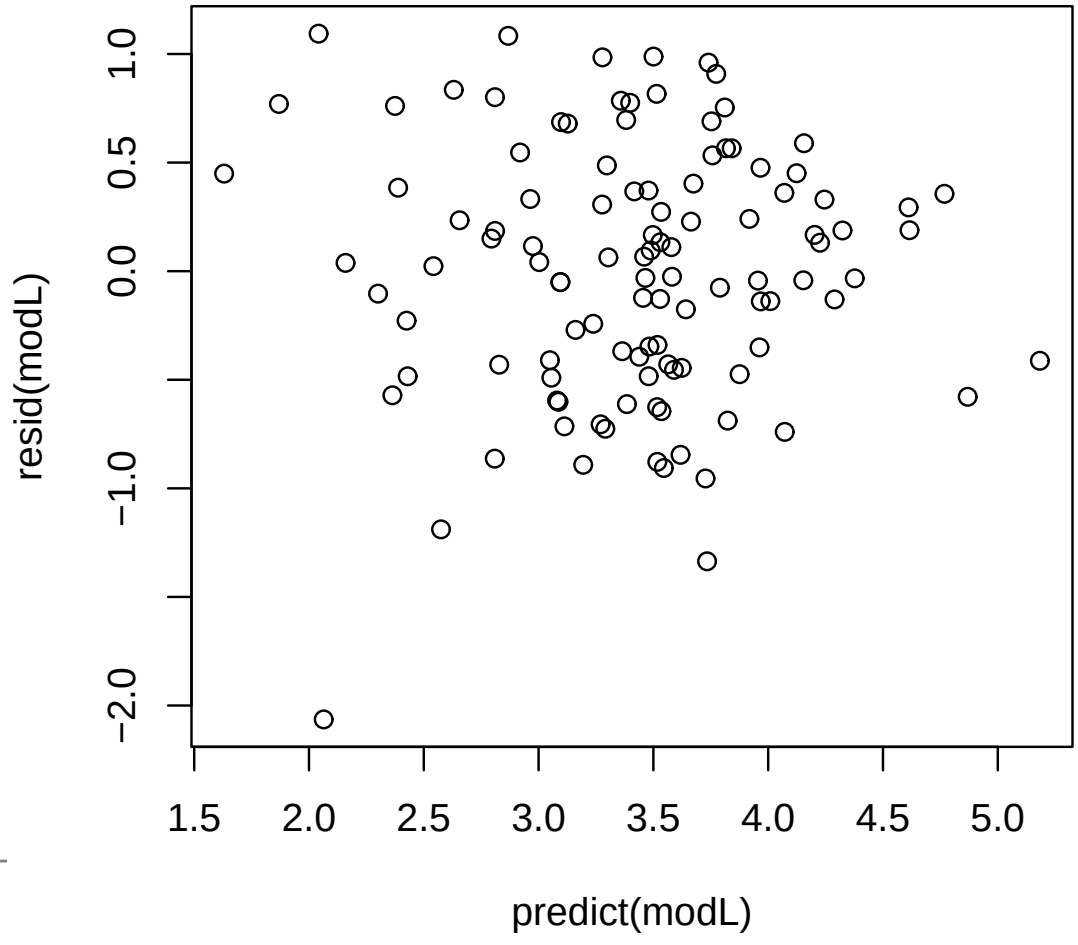
- Modell mit nichtsignifikaten Einflüssen (die nicht inhaltlich notwendig sind), gelten als zu kompliziert und werden daher nicht in die Auswahl mit einbezogen.
- Modelle mit fehlenden Voraussetzungen können nicht direkt eingesetzt werden.
- Modellen mit höherem R^2 wird der Vorzug gegeben.

Kriterien für die Modellwahl

- Modell mit nichtsignifikaten Einflüssen (die nicht inhaltlich notwendig sind), gelten als zu kompliziert und werden daher nicht in die Auswahl mit einbezogen.
- Modelle mit fehlenden Voraussetzungen können nicht direkt eingesetzt werden.
- Modellen mit höherem R^2 wird der Vorzug gegeben.
- Modelle mit einer inhaltlichen Interpretation wird der Vorzug gegeben.

Residuals vs. Predicted

```
> plot(predict(modL), resid(modL))
```

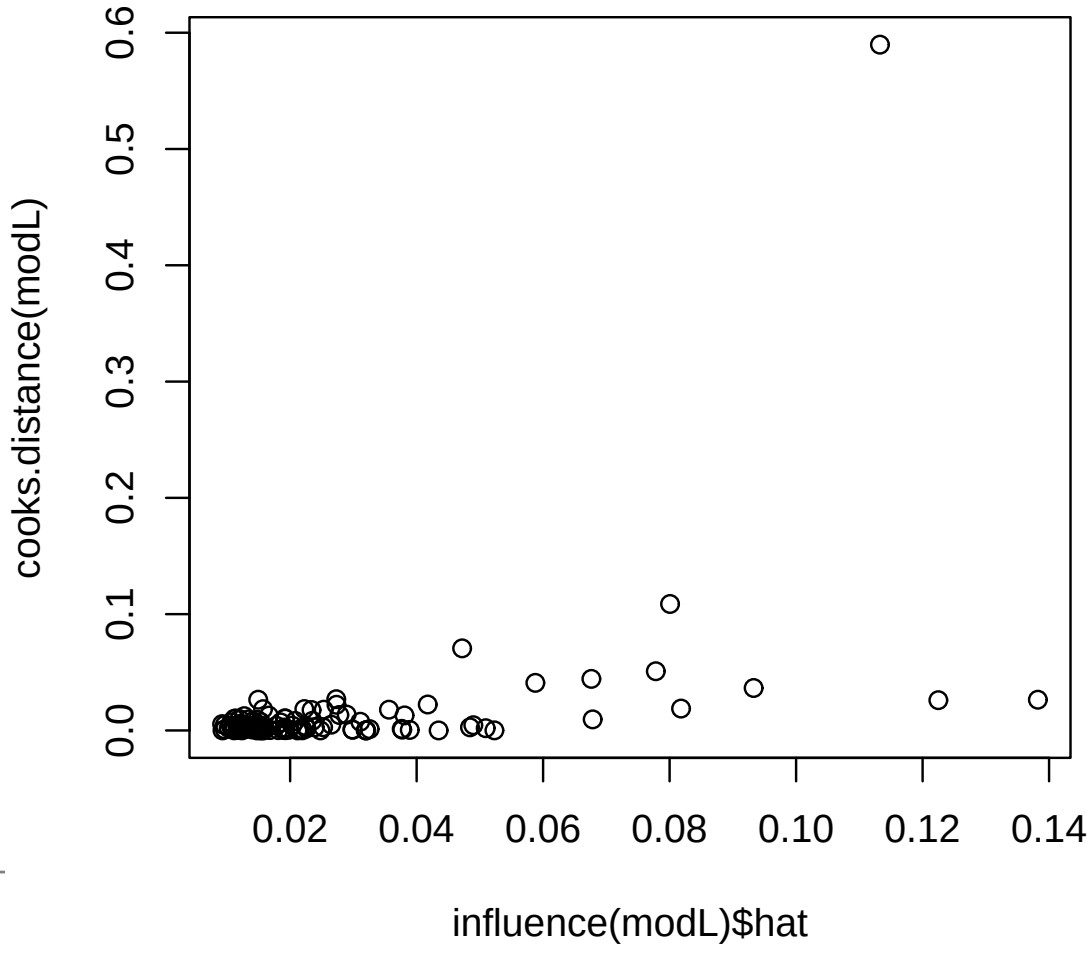


Interpretation

- Nun homoskedastisch.
- Ein Ausreißer ist deutlich sichtbar.

Hebelwirkungen

```
> plot(influence(modL)$hat, cooks.distance(modL))
```



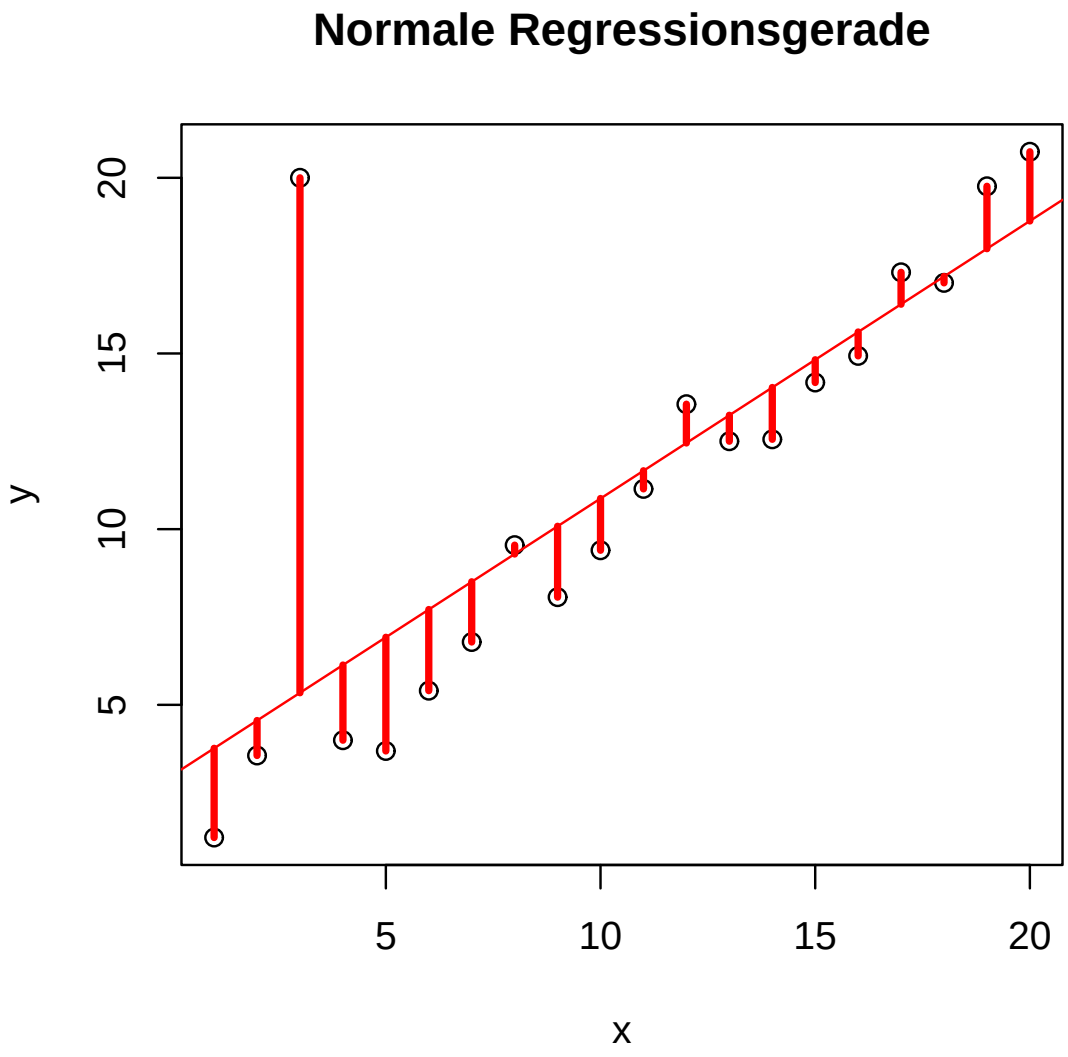
Möglichkeiten bei Ausreißern

- Robuste Schätzung: Computer ignoriert Ausreißer
- Erkannte Ausreißer entfernen: Problematisch, z.B. wenn man auf diese Weise die Lagerstätte entfernt.
- Den Grund des Ausreißers erforschen.

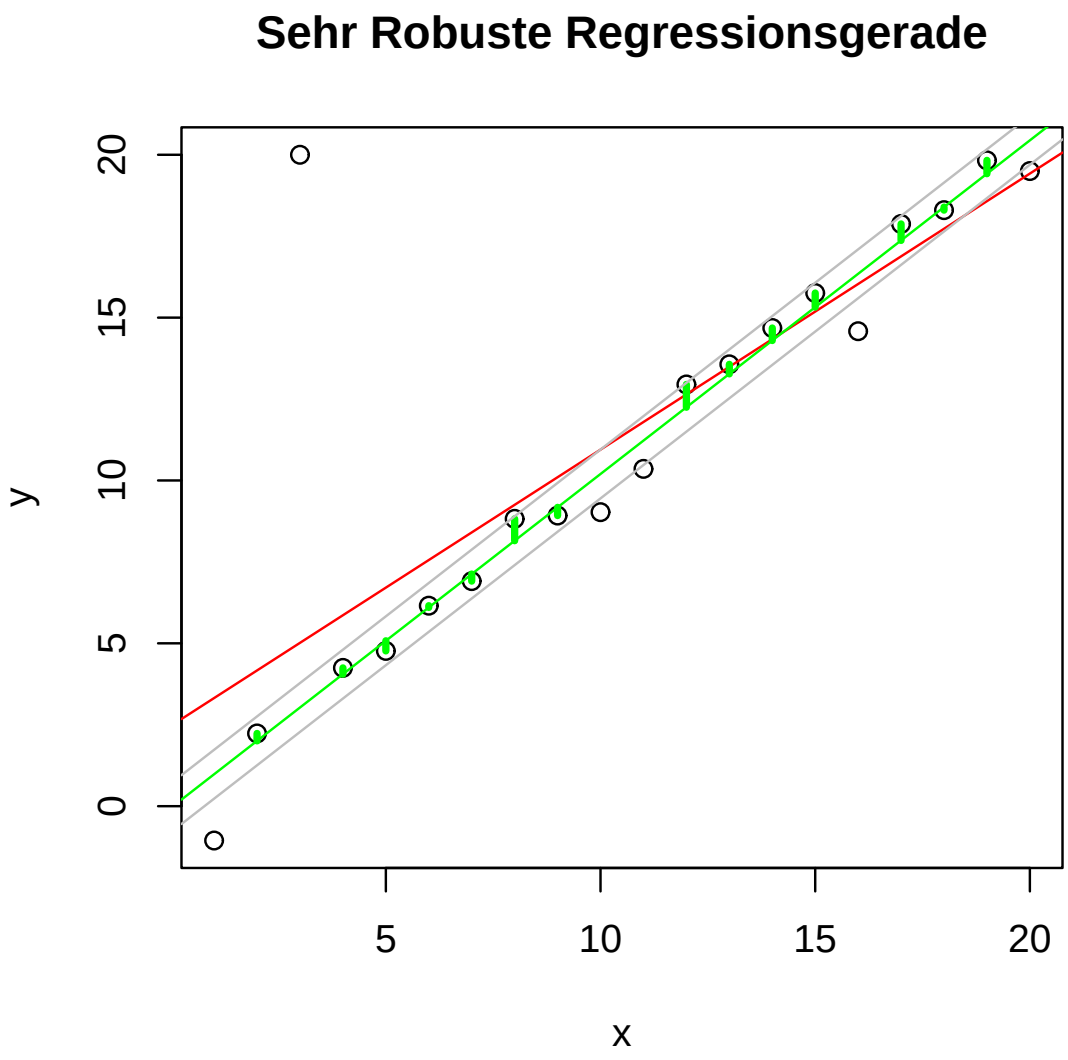
Robuste Schätzung

```
> require(robust)

[1] TRUE

> modLR <- lmRob(log(Ozone) ~ log(Wind) + log(Solar.R), data = ozo
> modLR

Call:
lmRob(formula = log(Ozone) ~ log(Wind) + log(Solar.R), data = ozon

Coefficients:
 (Intercept)    log(Wind) log(Solar.R)
   4.2435732   -1.1560500    0.3560017

Degrees of freedom: 111 total; 108 residual
Residual standard error: 0.6061698

> R2(modLR)
```

Prinzip der robusten Schätzung



Prinzip der robusten Schätzung

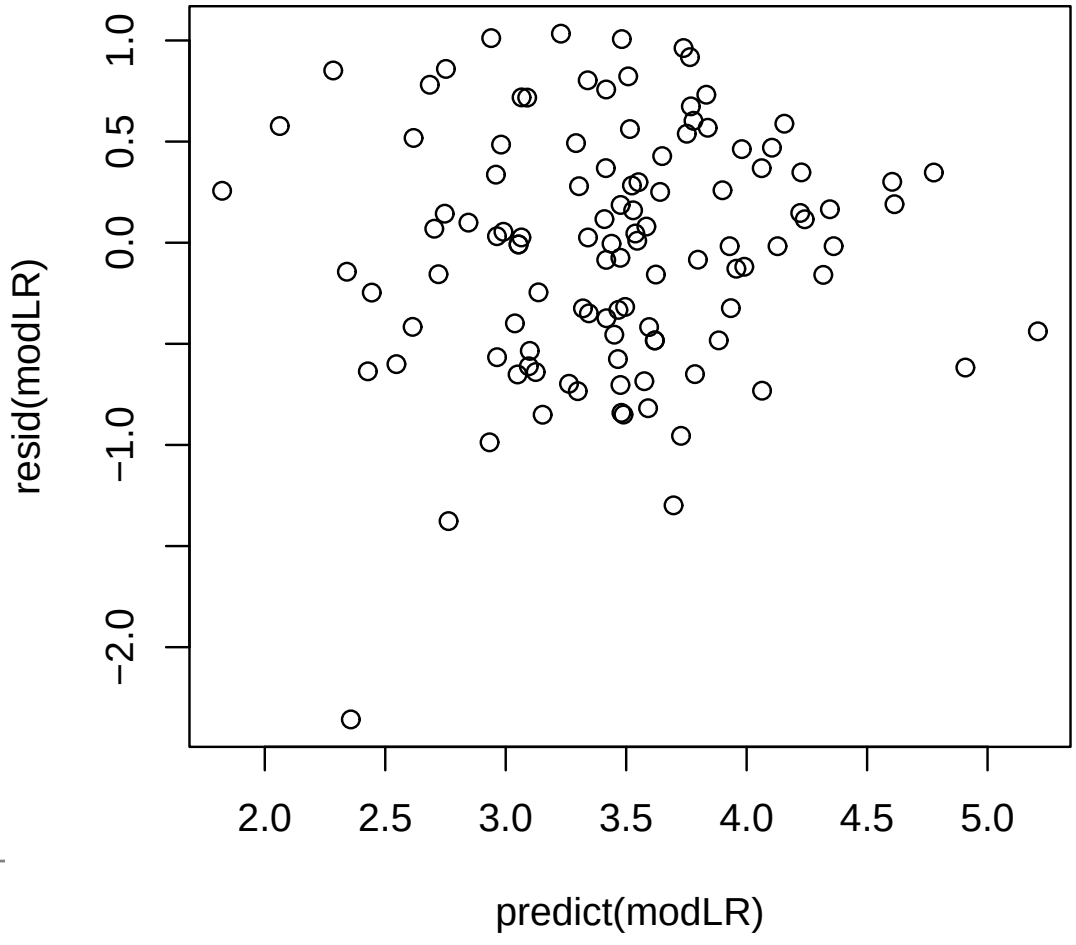


Varianzanalysetabelle

```
> anova(modLR)

Terms added sequentially (first to last)

             Chisq Df RobustF     Pr(F)
(Intercept)          1
log(Wind)            1  63.789 4.441e-16 ***
log(Solar.R)         1  20.528 3.892e-06 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.â
```

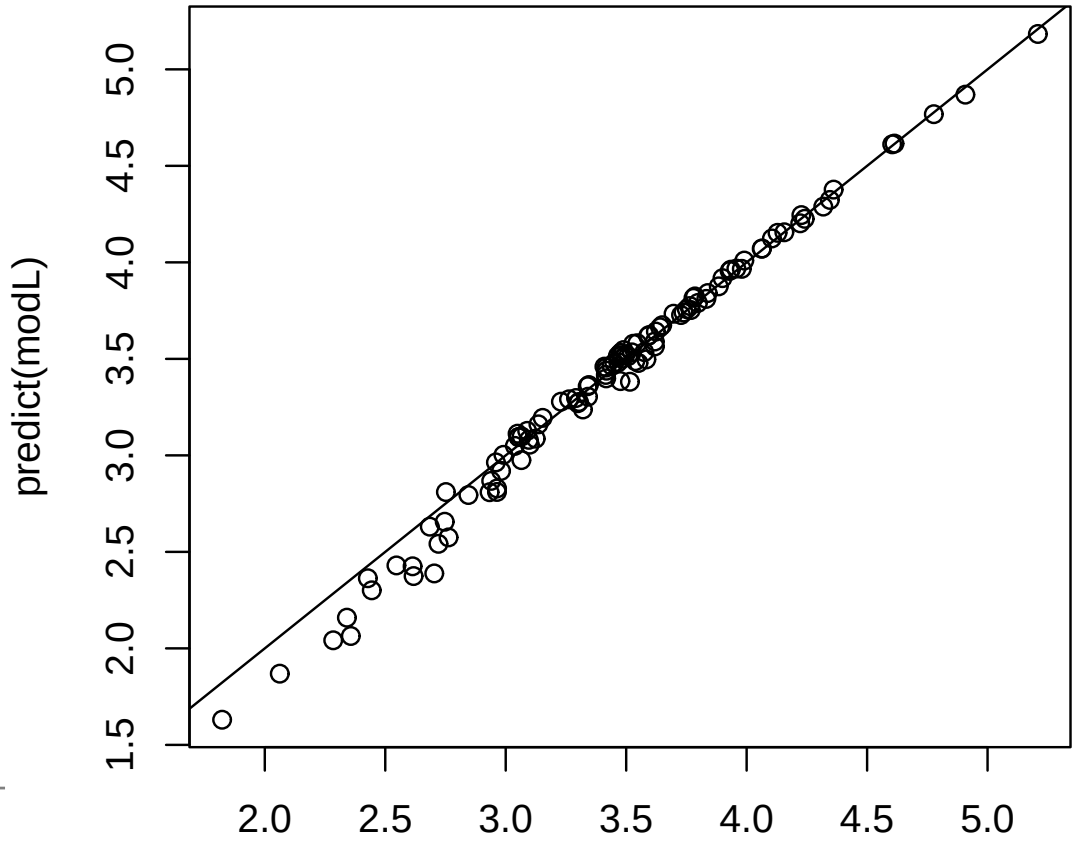
Residuals vs. Predicted

```
> plot(predict(modLR), resid(modLR))
```



Effekt der Robustheit

```
> plot(predict(modLR), predict(modL))
> abline(0, 1)
```



Qualität des Ergebnisses

```
> plot(predict(modLR), log(ozon$Ozone))
> abline(0, 1)
```

Qualität des Vorhersage

```
> plot(exp(predict(modLR)), ozon$Ozone)
> abline(0, 1)
```

Vergleich der Erklärungskraft

```
> R2(mod)

[1] 0.4494936

> R2(modL)

[1] 0.5457572

> R2(modLR)

[1] 0.5374841
```

Ergebnisse

- Unser bevorzugtes Modell:

$$\ln(\text{“}\texttt{Ozon}\text{”}_i) = 4.24 - 1.16 \cdot \ln(\text{“}\texttt{Wind}\text{”}_i) + 0.36 \cdot \ln(\text{“}\texttt{Solar.R}\text{”}_i) + \epsilon_i$$

$\epsilon_i \sim N(0, 0.6061698)$.
Robust geschätzt, da ein Ausreißer vorhanden.

- $R^2 = 0.537$
- Unser Modell erklärt die Hälfte der (geometrischen) Ozonvariabilität.
- Wind reduziert Ozonbelastung.
- Sonneneinstrahlung erhöht Ozonbelastung.
- Allerdings sind die Unterschiede zum additiven Modell minimal!

Das Log-Modell

- Unser bevorzugtes Modell:

$$\ln(\text{“}\texttt{Ozon}\text{”}_i) = 4.24 - 1.16 \cdot \ln(\text{“}\texttt{Wind}\text{”}_i) + 0.36 \cdot \ln(\text{“}\texttt{Solar.R}\text{”}_i) + \epsilon_i$$

$$\epsilon_i \sim N(0, 0.6061698).$$

- e^x auf beide Seiten angewendet:

$$\text{“}\texttt{Ozon}\text{”}_i = e^{4.24} \cdot e^{-1.16 \cdot \ln(\text{“}\texttt{Wind}\text{”}_i)} \cdot e^{0.36 \cdot \ln(\text{“}\texttt{Solar.R}\text{”}_i)} e_i^{\epsilon}$$

$$\epsilon_i \sim N(0, 0.6061698).$$

- und als multiplikatives Modell umformuliert:

$$\text{“}\texttt{Ozon}\text{”}_i = 69.7 \cdot \text{“}\texttt{Wind}\text{”}_i^{-1.16} \cdot \text{“}\texttt{Solar.R}\text{”}_i^{0.36} \cdot \epsilon_i$$

$$\epsilon_i \sim LogN(0, 0.6061698).$$ (Lognormalverteilung)

Zusammenfassung: Multiple Regression

Fast Alles geht genauso, wie bei der einfachen linearen Regression

- Schätzung durch minimale Quadrate
- Test in Varianzanalysetabelle
- Bewertung der Wichtigkeit durch R^2
- Diagnostische Methoden: Residuen, Hebelwirkung, Cook-Distanzen
- Robuste Schätzung
- Nicht: Einfache Visualisierung des Modells

Interaktion in der multiplen Regression

Konzept: Interaktion

- Möglicherweise wirkt die Sonneneinstrahlung ja bei verschiedenen Windverhältnissen verschieden.

Konzept: Interaktion

- Möglicherweise wirkt die Sonneneinstrahlung ja bei verschiedenen Windverhältnissen verschieden.
- Modell z.B.

$$\ln(\text{“}\texttt{Ozone}\text{”}) = a + b \cdot \ln(\text{“}\texttt{Wind}\text{”}) + (c + d \cdot \ln(\text{“}\texttt{Wind}\text{”})) \cdot \ln(\text{“}\texttt{Solar.R}\text{”}) + \epsilon$$

Konzept: Interaktion

- Möglicherweise wirkt die Sonneneinstrahlung ja bei verschiedenen Windverhältnissen verschieden.
- Modell z.B.

$$\ln(\text{“}\texttt{Ozone}\text{”}) = a + b \cdot \ln(\text{“}\texttt{Wind}\text{”}) + (c + d \cdot \ln(\text{“}\texttt{Wind}\text{”})) \cdot \ln(\text{“}\texttt{Solar.R}\text{”}) + \epsilon$$

- Ausmultiplizieren:

$$\ln(\text{“}\texttt{Ozone}\text{”}) = a + b \cdot \ln(\text{“}\texttt{Wind}\text{”}) + c \cdot \ln(\text{“}\texttt{Solar.R}\text{”}) + d \cdot \underbrace{\ln(\text{“}\texttt{Wind}\text{”}) \cdot \ln(\text{“}\texttt{Solar.R}\text{”})}_{\text{Interaktion}}$$

Konzept: Interaktion

- Möglicherweise wirkt die Sonneneinstrahlung ja bei verschiedenen Windverhältnissen verschieden.
- Modell z.B.

$$\ln(\text{“}\texttt{Ozone}\text{”}) = a + b \cdot \ln(\text{“}\texttt{Wind}\text{”}) + (c + d \cdot \ln(\text{“}\texttt{Wind}\text{”})) \cdot \ln(\text{“}\texttt{Solar.R}\text{”}) + \epsilon$$

- Ausmultiplizieren:

$$\ln(\text{“}\texttt{Ozone}\text{”}) = a + b \cdot \ln(\text{“}\texttt{Wind}\text{”}) + c \cdot \ln(\text{“}\texttt{Solar.R}\text{”}) + d \cdot \underbrace{\ln(\text{“}\texttt{Wind}\text{”}) \cdot \ln(\text{“}\texttt{Solar.R}\text{”})}_{\text{Interaktion}}$$

-

$$\ln(\text{“}\texttt{Ozone}\text{”}) = a + (b + d \cdot \ln(\text{“}\texttt{Solar.R}\text{”})) \cdot \ln(\text{“}\texttt{Wind}\text{”}) + c \cdot \ln(\text{“}\texttt{Solar.R}\text{”}) + \epsilon$$

Interaktion im Computer angeben

```
> modLI <- lm(log(Ozone) ~ log(Wind) + log(Solar.R) + log(Wind) *
+     log(Solar.R), data = ozon)
> modLI

Call:
lm(formula = log(Ozone) ~ log(Wind) + log(Solar.R) + log(Wind) *

Coefficients:
           (Intercept)              log(Wind)              log(Sola
                1.3648                 -0.1385                    0.
log(Wind):log(Solar.R)
               -0.1922
```

. . . und Abkürzen

```
> modLI <- lm(log(Ozone) ~ log(Wind) * log(Solar.R), data = ozon)
> modLI

Call:
lm(formula = log(Ozone) ~ log(Wind) * log(Solar.R), data = ozon)

Coefficients:
            (Intercept)                 log(Wind)                 log(Sola
                 1.3648                   -0.1385                        0.
log(Wind):log(Solar.R)
                -0.1922
```

R^2 vergleichen

```
> R2(mod)

[1] 0.4494936

> R2(modL)

[1] 0.5457572

> R2(modLI)

[1] 0.5494746
```

Mit mehr Parametern steigt R^2 grundsätzlich an!

Varianzanalysetabelle

```
> anova(modLI)

Analysis of Variance Table

Response: log(Ozone)
                         Df Sum Sq Mean Sq F value    Pr(>F)
log(Wind)                 1 28.971  28.971 83.4316 4.639e-15 ***
log(Solar.R)              1 16.038  16.038 46.1860 6.318e-10 ***
log(Wind):log(Solar.R)    1  0.307   0.307  0.8829    0.3495
Residuals               107 37.155   0.347
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.â
```

Ergebnis

- **Ergebnis:** Eine Interaktion konnte nicht nachgewiesen werden.

Ergebnis

- **Ergebnis:** Eine Interaktion konnte nicht nachgewiesen werden.
- Es gibt keinen Hinweis in den Daten, dass der Einfluß des Windes von der Sonneneinstrahlung abhängt.

Ergebnis

- **Ergebnis:** Eine Interaktion konnte nicht nachgewiesen werden.
- Es gibt keinen Hinweis in den Daten, dass der Einfluß des Windes von der Sonneneinstrahlung abhängt.
- Es gibt keinen Hinweis in den Daten, dass der Einfluß der Sonneneinstrahlung vom Wind abhängt.

Scheinzusammenhänge

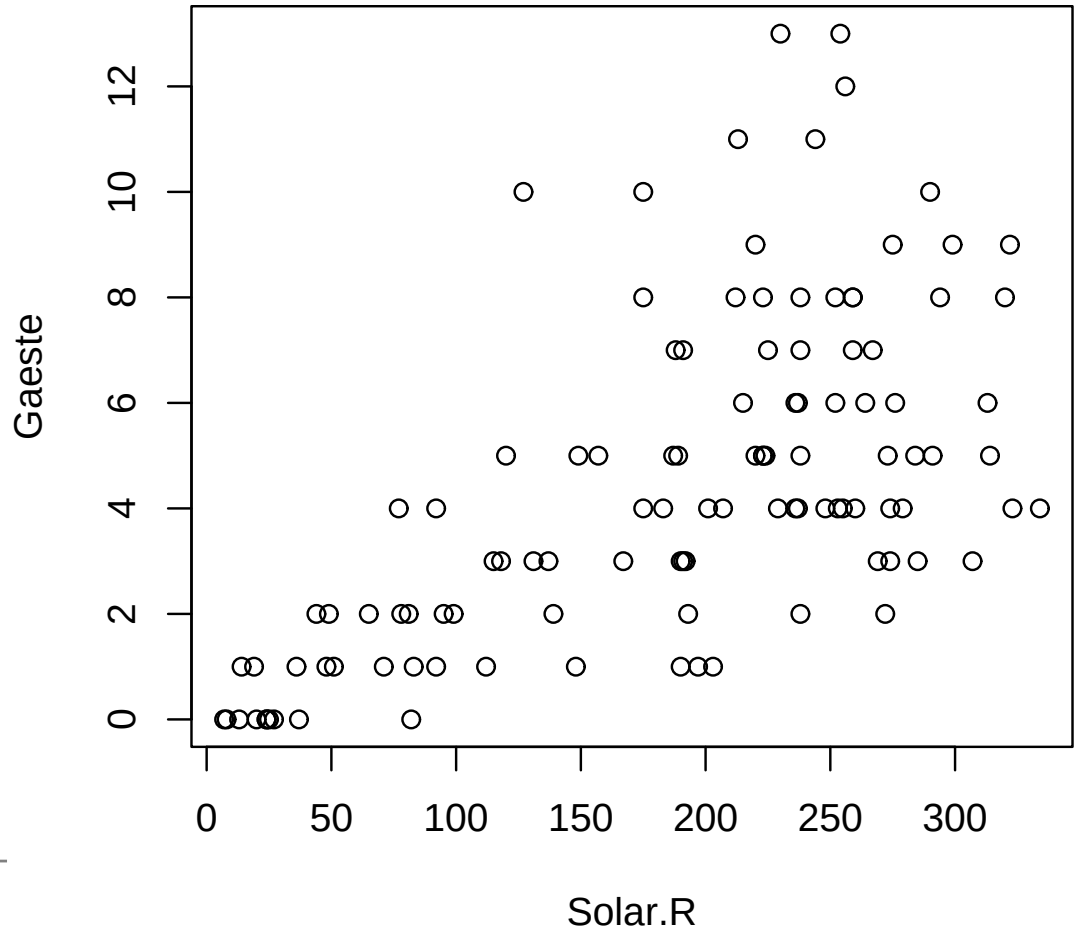
Problem: Scheinzusammenhang

Angenommen wir hätten auch noch die Anzahl der Badegäste im Strandhotel gezählt:

```
> ozon$Gaeste <- rpois(nrow(ozon), ozon$Solar.R/40)
> ozon$Gaeste
  [1]  8  3  2  5 11  0  0  5 11 12  1 13  3  1 11  1  0  5  0  3
 [26]  6  8  5  6  5  0  1  4  6  6  7  3 11  8  6  2  6  7  6  2
 [51]  0  7  6  2  6  7  5  6  7  1  1  4  7  8  9  6  5  4  5  2
 [76]  6  7  5  5  2  9 10  4  3  6  3  4  3  5 10  5  6  7  9  3
[101]  6  7  6  0  5  1  1  5  7  2  3
```

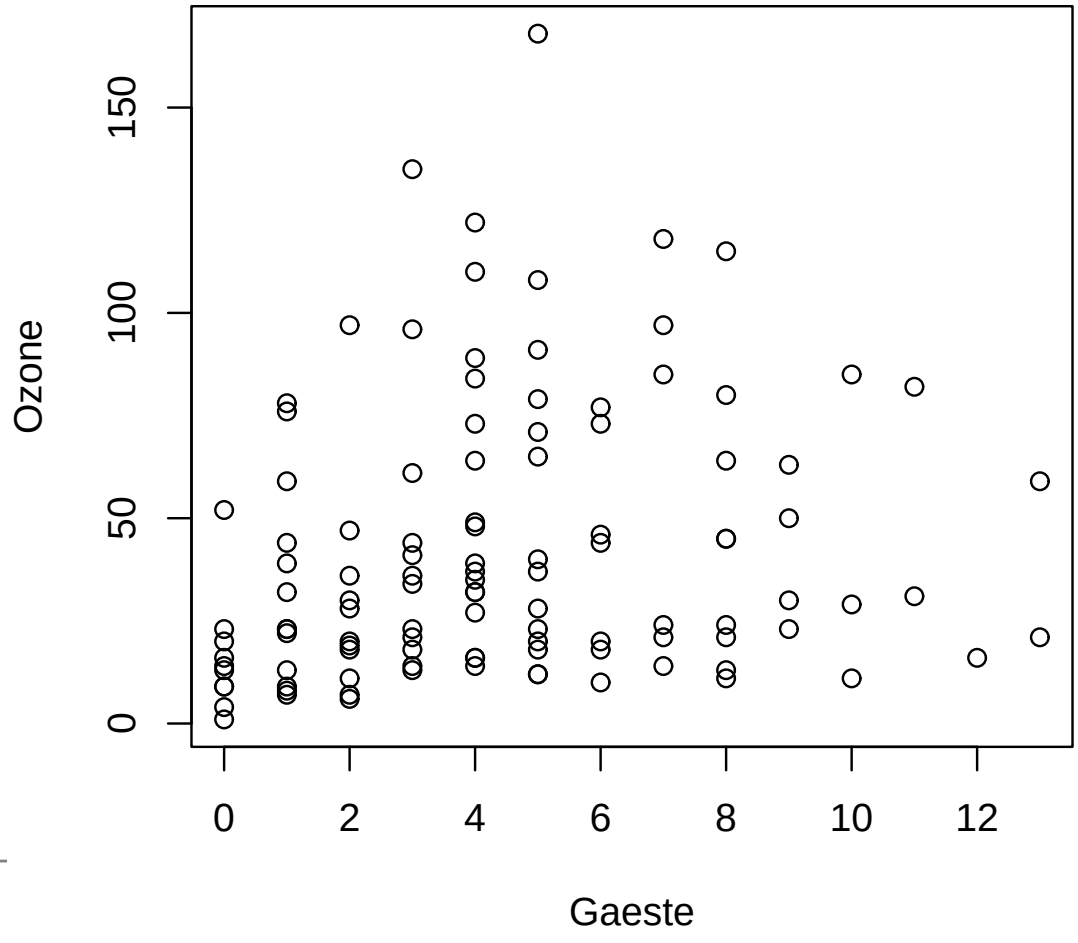
Sonne und Gaeste

```
> with(ozon, plot(Solar.R, Gaeste))
```



Machen Badegäste Ozon?

```
> plot(Ozone ~ Gaeste, data = ozon)
```



Regression mit Gaesten

```
> modXX <- lm(Ozone ~ Gaeste, data = ozon)
> modXX

Call:
lm(formula = Ozone ~ Gaeste, data = ozon)

Coefficients:
(Intercept)       Gaeste
     33.048        2.046
```

Varianzanalysetabelle

```
> anova(modXX)

Analysis of Variance Table

Response: Ozone
           Df Sum Sq Mean Sq F value  Pr(>F)
Gaeste      1   4543    4543  4.2233 0.04226 *
Residuals 109 117259    1076
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.â
```

Überraschendes Ergebnis

- Es besteht ein Zusammenhang zwischen Ozon und Badegästen.
- Natürlich bewirken Badegäste kein Ozon.
- Vielmehr bewirkt die Sonne, sowohl Ozon als auch Badegäste.

Modell zusammen mit Sonneneinstrahlung

```
> modXX1 <- lm(log(Ozone) ~ Gaeste + Solar.R, data = ozon)
> anova(modXX1)

Analysis of Variance Table

Response: log(Ozone)
           Df Sum Sq Mean Sq F value    Pr(>F)
Gaeste      1  6.047   6.047  10.005  0.002027 **
Solar.R     1 11.145  11.145  18.439 3.844e-05 ***
Residuals 108 65.277   0.604
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.â
```

Modell zusammen mit Sonneneinstrahlung

```
> modXX1 <- lm(log(Ozone) ~ Gaeste + Solar.R, data = ozon)
> anova(modXX1)

Analysis of Variance Table

Response: log(Ozone)
           Df Sum Sq Mean Sq F value    Pr(>F)
Gaeste      1  6.047   6.047  10.005  0.002027 **
Solar.R     1 11.145  11.145  18.439 3.844e-05 ***
Residuals 108 65.277   0.604
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.â
```

Die Anova-Tests

Diese ANOVA Tabelle enthält zwei Tests:

$$H_0 : Y_i = a + \epsilon_i$$

Test Gaeste $\quad vs.$

$$H_1 : Y_i = a + b \cdot \text{“}\texttt{Gaeste}\text{”}_i + \epsilon_i$$

Test Solar.R $\quad vs.$

$$H_2 : Y_i = a + b \cdot \text{“}\texttt{Gaeste}\text{”} +_i + c \cdot \text{“}\texttt{Solar.R}\text{”}_i + \epsilon_i$$

In jeder Zeile der Tabelle wird ein Test durchgeführt.

Welcher Nachweis fehlt?

Wir haben bewiesen:

- Ozon nicht konstant ist, sondern irgendwie bei mehr Gaesten mehr wird.
- Ozon nicht nur von Gaeste abhängt, sondern auch von der Einstrahlung.
- Wir haben nicht überprüft, ob eine Abhängigkeit von den Gästen auch dann besteht, wenn die Sonneneinstrahlung als Einfluß bekannt ist.

Modell in umgekehrter Reihenfolge

```
> modXX1a <- lm(log(Ozone) ~ Solar.R + Gaeste, data = ozon)
> anova(modXX1a)

Analysis of Variance Table

Response: log(Ozone)
           Df Sum Sq Mean Sq F value    Pr(>F)
Solar.R     1 17.157  17.157 28.3852 5.478e-07 ***
Gaeste      1  0.036   0.036  0.0595    0.8077
Residuals 108 65.277   0.604
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.â
```

Die Gaeste können also nichts erklären, was nicht bereits durch die Sonneeinstrahlung erklärt ist.

Verdacht

- Vielleicht ist das ja auch mit dem Wind so.
- Chemisch gesehen entsteht Ozon bei Sonneneinstrahlung.
- Viel Wind geht einher mit wenig Einstrahlung.
- Wind kann Ozon doch weder produzieren noch vernichten, oder?
- Vielleicht haben wir hier auch einen Scheinzusammenhang.

Wir hatten,. . .

```
> anova(lm(log(Ozone) ~ log(Wind) + log(Solar.R), data = ozon))

Analysis of Variance Table

Response: log(Ozone)
              Df Sum Sq Mean Sq F value    Pr(>F)
log(Wind)      1 28.971  28.971  83.522 4.212e-15 ***
log(Solar.R)   1 16.038  16.038  46.236 6.024e-10 ***
Residuals    108 37.461   0.347
---
Signif. codes:  0 âÃŸ***âÃŹ 0.001 âÃŸ**âÃŹ 0.01 âÃŸ*âÃŹ 0.05 âÃŸ.â
```

. . . und wechseln die Reihenfolge

```
> anova(lm(log(Ozone) ~ log(Solar.R) + log(Wind), data = ozon))

Analysis of Variance Table

Response: log(Ozone)
              Df Sum Sq Mean Sq F value    Pr(>F)
log(Solar.R)   1 23.899  23.899  68.901 3.239e-13 ***
log(Wind)      1 21.109  21.109  60.857 4.149e-12 ***
Residuals    108 37.461   0.347
---
Signif. codes:  0 âÃŸ***âÃŹ 0.001 âÃŸ**âÃŹ 0.01 âÃŸ*âÃŹ 0.05 âÃŸ.â
```

Ergebnis

- Die Sonneneinstrahlung erklärt den Einfluss des Windes nicht.
- **Regel:** Um offensichtliche Scheinzusammenhänge auszuschließen, die sich durch bekannte Einflußfaktorenerklären lassen, sollte zur Kontrolle immer nochmal eine Modell überprüft werden, in dem der Einfluss zuletzt steht.
- **Regel:** Bei Beobachtungen ist durch den statistischen Nachweis eines Einflusses noch kein **kausaler** Zusammenhang nachgewiesen.

Varianzanalyse im Detail

Die Varianzanalyse kenne wir als:

Mehrstichprobentest, auf Gleichheit der Mittelwerte, bei Normalverteilung.

Wiederholung Varianzanalyse

Unser Blick auf die **Varianzanalyse** war:
Wir haben mehrere normalverteilte Stichproben mit gleicher Varianz:

$$
\begin{aligned}
&X_i,\ i = 1, \ldots, n_X, \quad X_i \sim N(\mu_X, \sigma^2) \\
&Y_i,\ i = 1, \ldots, n_Y, \quad Y_i \sim N(\mu_Y, \sigma^2) \\
&\vdots \\
&Z_i,\ i = 1, \ldots, n_Y, \quad Z_i \sim N(\mu_Z, \sigma^2)
\end{aligned}
$$

Die Varianzanalyse testet ob alle Mittelwerte gleich sind:

$H_0 : \mu_X = \mu_Y = \ldots = \mu_Z \quad vs. \quad H_1$: ganz oder teilweise verschie

Stichprobe als Kategorie

Ein anderer Blick auf die **Varianzanalyse** ist:

- Wir haben eine reelle Variable Y_i
- und eine kategorielle Variable $X_i \in \{1, \ldots, p\}$, die sagt aus welcher Gruppe/Stichprobe das statistische Individum i kommt

Wir verwenden dann die Bezeichnung:

$$\mu_k = \text{Mittelwert der Grundgesamtheit } k$$

Varianzanalyse als lineares Modell

Unsere Voraussetzung war dann,

$$Y_i \sim N(\mu_{X_i}, \sigma^2)$$

was man aber mit $\epsilon_i = Y_i - \mu_{X_i}$ schreiben kann als:

$$Y_i = a + \mu_{X_i} + \epsilon_i, \text{ mit } \epsilon_i \sim N(0, \sigma^2)$$

und (mit $a = 0$) das ist:

$$Y_i = a + \mu_1 f_1(X_i) + \mu_2 f_2(X_i) + \ldots + \epsilon_i$$

mit $f_k(X) = 1$ falls $X = k$ und 0 sonst.

Beispiel: Aquifer

```
> Aqui <- read.table("Aquifers.txt")
> Aqui$trans <- exp(Aqui$logT)
> Aqui

          logT Teufe  Type  trans
1  -3.5755508 78.64 Poren  0.028
2  -2.6172958 49.00 Poren  0.073
3  -2.2072749 47.00 Poren  0.110
4  -1.9379420 43.67 Poren  0.144
5  -1.7719568 37.00 Poren  0.170
6  -0.8209806 23.50 Poren  0.440
7   0.4700036  9.00 Poren  1.600
8   0.5877867 80.50 Kluft  1.800
9   1.4586150 21.25 Kluft  4.300
10  1.8196988 43.50 Kluft  6.170
11  2.5802168 29.50 Kluft 13.200
12  3.4657359 16.50 Kluft 32.000
```

Daten ansehen

```
> boxplot(logT ~ Type, data = Aqui)
```

```
> Hebelwirkung <- function(Modell) {
+     lm.influence(Modell)$hat
+ }
> Auswertung <- function(Formel, Daten, Name, robust = FALSE) {
+     Daten <<- Daten
+     if (robust)
+         Modell <- lmRob(formula = Formel, data = Daten)
+     else Modell <- lm(formula = Formel, data = Daten)
+     cat("R^2=", R2(Modell), "\n")
+     print(anova(Modell))
+     cat("Estimated Parameters of the model:\n")
+     print(coef(Modell))
+     if (robust)
+         par(mfrow = c(2, 2))
+     else par(mfrow = c(2, 3))
+     plot(Modell$fitted.values, resid(Modell), xlab = "Vorhersagen",
+         ylab = "Residuen")
+     title(Name, paste("R^2=", R2(Modell)))
+     fr <- model.frame(Formel, Daten)
+     if (is.factor(fr[[2]])) {
+         plot(fr[[2]], fr[[1]], xlab = names(fr)[2], ylab = names(fr
+             pch = 20)
+         points(fr[[2]], predict(Modell), pch = 20, col = "red")
+     }
+     else {
+         plot(fr[[2]], fr[[1]], xlab = names(fr)[2], ylab = names(fr
+             pch = 20)
+         points(fr[[2]], predict(Modell), pch = 20, col = "red")
+         segments(fr[[2]], predict(Modell), fr[[2]], fr[[1]],
+             col = "yellow")
+     }
+     qqnorm(resid(Modell), ylab = "Sample Quantiles of residuals")
+     qqline(resid(Modell))
+     title("", Name)
+     if (!robust) {
+         Hebelwirkungen <- Hebelwirkung(Modell)
+         plot(predict(Modell), Hebelwirkungen)
+         title("Hebelwirkung", Name)
+         plot(predict(Modell), cooks.distance(Modell))
```

```
+         title("Cook Distanzen", Name)
+         hist(cooks.distance(Modell), xlab = "Cooks Distance",
+             ylab = "Anzahl")
+         title("", Name)
+     }
+ }
```

Varianzanalyse des Gesteinseinfluss

```
> Auswertung(logT ~ Type, Aqui, "Modell 1")

R^2= 0.7438713
Analysis of Variance Table


Response: logT
          Df Sum Sq Mean Sq F value    Pr(>F)
Type       1 55.118  55.118  31.947 0.0001483 ***
Residuals 11 18.978   1.725
---
Signif. codes:  0 âÃŸ***âÃŹ 0.001 âÃŸ**âÃŹ 0.01 âÃŸ*âÃŹ 0.05 âÃŸ.â
Estimated Parameters of the model:
(Intercept)   TypePoren
   2.350285   -4.130427
```

Beobachtungen zur Varianzanalyse

- Der Computer hat den Type implizit als kategoriell erkannt und statt einer Regressionsanalyse eine Varianzanalyse gerechnet.

Beobachtungen zur Varianzanalyse

- Der Computer hat den Type implizit als kategoriell erkannt und statt einer Regressionsanalyse eine Varianzanalyse gerechnet.
- Es fehlt der Parameter "`TypeKluft`":

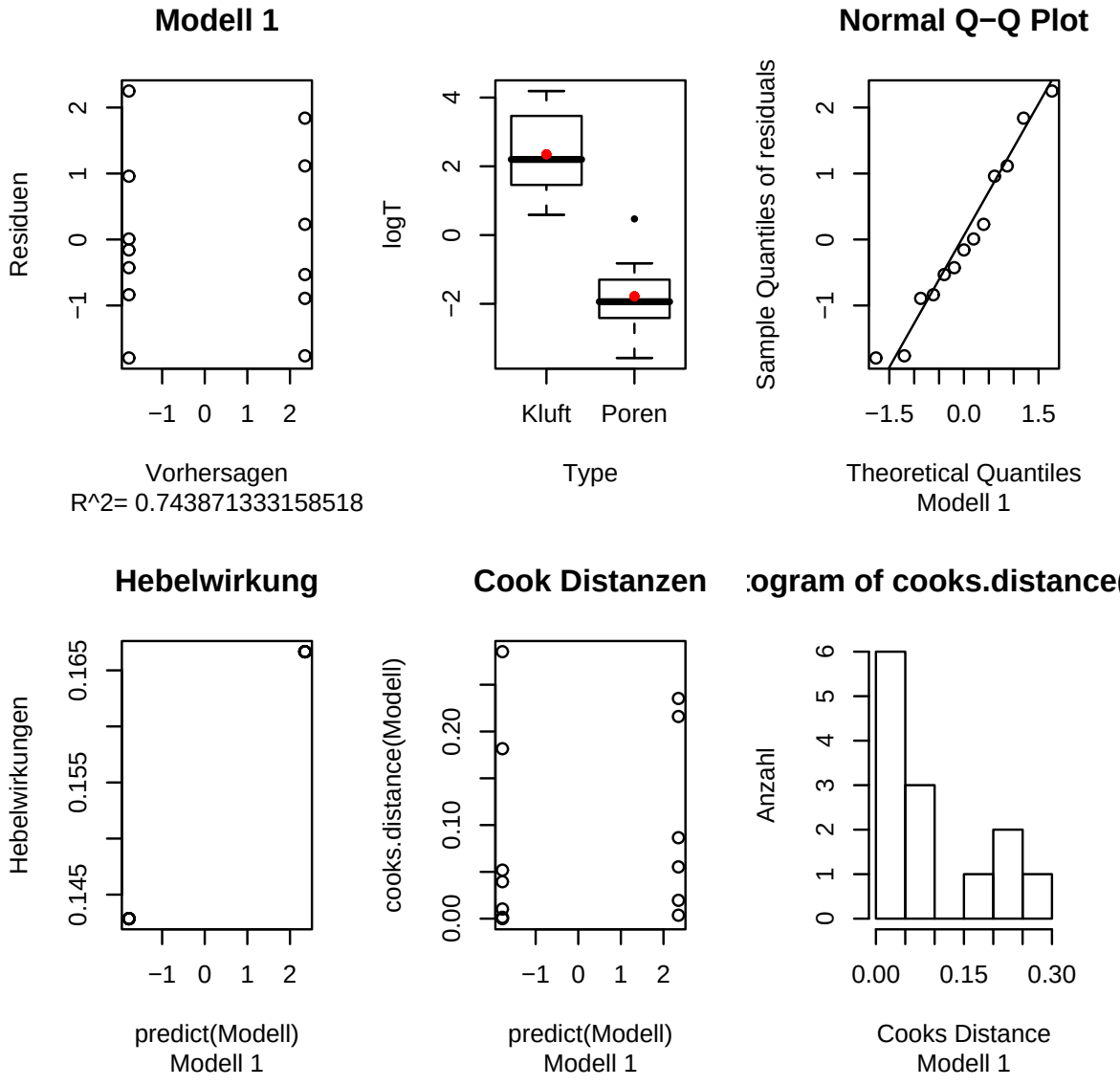
Beobachtungen zur Varianzanalyse

- Der Computer hat den Type implizit als kategoriell erkannt und statt einer Regressionsanalyse eine Varianzanalyse gerechnet.
- Es fehlt der Parameter "`TypeKluft`":
- Der Computer entfernt implizit den Einfluss der letzten Kategorie aus dem Modell:

$$\text{“}\texttt{logT}\text{”} = a + b f_1(\text{“}\texttt{Type}\text{”}) + \epsilon$$

Also: $\mu_{\text{Poren}} = a + b = 1.78$ und $\mu_{\text{Kluft}} = a + 0 = 2.35$

Diagnostik


Modell 1
Residuen
Vorhersagen
R^2= 0.743871333158518
logT
Kluft
Poren
Type
Normal Q−Q Plot
Sample Quantiles of residuals
Theoretical Quantiles
Modell 1
Hebelwirkung
Hebelwirkungen
predict(Modell)
Modell 1
Cook Distanzen
cooks.distance(Modell)
predict(Modell)
Modell 1
ogram of cooks.distance(
Anzahl
Cooks Distance
Modell 1

Allgemeines Lineares Modell

Regression des Teufeneinfluss

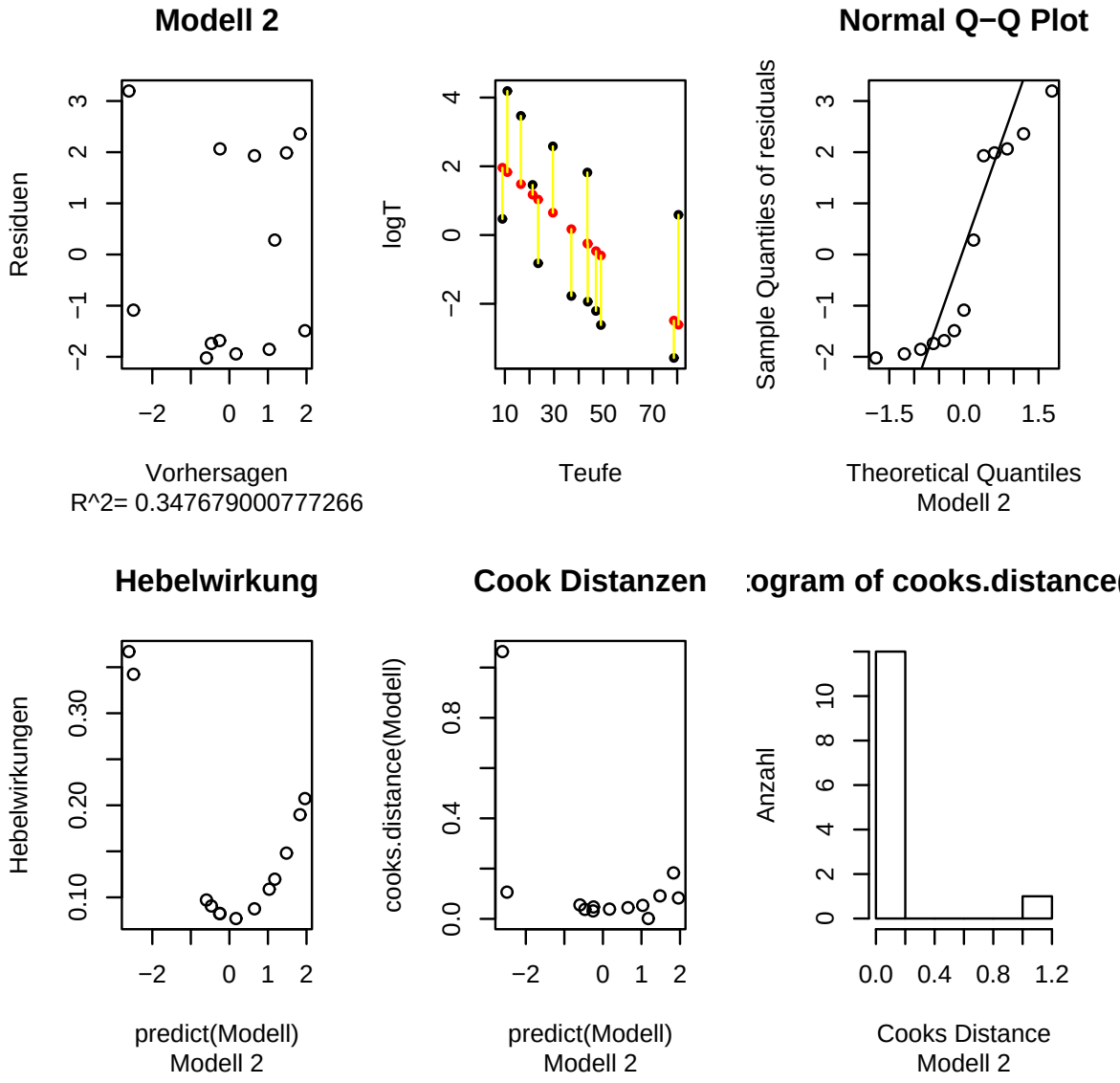
```
> Auswertung(logT ~ Teufe, Aqui, "Modell 2")

R^2= 0.347679
Analysis of Variance Table

Response: logT
          Df Sum Sq Mean Sq F value  Pr(>F)
Teufe      1 25.762  25.762  5.8629 0.03393 *
Residuals 11 48.335   4.394
---
Signif. codes:  0 âÃŸ***âÃŹ 0.001 âÃŸ**âÃŹ 0.01 âÃŸ*âÃŹ 0.05 âÃŸ.â
Estimated Parameters of the model:
(Intercept)       Teufe
 2.53348382 -0.06385867
```

Modell

`logT~Teufe`

$$\text{“}\texttt{logT}\text{”} = a + b \cdot \text{“}\texttt{Teufe}\text{”}_i + \epsilon_i$$

Diagnostik


Modell 2
Residuen
Vorhersagen
R^2= 0.347679000777266
logT
Teufe
Normal Q−Q Plot
Sample Quantiles of residuals
Theoretical Quantiles
Modell 2
Hebelwirkung
Hebelwirkungen
predict(Modell)
Modell 2
Cook Distanzen
cooks.distance(Modell)
predict(Modell)
Modell 2
ogram of cooks.distance(
Anzahl
Cooks Distance
Modell 2

Kombinierter Einfluss

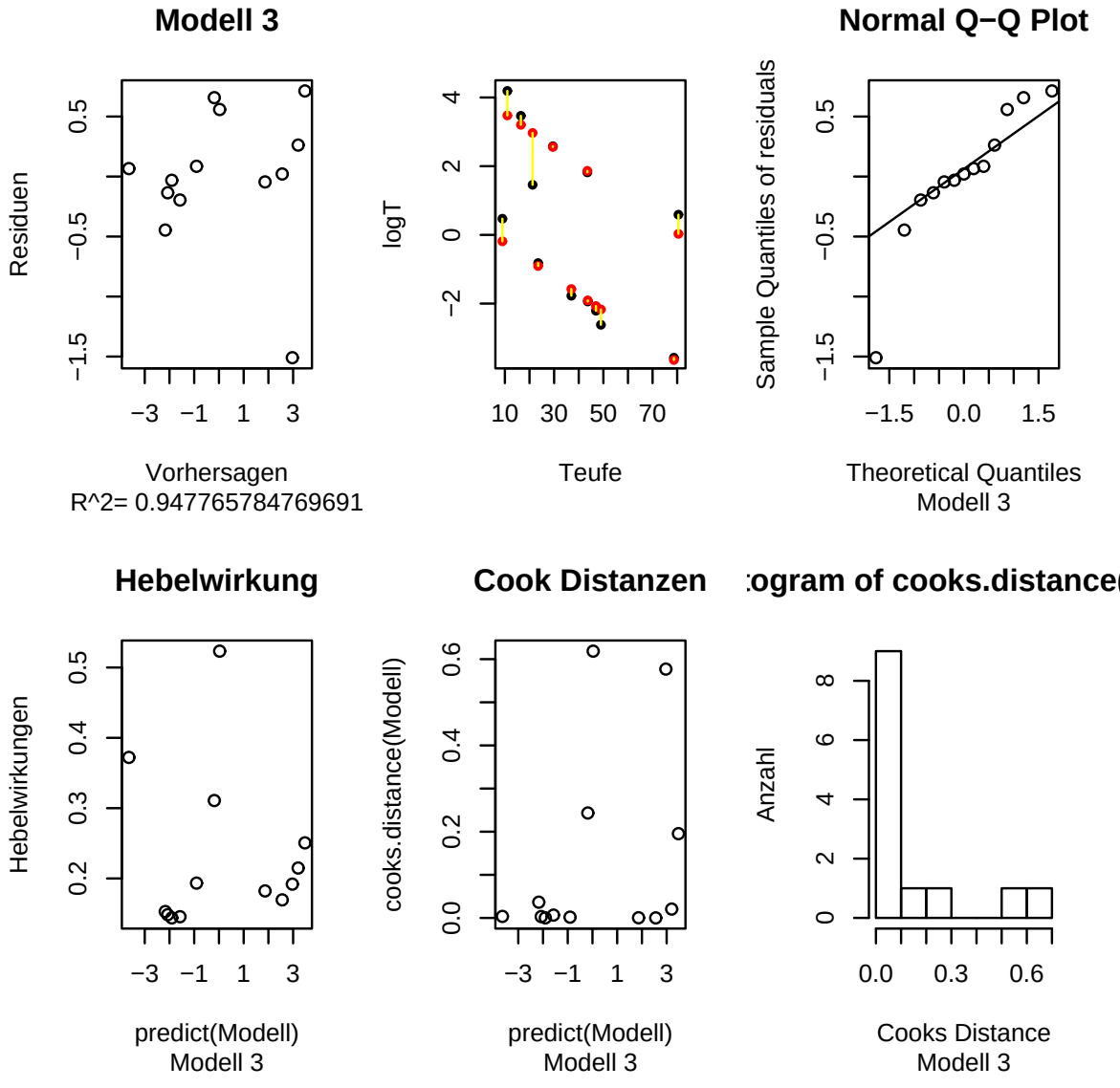
Wir können beide Einflüsse in das Modell aufnehmen:
`logT~Teufe+Type`

$$\text{“}\texttt{logT}\text{”} = a + b \cdot \text{“}\texttt{Teufe}\text{”}_i + c \cdot f_1(\text{“}\texttt{Type}\text{”}) + d \cdot f_2(\text{“}\texttt{Type}\text{”})\epsilon_i$$

Lineares Modell für logT

```
> Auswertung(logT ~ Teufe + Type, Aqui, "Modell 3")

R^2= 0.9477658
Analysis of Variance Table

Response: logT
          Df Sum Sq Mean Sq F value    Pr(>F)
Teufe      1 25.762  25.762  66.561 9.911e-06 ***
Type       1 44.464  44.464 114.884 8.387e-07 ***
Residuals 10  3.870   0.387
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.â
Estimated Parameters of the model:
(Intercept)       Teufe   TypePoren
 4.02234151 -0.04960366 -3.76299388
```

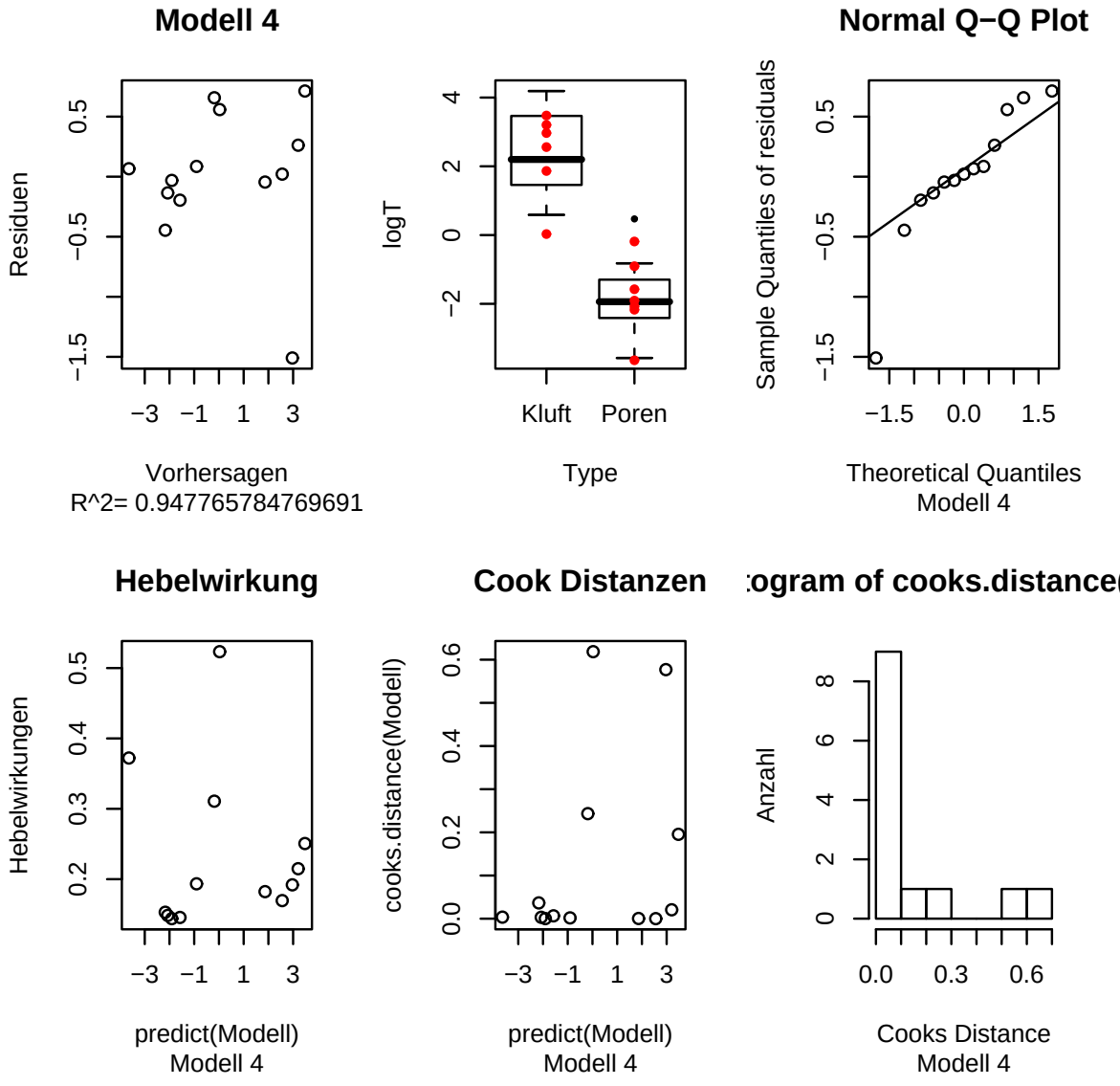
Diagnostik


Modell 3
Residuen
Vorhersagen
R^2= 0.947765784769691
logT
Teufe
Normal Q−Q Plot
Sample Quantiles of residuals
Theoretical Quantiles
Modell 3
Hebelwirkung
Hebelwirkungen
predict(Modell)
Modell 3
Cook Distanzen
cooks.distance(Modell)
predict(Modell)
Modell 3
ogram of cooks.distance(
Anzahl
Cooks Distance
Modell 3

. . . und in umgekehrter Reihenfolge

```
> Auswertung(logT ~ Type + Teufe, Aqui, "Modell 4")

R^2= 0.9477658
Analysis of Variance Table

Response: logT
          Df Sum Sq Mean Sq F value    Pr(>F)
Type       1 55.118  55.118 142.411 3.078e-07 ***
Teufe      1 15.108  15.108  39.035 9.533e-05 ***
Residuals 10  3.870   0.387
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.â
Estimated Parameters of the model:
(Intercept)   TypePoren       Teufe
 4.02234151 -3.76299388 -0.04960366
```

Diagnostik


Modell 4
Residuen
Vorhersagen
R^2= 0.947765784769691
logT
Kluft
Poren
Type
Normal Q−Q Plot
Sample Quantiles of residuals
Theoretical Quantiles
Modell 4
Hebelwirkung
Hebelwirkungen
predict(Modell)
Modell 4
Cook Distanzen
cooks.distance(Modell)
predict(Modell)
Modell 4
ogram of cooks.distance(
Anzahl
Cooks Distance
Modell 4

Ergebnis

- Die Leitfähigkeit des Aquifers hängt sowohl von der Teufe und als auch vom Gesteinstyp ab.
- Die Leitfähigkeit ist dadurch fast vollständig erklärt.

Diskret-Stetig Interaktion

Frage: Ist der Teufeneinfluss vielleicht unterschiedlich zwischen den Gesteinstypen?
Wir fügen eine Interaktion zum Modell hinzu:

$$\text{“}\texttt{logT}\text{”} = \ldots + e \cdot \text{“}\texttt{Teufe}\text{”} \cdot f_1(\text{“}\texttt{Type}\text{”}) + f \cdot \text{“}\texttt{Teufe}\text{”} \cdot f_2(\text{“}\texttt{Type}\text{”}) + \epsilon$$

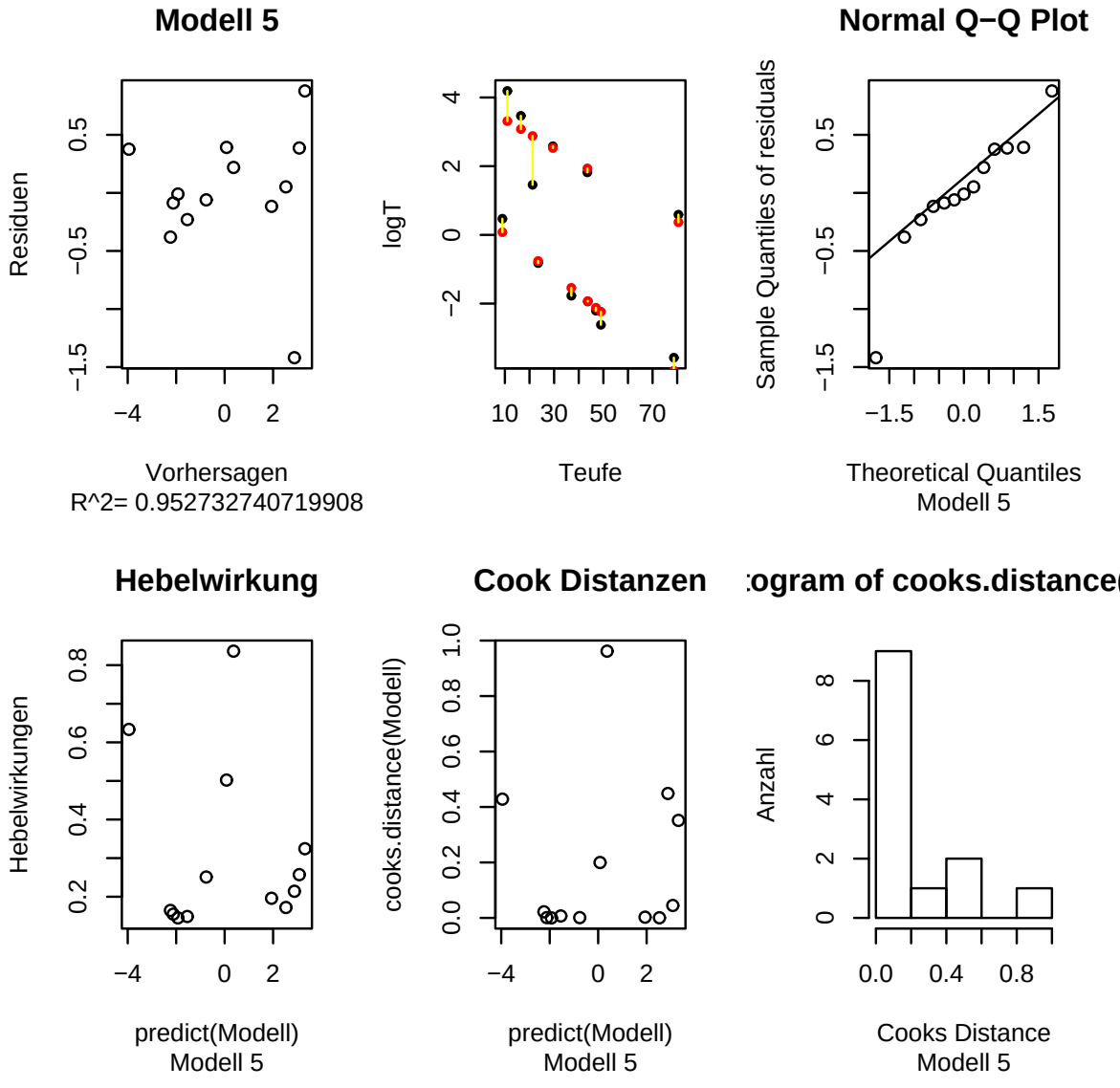
. . . und mit Interaktion

```
> Auswertung(logT ~ Teufe * Type, Aqui, "Modell 5")

R^2= 0.9527327
Analysis of Variance Table


Response: logT
            Df Sum Sq Mean Sq  F value    Pr(>F)
Teufe        1 25.762  25.762  66.2004 1.933e-05 ***
Type         1 44.464  44.464 114.2605 2.049e-06 ***
Teufe:Type   1  0.368   0.368   0.9457    0.3562
Residuals    9  3.502   0.389
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.â
Estimated Parameters of the model:
    (Intercept)           Teufe       TypePoren Teufe:TypePoren
     3.77785142     -0.04235056     -3.17872499     -0.01551707
```

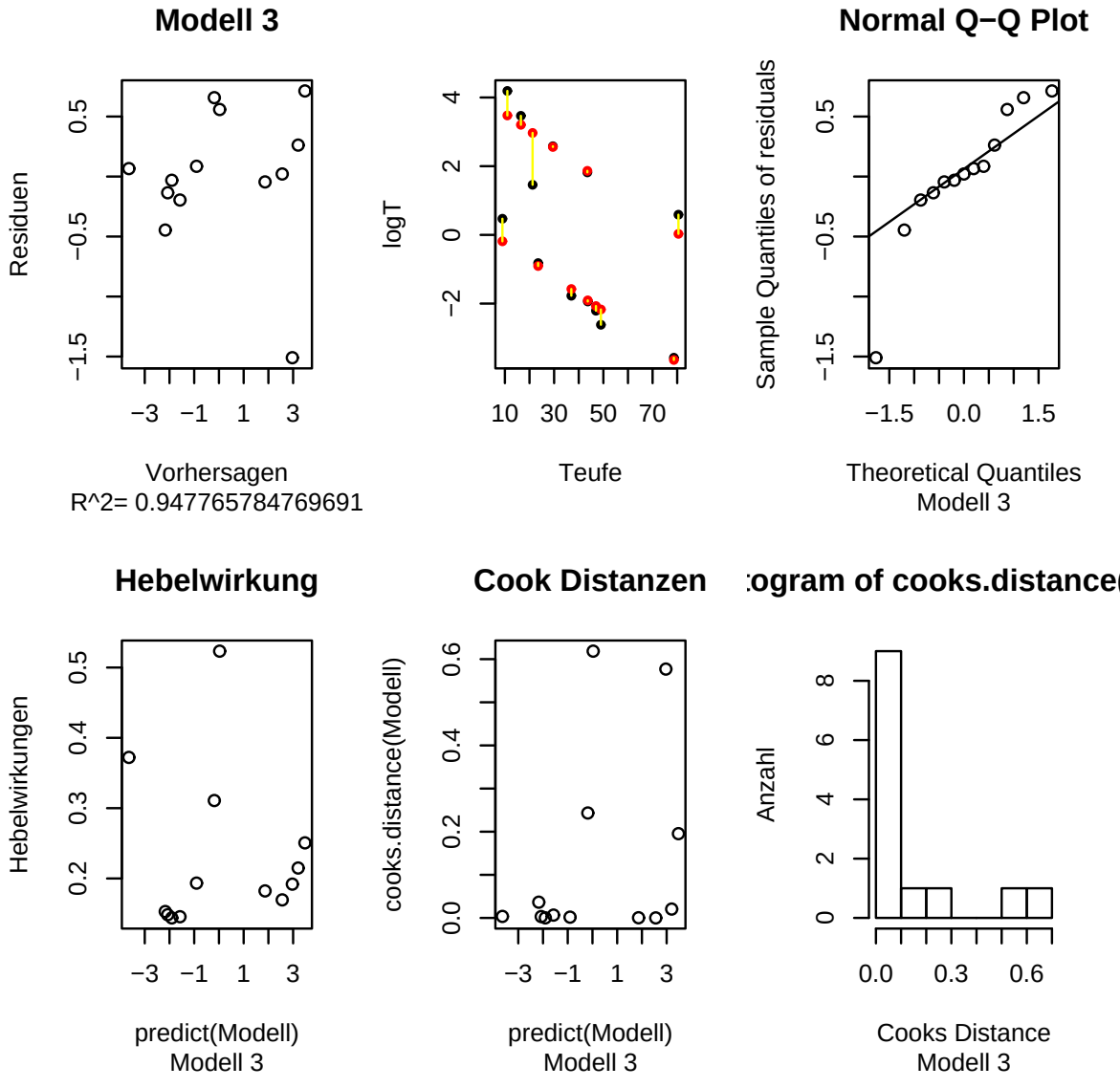
Ergebnis

Interaktion wird nicht benötigt.

Diagnostik mit Interaktion



Diagnostik ohne Interaktion



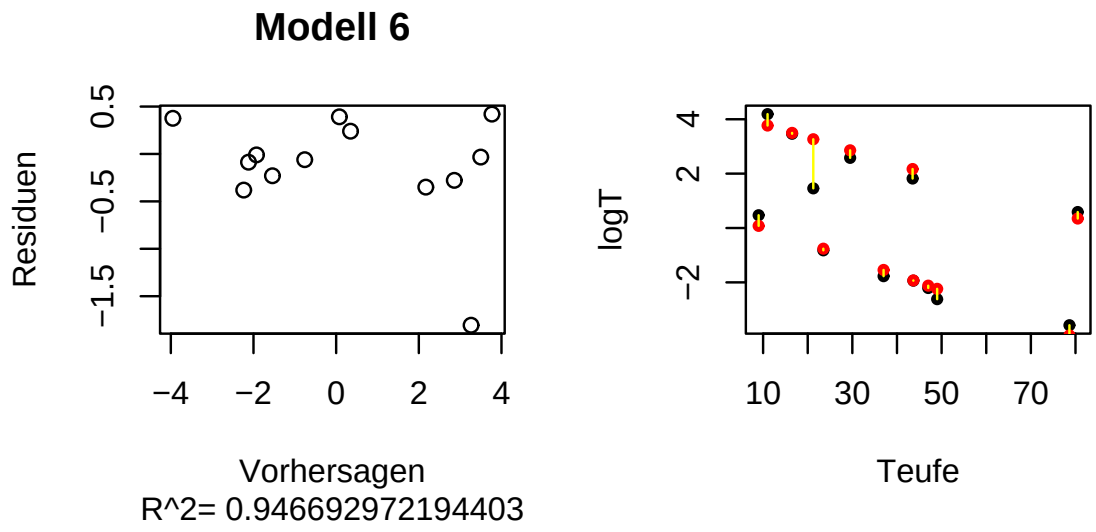
Robuste Schätzung

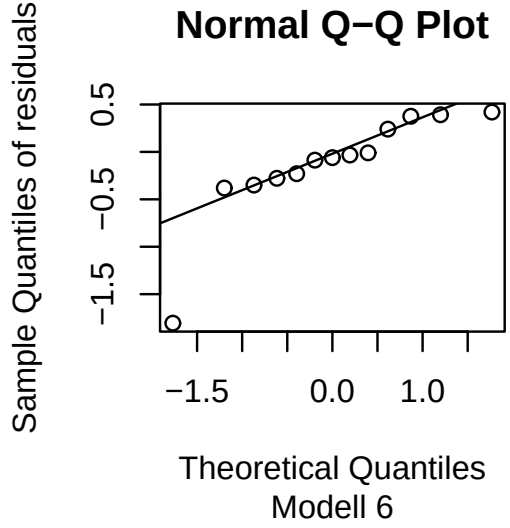
Die Robuste Methodik kann auch hier angewendet werden.

. . . und mit robuster Schätzung

```
> Auswertung(logT ~ Teufe * Type, Aqui, "Modell 6", robust = TRUE)

R^2= 0.946693

Terms added sequentially (first to last)

             Chisq Df RobustF     Pr(F)
(Intercept)      1
Teufe            1   3.944  0.042992 *
Type             1  51.032 3.348e-13 ***
Teufe:Type       1   6.851  0.007648 **
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.â
Estimated Parameters of the model:
    (Intercept)           Teufe       TypePoren Teufe:TypePoren
    4.310785926   -0.049231142   -3.711659491   -0.008636487
```

Diagnostik


Modell 6
Residuen
0.5
−0.5
−1.5
−4
−2
0
2
4
Vorhersagen
R^2= 0.946692972194403
logT
4
2
−2
10
30
50
70
Teufe


Normal Q−Q Plot
Sample Quantiles of residuals
0.5
−0.5
−1.5
−1.5
0.0
1.0
Theoretical Quantiles
Modell 6