Statistik

Vorlesung 8 (Lineare Modelle)

K.Gerald van den Boogaart
http://www.stat.boogaart.de

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:

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

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:

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

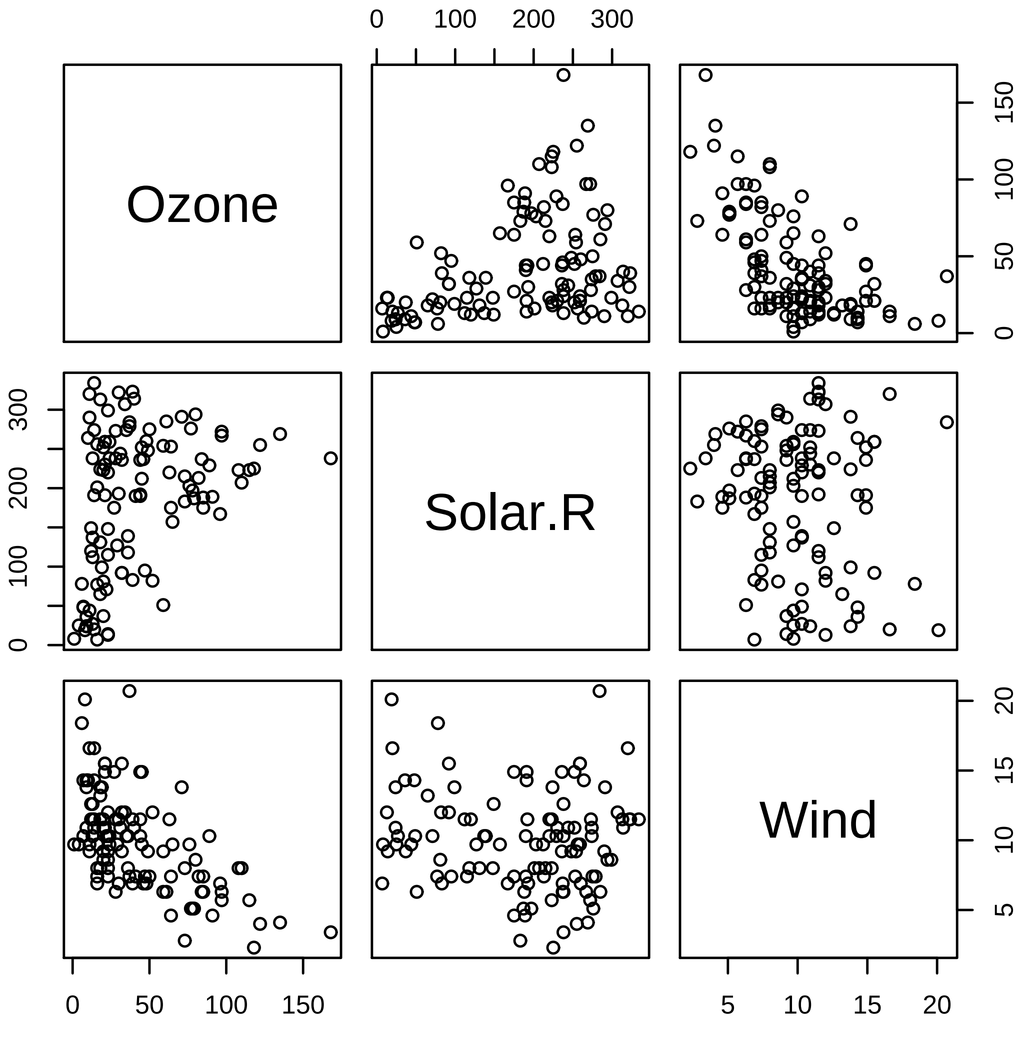
$\epsilon \sim N(0, \sigma^2)$,
als Lineares Modell:

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

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

$$f_1(X) = X$$

Beispiel: Ozon (städtisch)



Beobachtungen

- Sonnenstrahlung erhöht das Ozon
- Wind reduziert das Ozon

Einstrahlung: Parameter

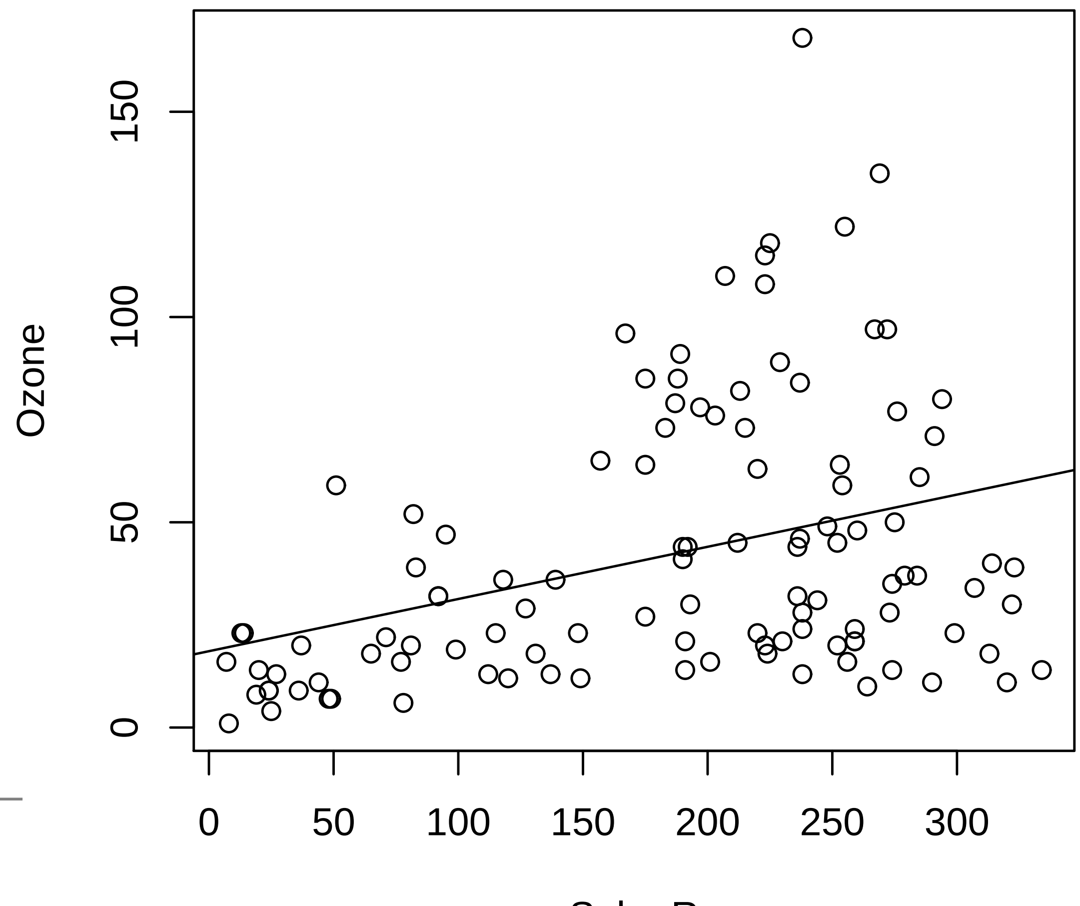
```
> mod <- lm(Ozone ~ Solar.R, data = ozon)
> mod

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


Coefficients:
(Intercept)      Solar.R
     18.599        0.127
```

Einstrahlung: Varianzanalysetabelle

```
> anova(mod)

Analysis of Variance Table


Response: Ozone
           Df Sum Sq Mean Sq F value  Pr(>F)
Solar.R     1  14780   14780    15.1 0.00018 ***
Residuals 109 107022     982
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ

> R2(mod)

[1] 0.1213
```

Einstrahlung: Regressionsgerade

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



Wind: Parameter

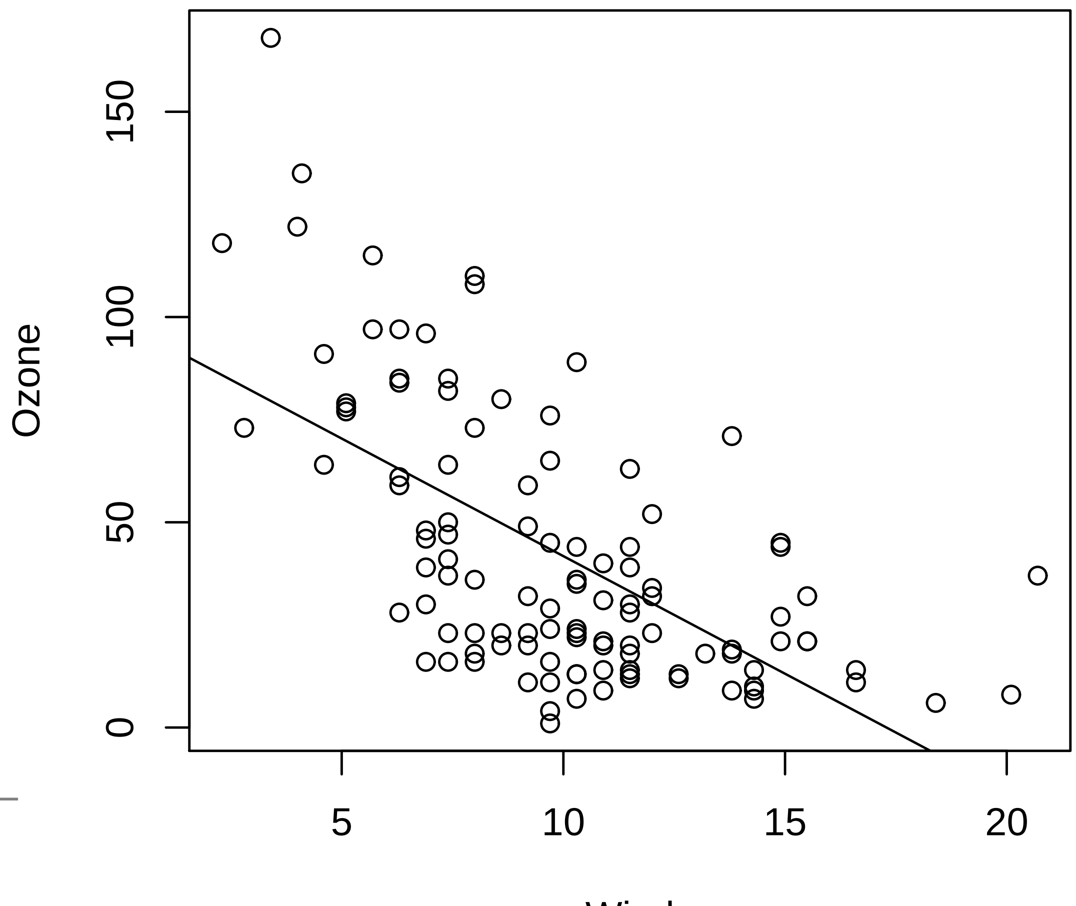
```
> mod <- lm(Ozone ~ Wind, data = ozon)
> mod

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


Coefficients:
(Intercept)         Wind
      99.04        -5.73
```

Wind: Tests

```
> anova(mod)

Analysis of Variance Table


Response: Ozone
           Df Sum Sq Mean Sq F value Pr(>F)
Wind        1  45694   45694    65.4  9e-13 ***
Residuals 109  76108     698
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ

> R2(mod)

[1] 0.3752
```

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:

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

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

$$\begin{aligned} f_1\left(\left(\begin{array}{c}\text{“}\texttt{Wind}\text{”}\\ \text{“}\texttt{Solar.R}\text{”}\end{array}\right)\right) &= \text{“}\texttt{Wind}\text{”} \\ f_2\left(\left(\begin{array}{c}\text{“}\texttt{Wind}\text{”}\\ \text{“}\texttt{Solar.R}\text{”}\end{array}\right)\right) &= \text{“}\texttt{Solar.R}\text{”} \end{aligned}$$

Parameter

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

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

Coefficients:
(Intercept)         Wind      Solar.R
     77.246       -5.402        0.100
```

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

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    73.6 7.7e-14 ***
Solar.R     1   9055    9055    14.6 0.00022 ***
Residuals 108  67053     621
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
```

Die Anova-Tests

Diese ANOVA Tabelle enthält zwei Tests:

$$
\begin{array}{rl}
 & H_0 : Y_i = a + \epsilon_i \\
\text{Test Wind} & vs. \\
 & H_1 : Y_i = a + b \cdot \text{“}\mathtt{Wind}\text{”}_i + \epsilon_i \\
\text{Test Solar.R} & vs. \\
 & H_2 : Y_i = a + b \cdot \text{“}\mathtt{Wind}\text{”}_i + c \cdot \text{“}\mathtt{Solar.R}\text{”}_i + \epsilon_i
\end{array}
$$

In jeder Zeile der Tabelle wird ein Test durchgeführt.

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    73.6 7.7e-14 ***
Solar.R     1   9055    9055    14.6 0.00022 ***
Residuals 108  67053     621
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
```

Bei Effekte sind also vorläufig nachgewiesen.

Voraussetzungen

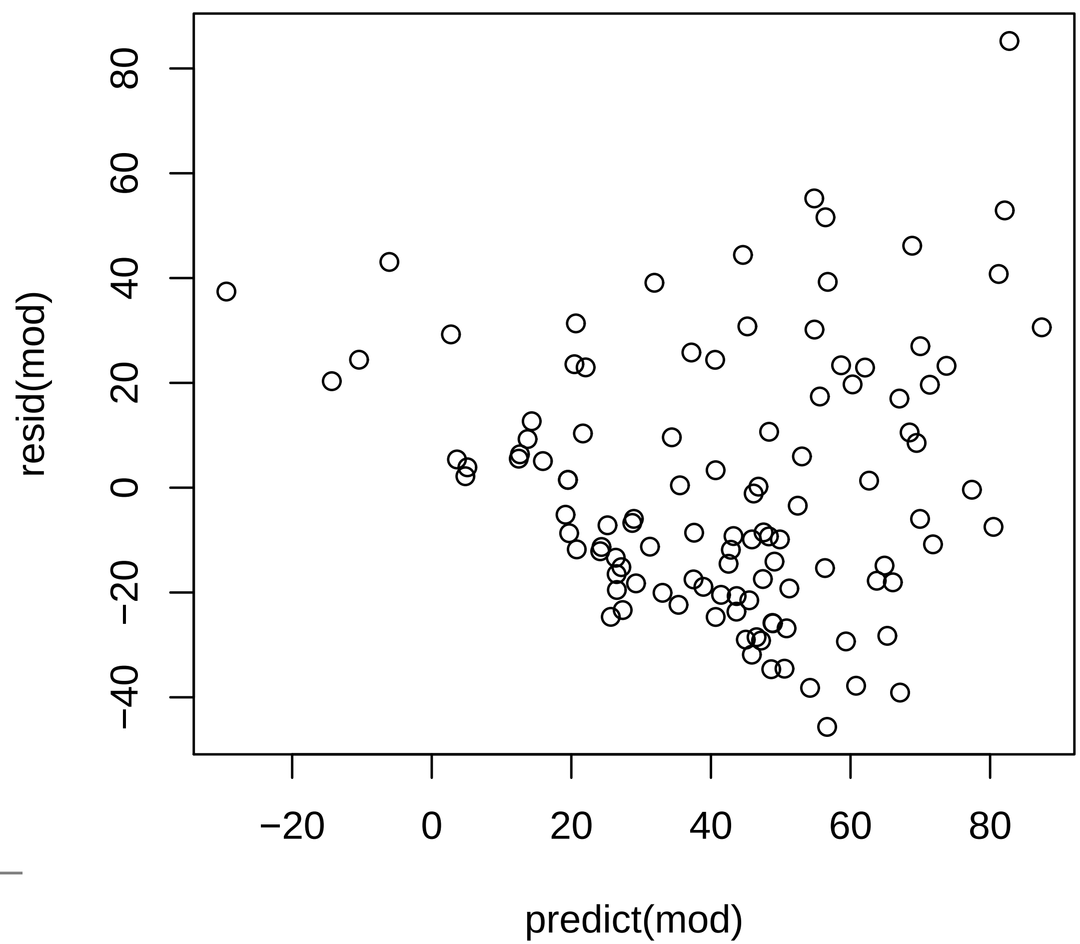
$$Y_i = \text{Formel} + \epsilon_i$$

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)
       3.694         -1.128         0.448
```

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

Varianzanalysetabelle

```
> anova(modL)

Analysis of Variance Table


Response: log(Ozone)
              Df Sum Sq Mean Sq F value  Pr(>F)
log(Wind)      1   29.0    29.0    83.5 4.2e-15 ***
log(Solar.R)   1   16.0    16.0    46.2 6.0e-10 ***
Residuals    108   37.5     0.3
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
```

Vergleich der Erklärungskraft

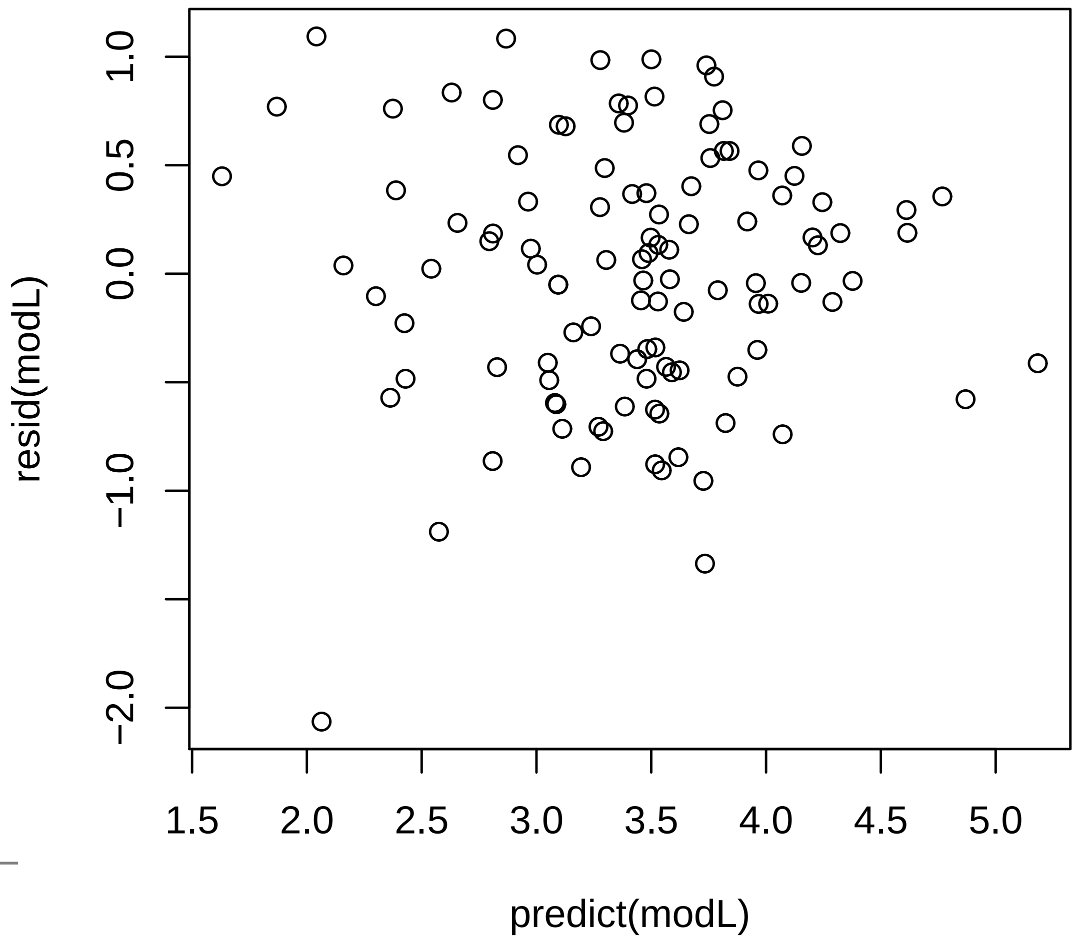
```
> R2(mod)
[1] 0.4495
> R2(modL)
[1] 0.5458
```

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))
```

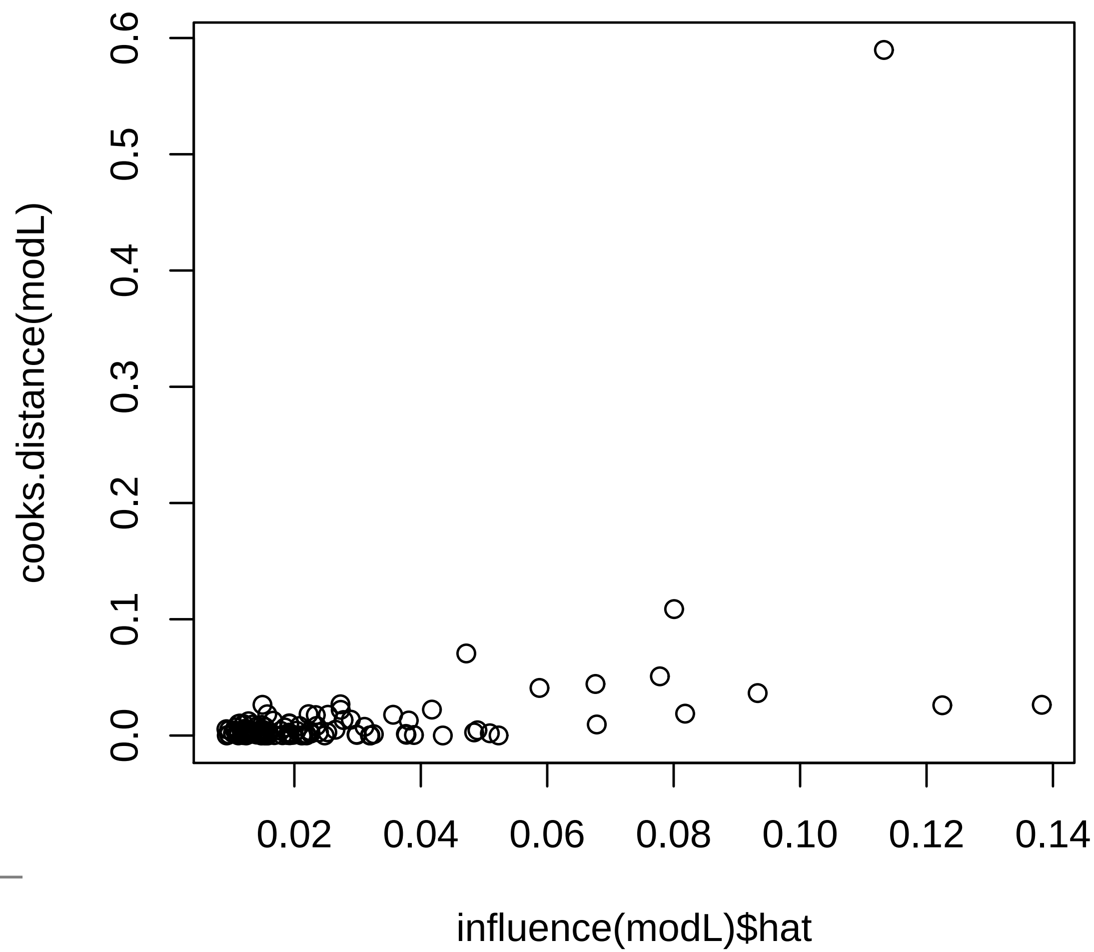


Interpretation

- Nun homoskedastisch.
- Ein Ausreißer ist deutlich sichtbar.

Hebelwirkungen

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



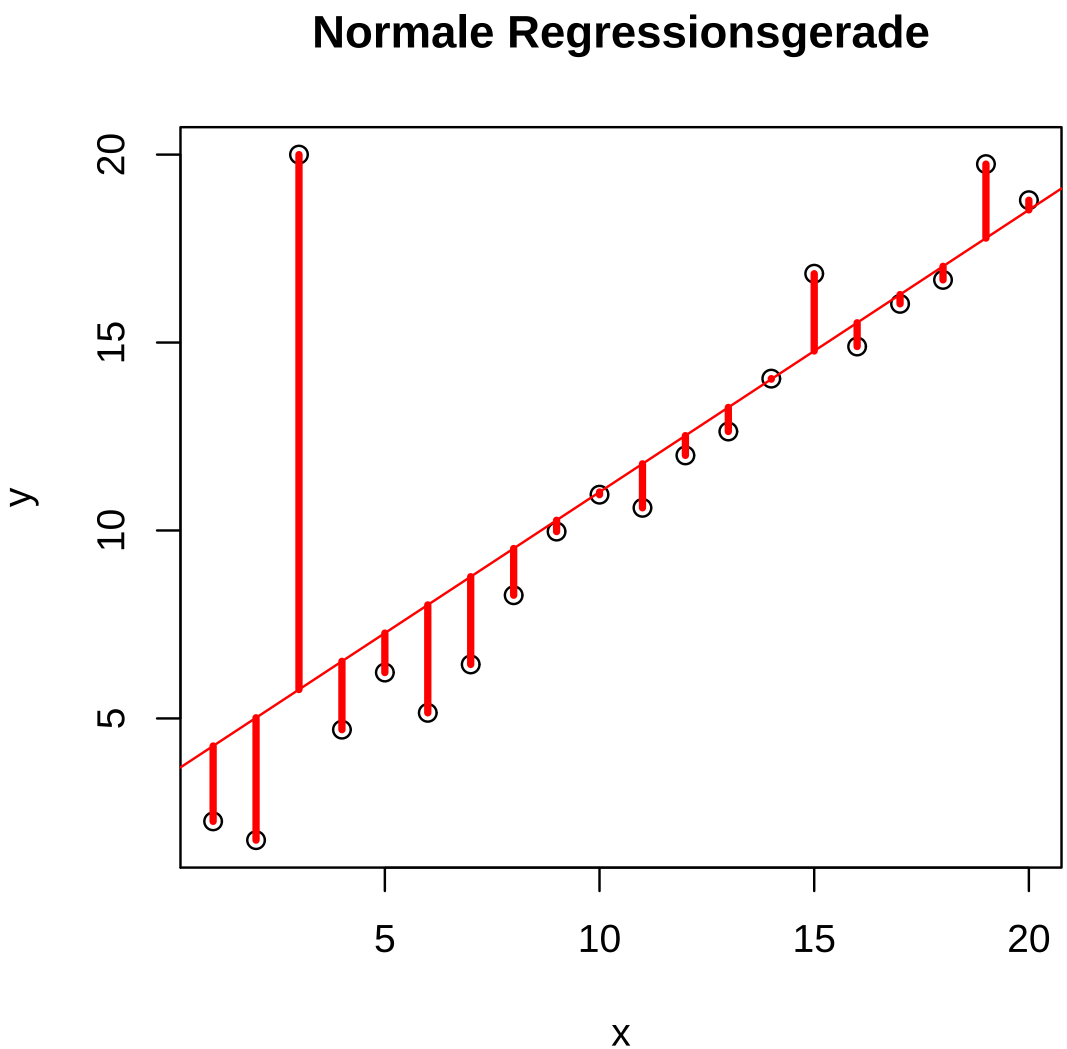
Möglichkeiten bei Ausreißern

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

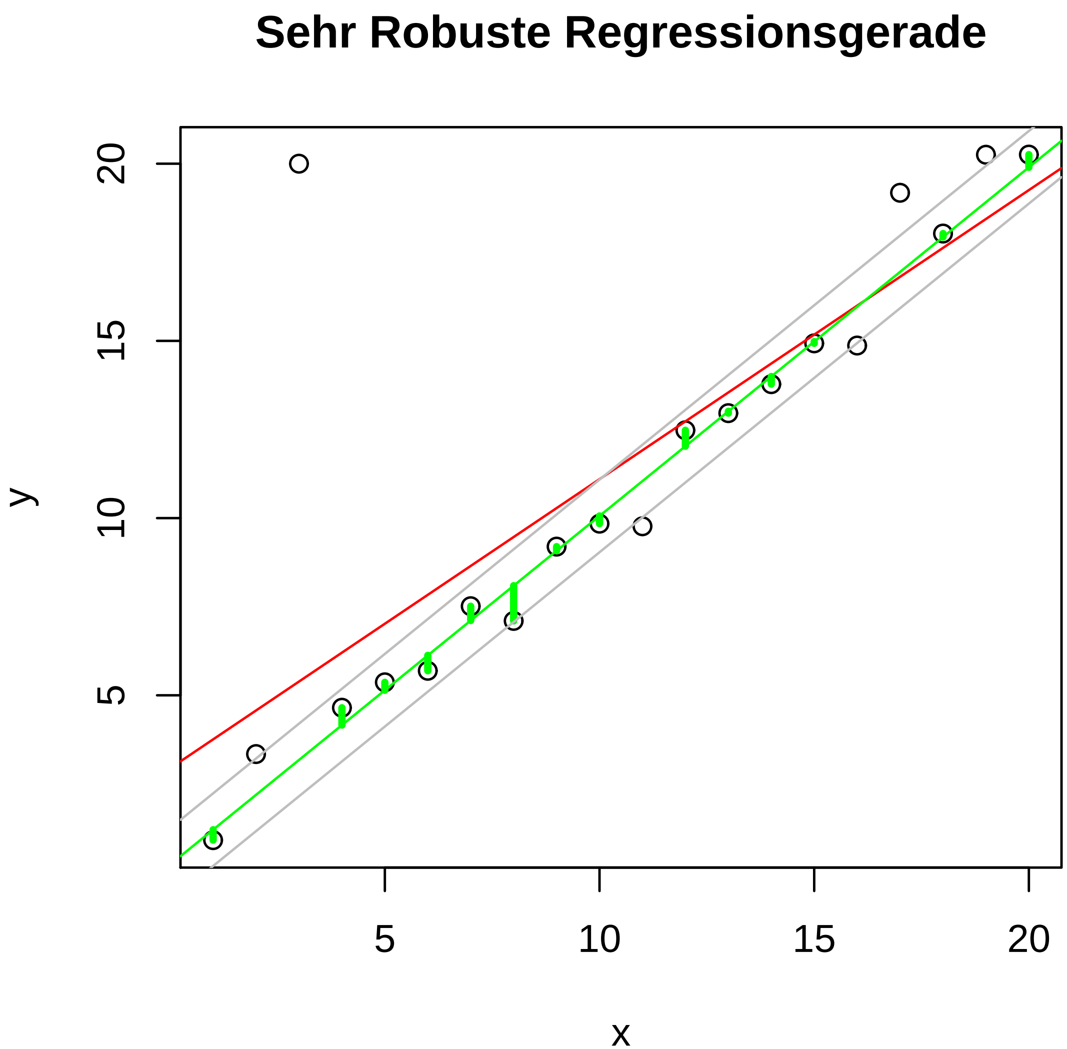
Robuste Schätzung

```
> require(robust)
> modLR <- lmRob(log(Ozone) ~ log(Wind) + log(Solar.R),
+       data = ozon)
> modLR

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

Coefficients:
 (Intercept)    log(Wind) log(Solar.R)
       4.244       -1.156        0.356

Degrees of freedom: 111 total; 108 residual
Residual standard error: 0.6062

> R2(modLR)

[1] 0.5375
```

Prinzip der robusten Schätzung



Prinzip der robusten Schätzung

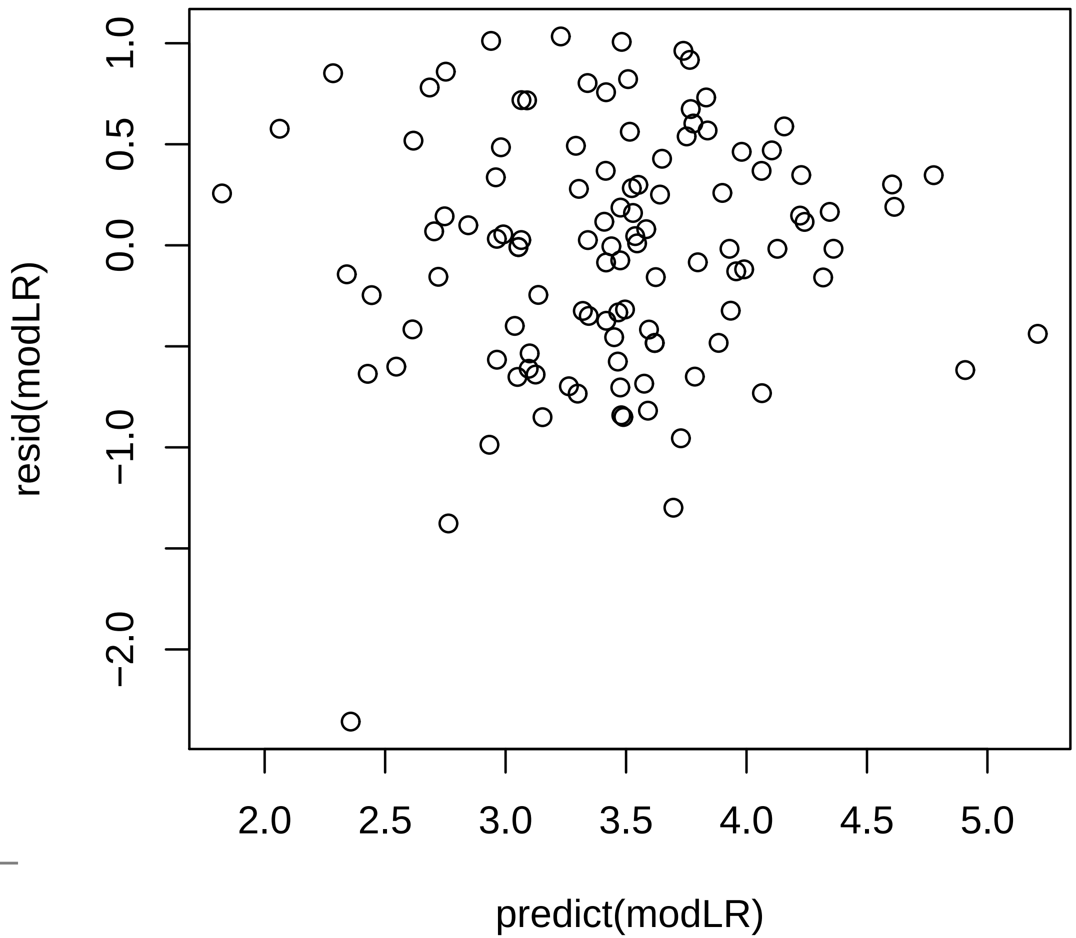


Varianzanalysetabelle

```
> anova(modLR)

Terms added sequentially (first to last)

              Chisq Df RobustF   Pr(F)
(Intercept)           1
log(Wind)             1    63.8 4.4e-16 ***
log(Solar.R)          1    20.5 3.9e-06 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
```

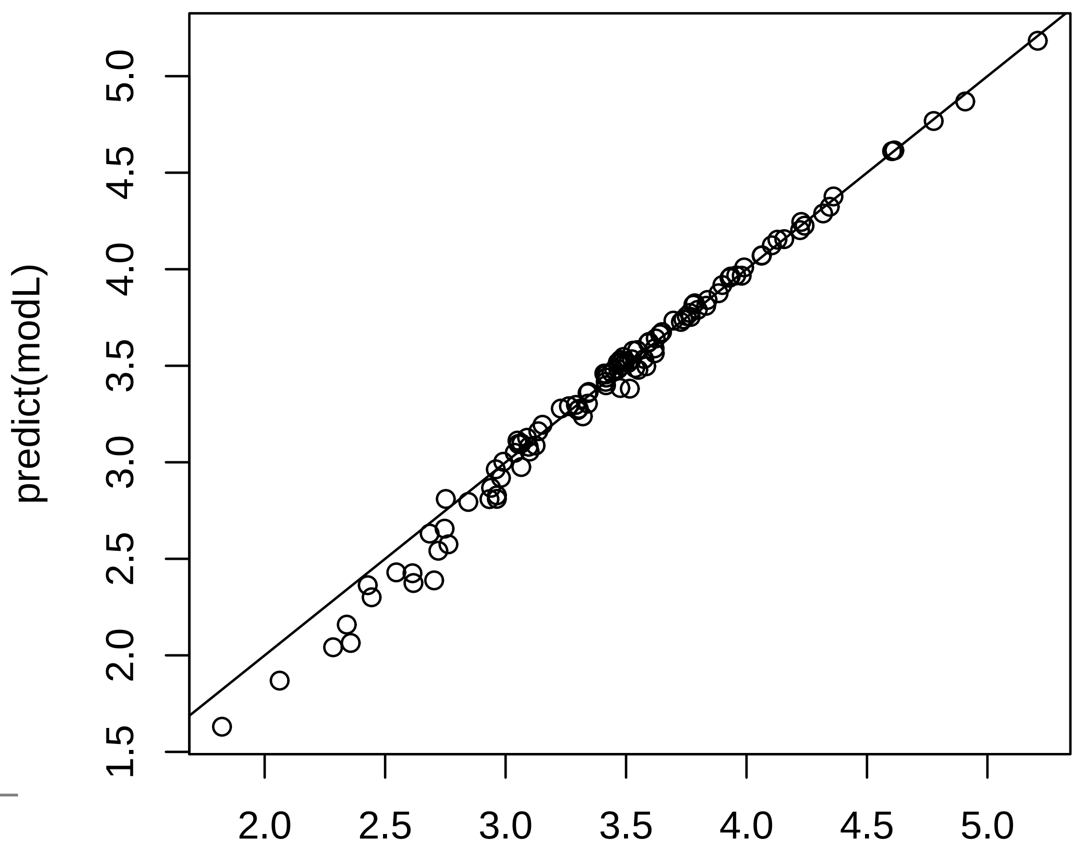
Residuals vs. Predicted

```
> 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] 0.4495
> R2(modL)
[1] 0.5458
> R2(modLR)
[1] 0.5375
```

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:

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

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

- e^x auf beide Seiten angewendet:

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

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

- und als multiplikatives Modell umformuliert:

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

  $\epsilon_i \sim LogN(0, 0.6061698)$. (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.
- 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)
                1.365                     -0.138
         log(Solar.R)  log(Wind):log(Solar.R)
                0.900                     -0.192
```

...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)
                 1.365                   -0.138
          log(Solar.R)  log(Wind):log(Solar.R)
                 0.900                   -0.192
```

R^2 vergleichen

```
> R2(mod)
[1] 0.4495
> R2(modL)
[1] 0.5458
> R2(modLI)
[1] 0.5495
```

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   29.0    29.0   83.43 4.6e-15
log(Solar.R)              1   16.0    16.0   46.19 6.3e-10
log(Wind):log(Solar.R)    1    0.3     0.3    0.88    0.35
Residuals               107   37.2     0.3

log(Wind)               ***
log(Solar.R)            ***
log(Wind):log(Solar.R)
Residuals
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
```

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

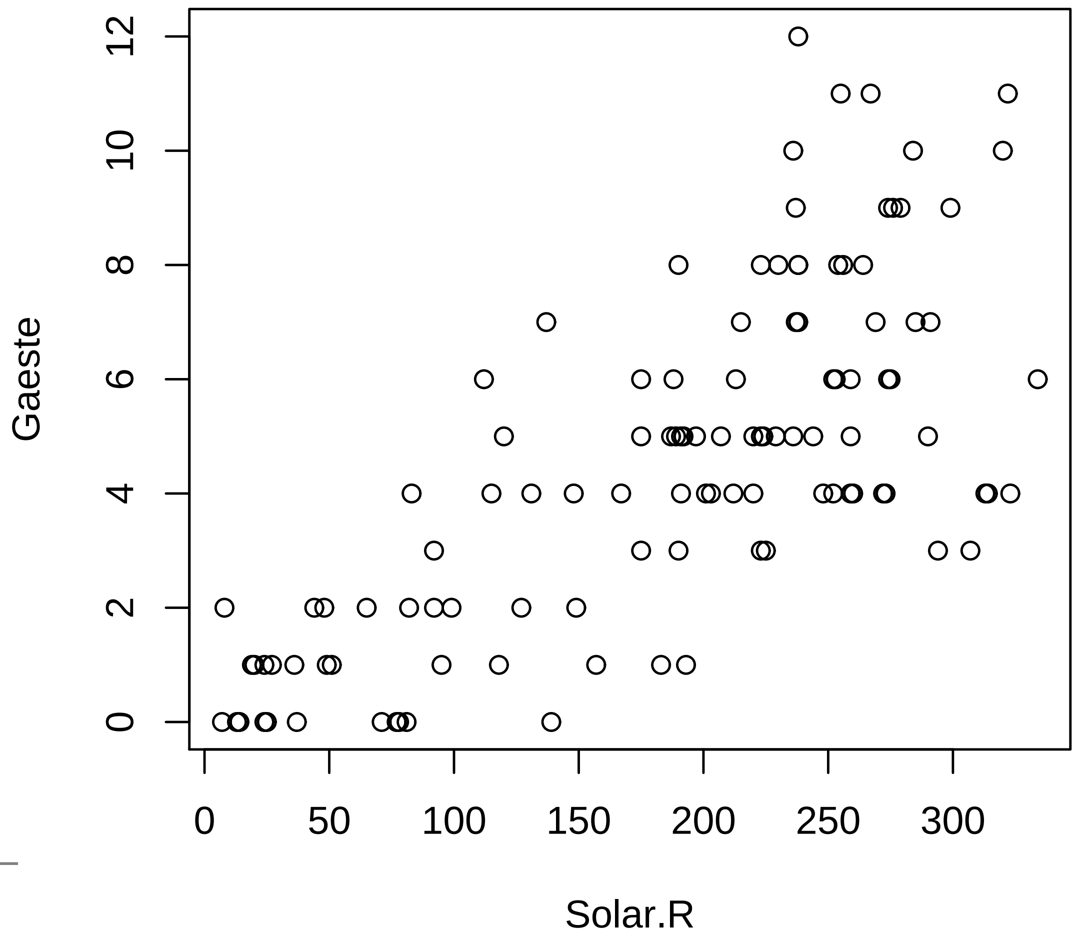
Problem: Scheinzusammenhang

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  5  1  2  5  2  0  3 11  2  1  9  5  3  6  1  0  9
 [19]  0  2  0  6  6  6  1  8  7  3  6  3  0  4  4 11  4 10
 [37]  2  7  6  7  7  6  4  6  2  6  8 11  5  7  1  6  6  0
 [55]  1  7  6  5  5  2  0  2  7  6  6  4  7  1  1  0  4 10
 [73]  5  7  2 11  7  3  7  4  3  3  3  4  7  4  1  6  3  4
 [91]  7  8  7  9  8  1  3  4  2  0  7  1  1  1  4  0  1  7
[109]  2  1  5
```

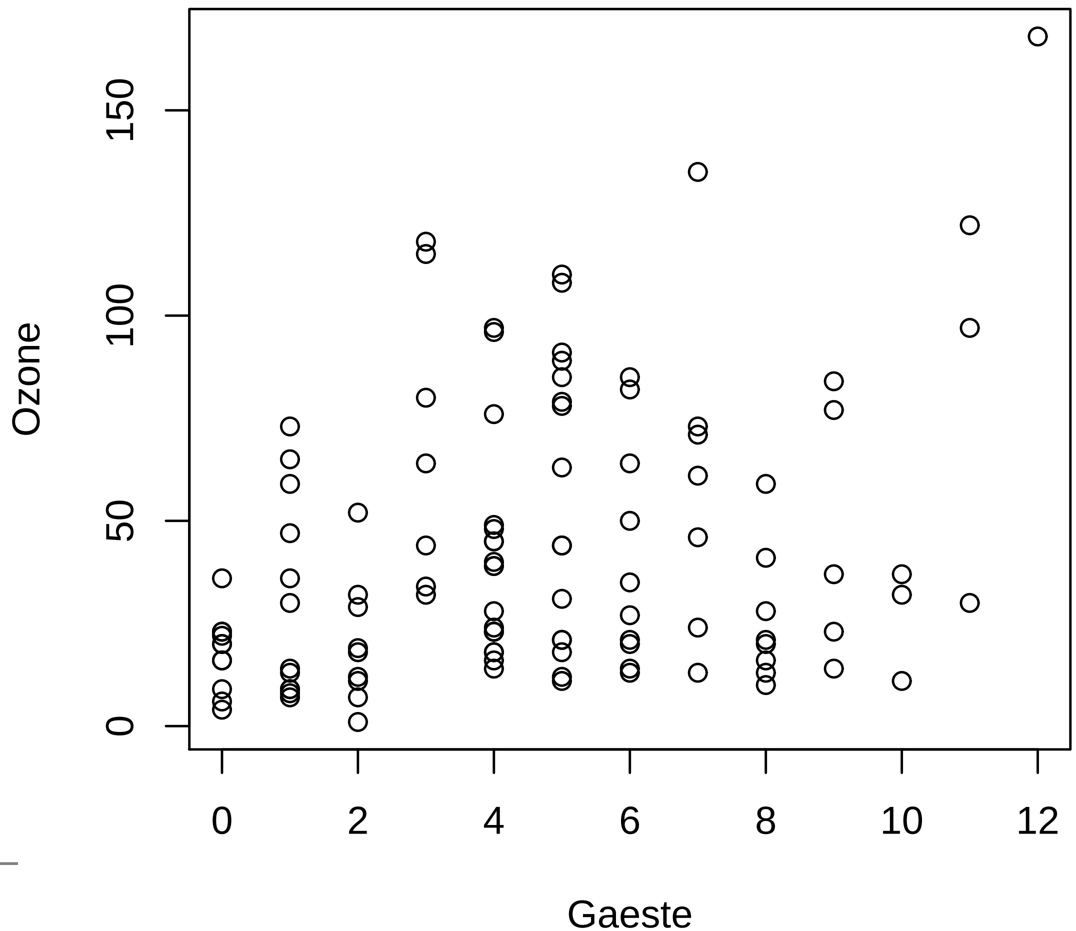
Sonne und Gaeste

```
> 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
      26.29         3.51
```

Varianzanalysetabelle

```
> anova(modXX)

Analysis of Variance Table


Response: Ozone
           Df Sum Sq Mean Sq F value  Pr(>F)
Gaeste      1  12190   12190    12.1 0.00072 ***
Residuals 109 109612    1006
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
```

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    8.5     8.5    14.1 0.00028 ***
Solar.R     1    8.6     8.6    14.3 0.00026 ***
Residuals 108   65.3     0.6
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
```

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    8.5     8.5    14.1 0.00028 ***
Solar.R     1    8.6     8.6    14.3 0.00026 ***
Residuals 108   65.3     0.6
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
```

Die Anova-Tests

Diese ANOVA Tabelle enthält zwei Tests:

$$
\begin{array}{ll}
 & H_0 : Y_i = a + \epsilon_i \\
\text{Test Gaeste} & vs. \\
 & H_1 : Y_i = a + b \cdot \text{“}\texttt{Gaeste}\text{”}_i + \epsilon_i \\
\text{Test Solar.R} & vs. \\
 & H_2 : Y_i = a + b \cdot \text{“}\texttt{Gaeste}\text{”} +_i + c \cdot \text{“}\texttt{Solar.R}\text{”}_i + \epsilon_i
\end{array}
$$

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   17.2    17.2   28.38 5.5e-07 ***
Gaeste      1  0.017   0.017    0.03    0.87
Residuals 108   65.3     0.6
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
```

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   29.0    29.0    83.5 4.2e-15 ***
log(Solar.R)   1   16.0    16.0    46.2 6.0e-10 ***
Residuals    108   37.5     0.3
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
```

. . . 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   23.9    23.9    68.9 3.2e-13 ***
log(Wind)      1   21.1    21.1    60.9 4.1e-12 ***
Residuals    108   37.5     0.3
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
```

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:

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

Die Varianzanalyse testet ob alle Mittelwerte gleich sind:

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

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:

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

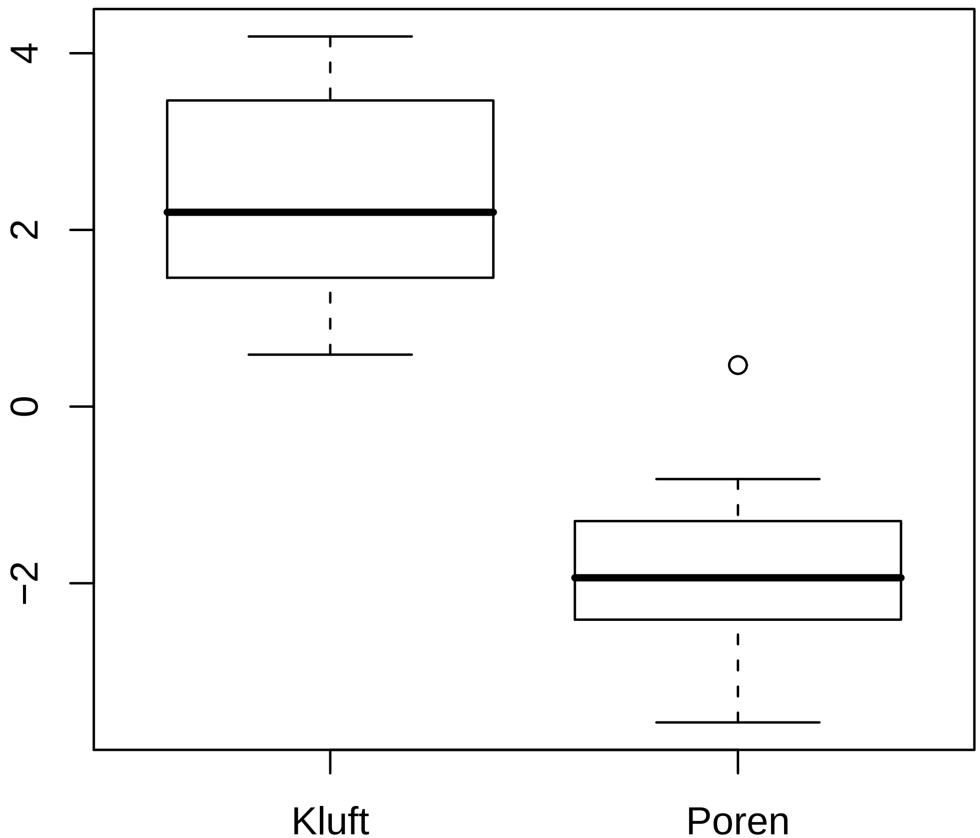
mit $f_k(X) = 1$ falls $X = k$ und 0 sonst.

Beispiel: Aquifer

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

      logT Teufe  Type  trans
1  -3.5756 78.64 Poren  0.028
2  -2.6173 49.00 Poren  0.073
3  -2.2073 47.00 Poren  0.110
4  -1.9379 43.67 Poren  0.144
5  -1.7720 37.00 Poren  0.170
6  -0.8210 23.50 Poren  0.440
7   0.4700  9.00 Poren  1.600
8   0.5878 80.50 Kluft  1.800
9   1.4586 21.25 Kluft  4.300
10  1.8197 43.50 Kluft  6.170
11  2.5802 29.50 Kluft 13.200
12  3.4657 16.50 Kluft 32.000
```

Daten ansehen

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



Varianzanalyse des Gesteinseinfluss

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

R^2= 0.7439
Analysis of Variance Table

Response: logT
          Df Sum Sq Mean Sq F value  Pr(>F)
Type       1   55.1    55.1    31.9 0.00015 ***
Residuals 11   19.0     1.7
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
Estimated Parameters of the model:
(Intercept)   TypePoren
      2.350      -4.130
```

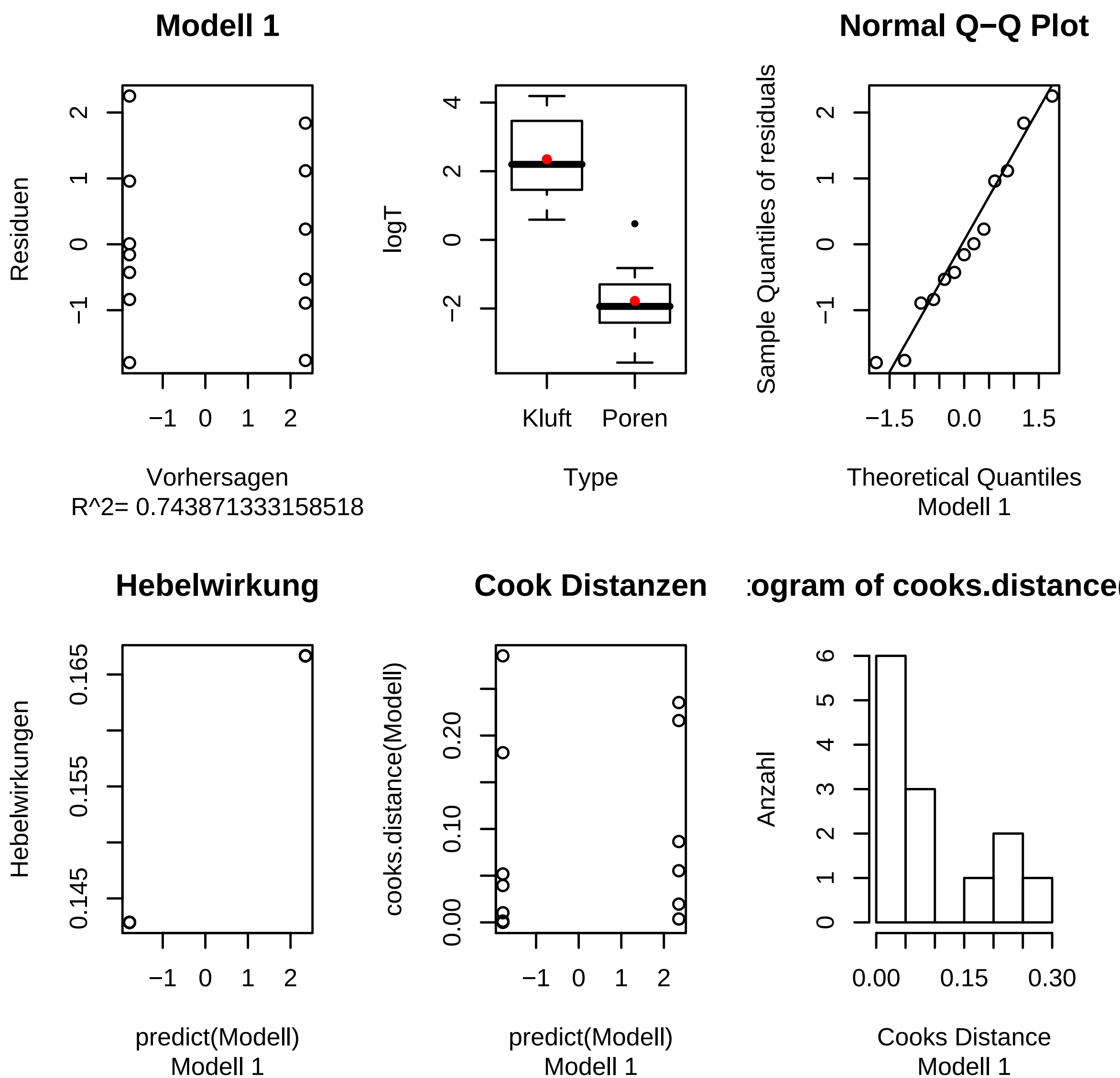
Beobachtungen zur Varianzanalyse

- 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:

$$\text{``}\texttt{logT}\text{''} = a + bf_1(\text{``}\texttt{Type}\text{''}) + \epsilon$$

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

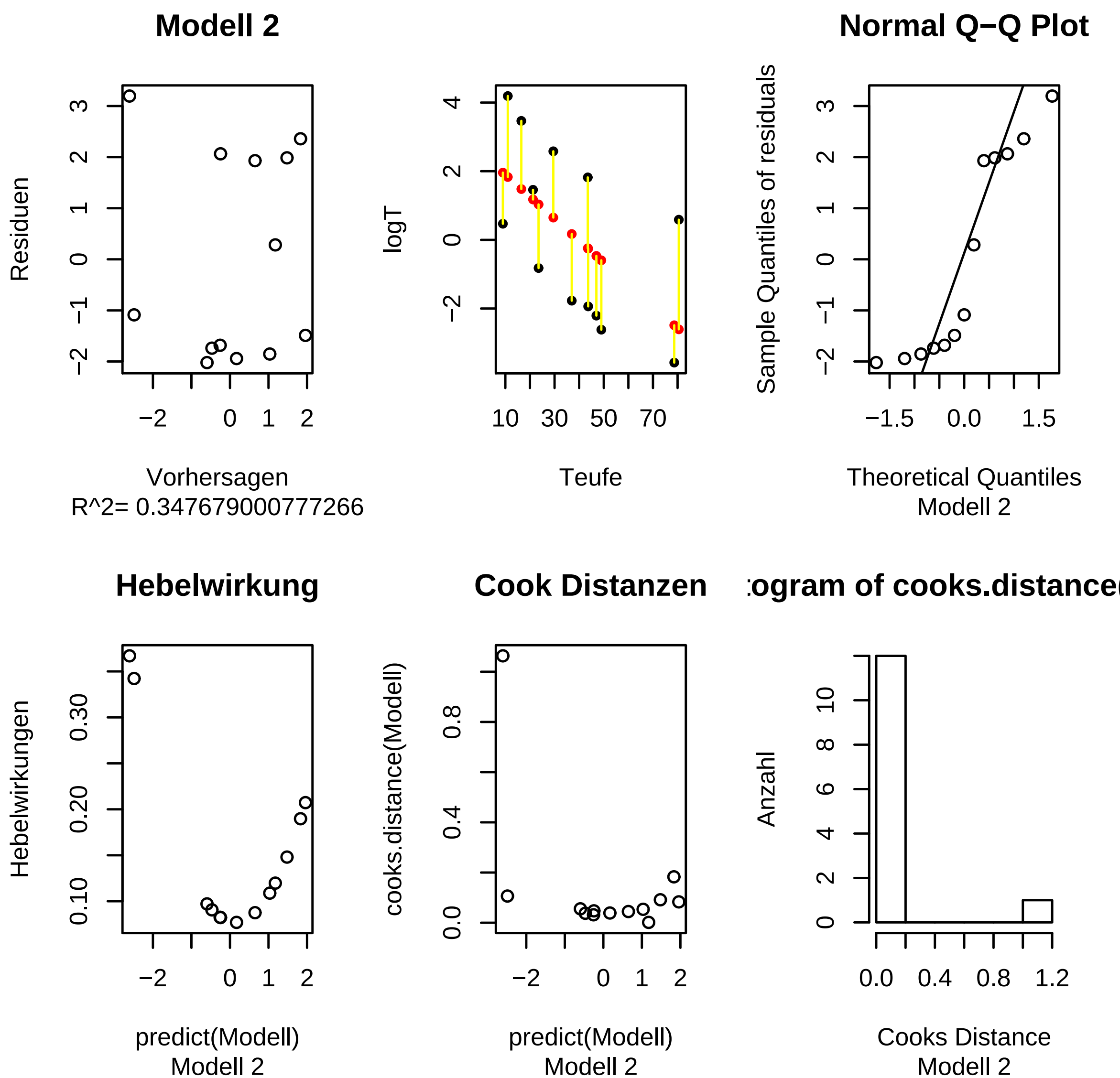
```
> Auswertung(logT ~ Teufe, Aqui, "Modell 2")

R^2= 0.3477
Analysis of Variance Table

Response: logT
          Df Sum Sq Mean Sq F value Pr(>F)
Teufe      1   25.8    25.8    5.86  0.034 *
Residuals 11   48.3     4.4
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
Estimated Parameters of the model:
(Intercept)       Teufe
    2.53348    -0.06386
```

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

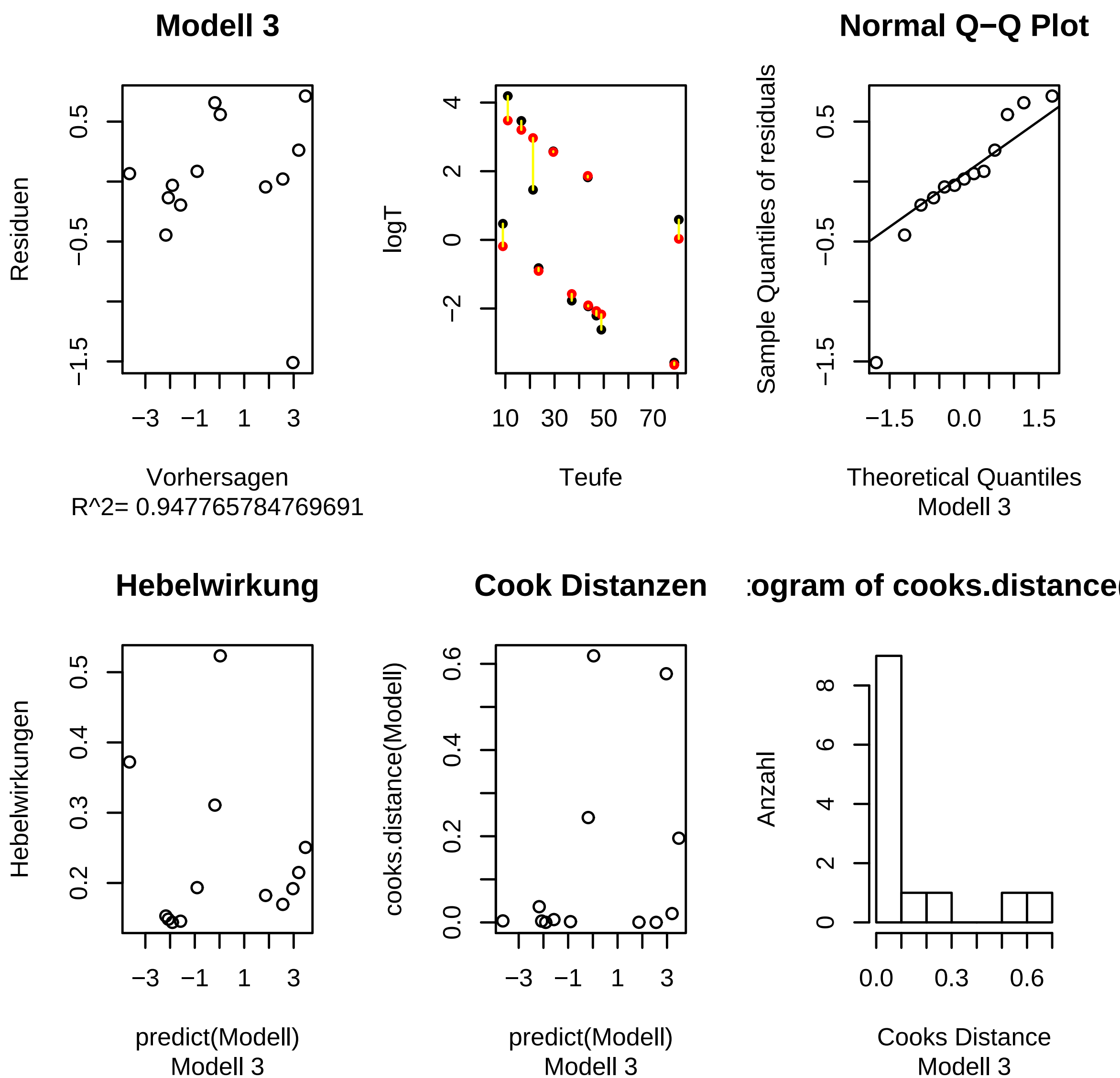
Wir können beide Einflüsse in das Modell aufnehmen:
`logT~Teufe+Type`

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

Lineares Modell für logT

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

R^2= 0.9478
Analysis of Variance Table


Response: logT
          Df Sum Sq Mean Sq F value  Pr(>F)
Teufe      1   25.8    25.8    66.6 9.9e-06 ***
Type       1   44.5    44.5   114.9 8.4e-07 ***
Residuals 10    3.9     0.4
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
Estimated Parameters of the model:
(Intercept)       Teufe   TypePoren
     4.0223     -0.0496     -3.7630
```

Diagnostik

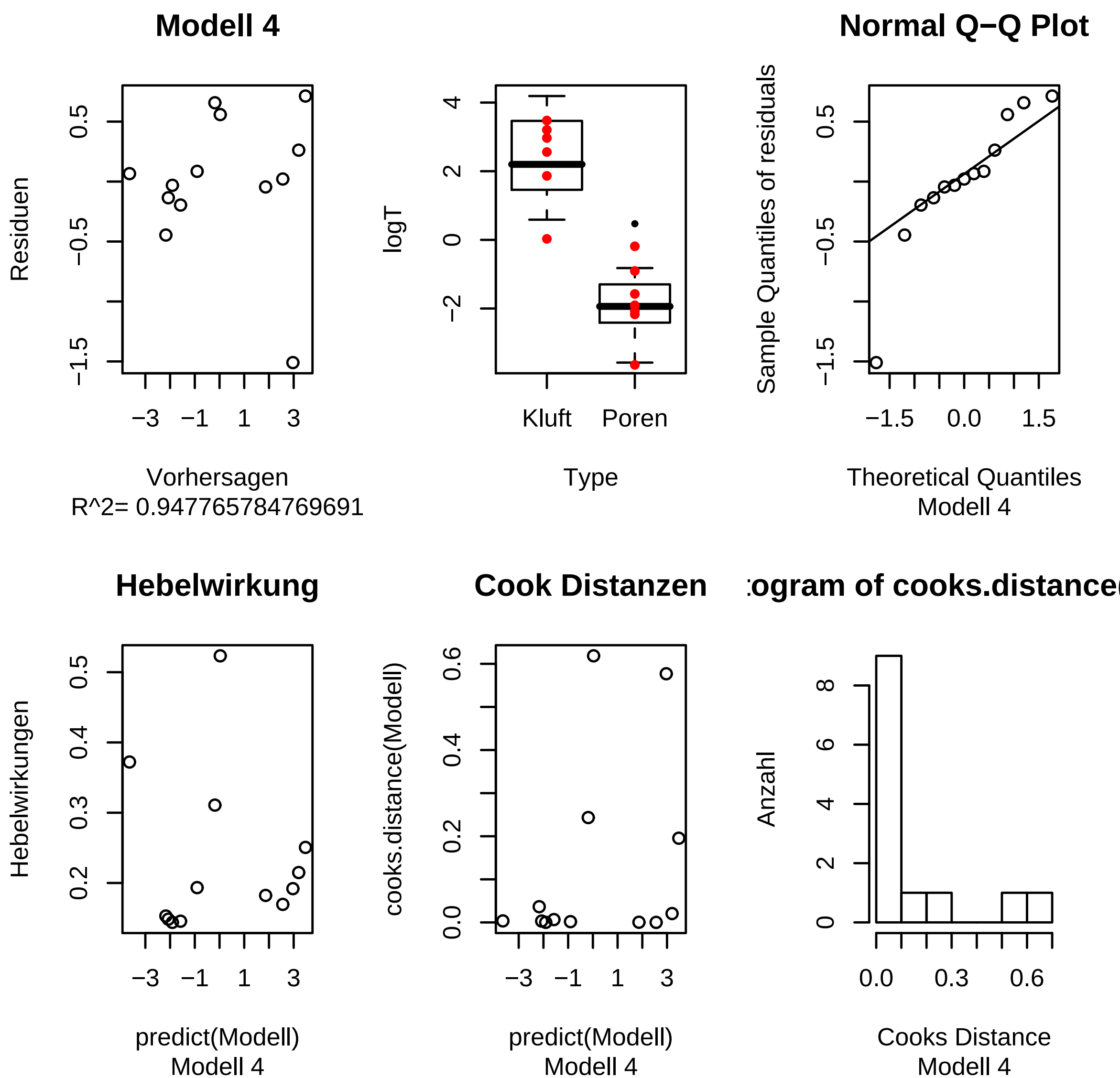


Modell 3
Residuen
0.5
−0.5
−1.5
−3 −1 1 3
Vorhersagen
R^2= 0.947765784769691
logT
4
2
0
−2
10 30 50 70
Teufe
Normal Q−Q Plot
Sample Quantiles of residuals
−1.5 0.0 1.5
Theoretical Quantiles
Modell 3
Hebelwirkung
Hebelwirkungen
0.5
0.4
0.3
0.2
predict(Modell)
Modell 3
Cook Distanzen
cooks.distance(Modell)
0.6
0.4
0.2
0.0
predict(Modell)
Modell 3
:ogram of cooks.distance(
Anzahl
8
6
4
2
0
0.0 0.3 0.6
Cooks Distance
Modell 3

…und in umgekehrter Reihenfolge

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

R^2= 0.9478
Analysis of Variance Table


Response: logT
          Df Sum Sq Mean Sq F value  Pr(>F)
Type       1   55.1    55.1     142 3.1e-07 ***
Teufe      1   15.1    15.1      39 9.5e-05 ***
Residuals 10    3.9     0.4
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
Estimated Parameters of the model:
(Intercept)   TypePoren       Teufe
     4.0223     -3.7630     -0.0496
```

Diagnostik



Modell 4
Residuen
0.5
−0.5
−1.5
−3
−1
1
3
Vorhersagen
R^2= 0.947765784769691
logT
4
2
0
−2
Kluft
Poren
Type
Normal Q−Q Plot
Sample Quantiles of residuals
−1.5
0.0
1.5
Theoretical Quantiles
Modell 4
Hebelwirkung
Hebelwirkungen
0.5
0.4
0.3
0.2
predict(Modell)
Modell 4
Cook Distanzen
cooks.distance(Modell)
0.6
0.4
0.2
0.0
predict(Modell)
Modell 4
:ogram of cooks.distance(
Anzahl
8
6
4
2
0
0.0
0.3
0.6
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:

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

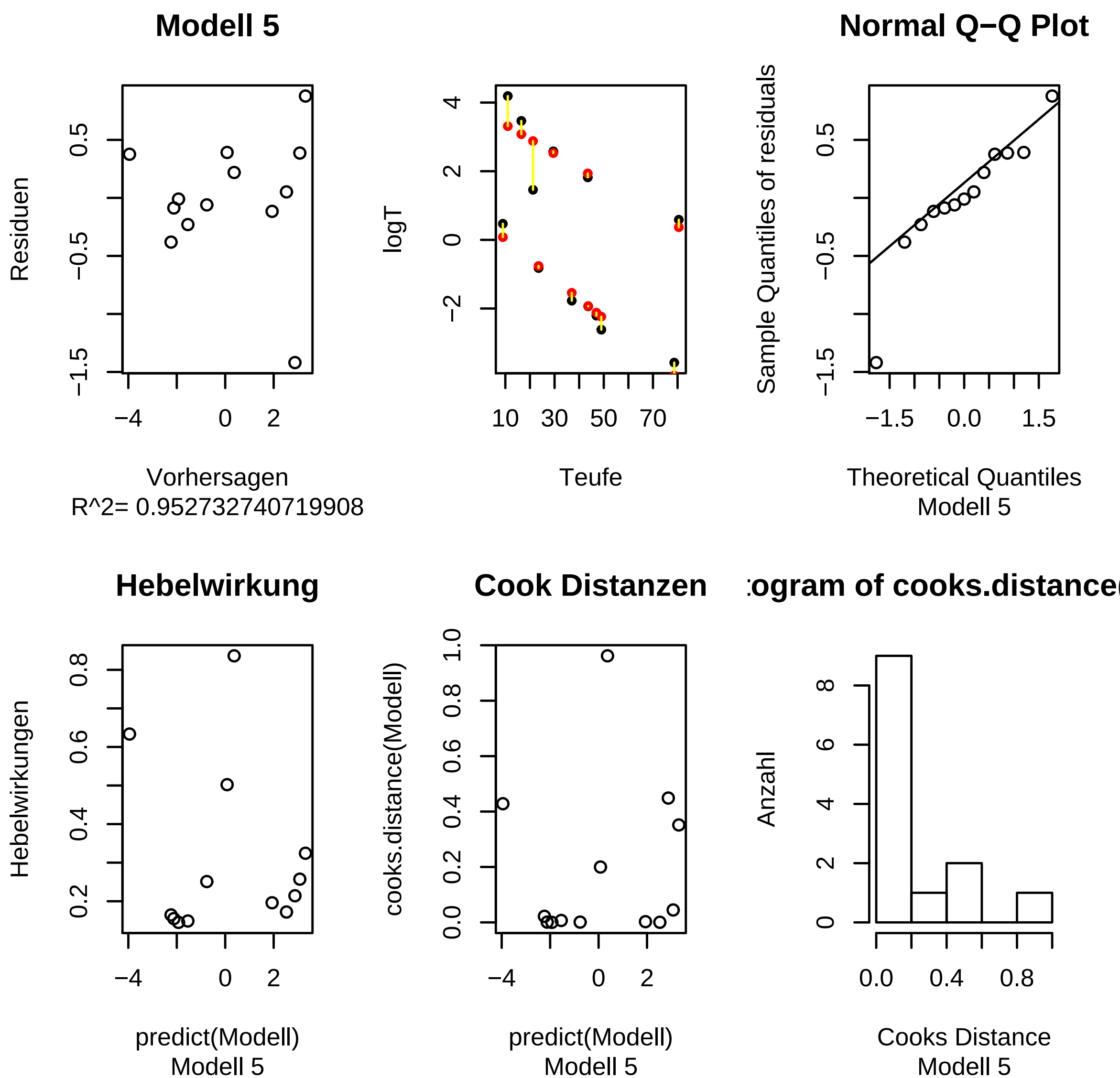
…und mit Interaktion

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

R^2= 0.9527
Analysis of Variance Table


Response: logT
           Df Sum Sq Mean Sq F value  Pr(>F)
Teufe       1   25.8    25.8   66.20 1.9e-05 ***
Type        1   44.5    44.5  114.26 2.0e-06 ***
Teufe:Type  1    0.4     0.4    0.95    0.36
Residuals   9    3.5     0.4
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
Estimated Parameters of the model:
    (Intercept)           Teufe       TypePoren
        3.77785        -0.04235        -3.17872
Teufe:TypePoren
```

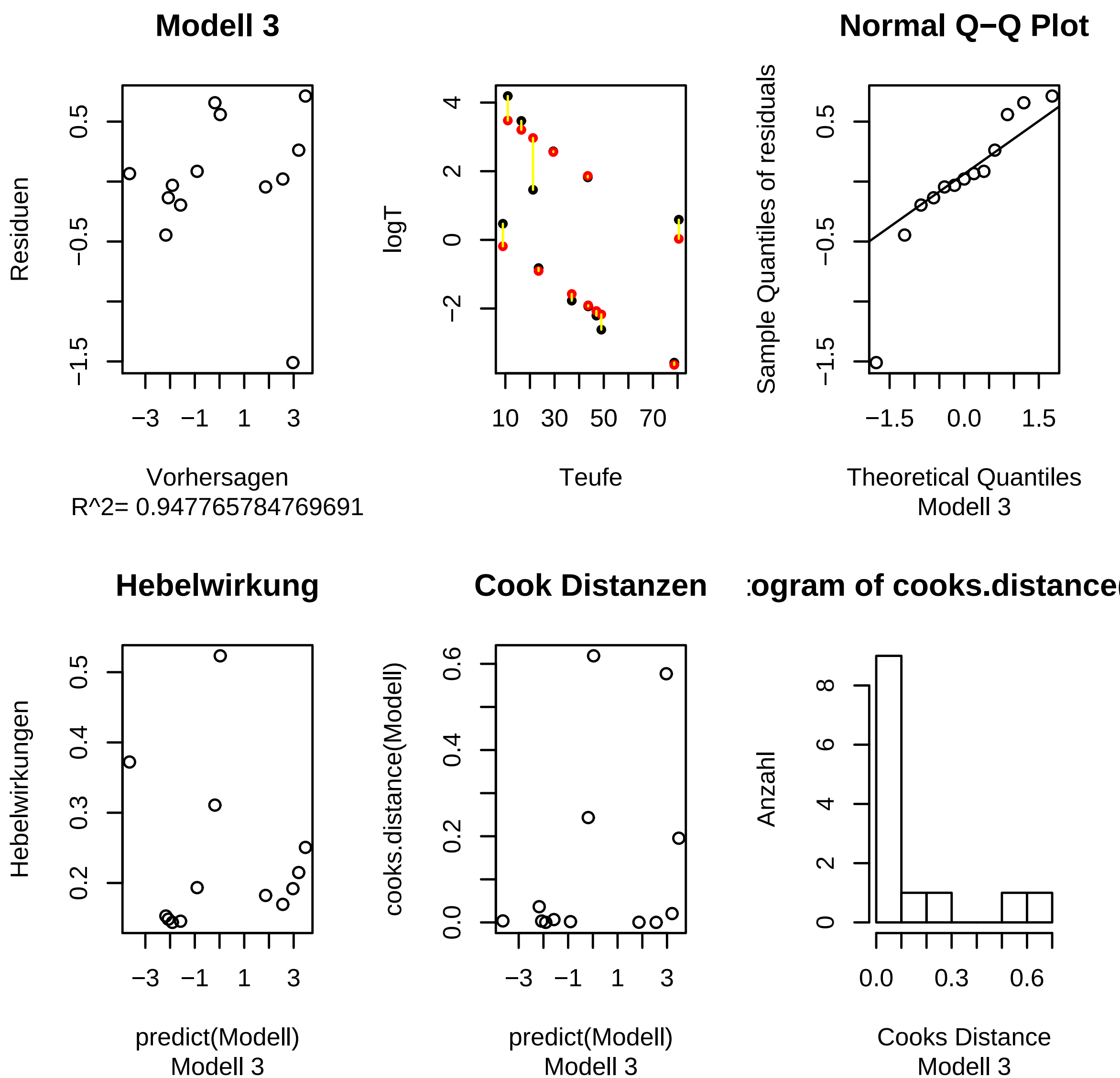
Ergebnis

Interaktion wird nicht benötigt.

Diagnostik mit Interaktion



Diagnostik ohne Interaktion



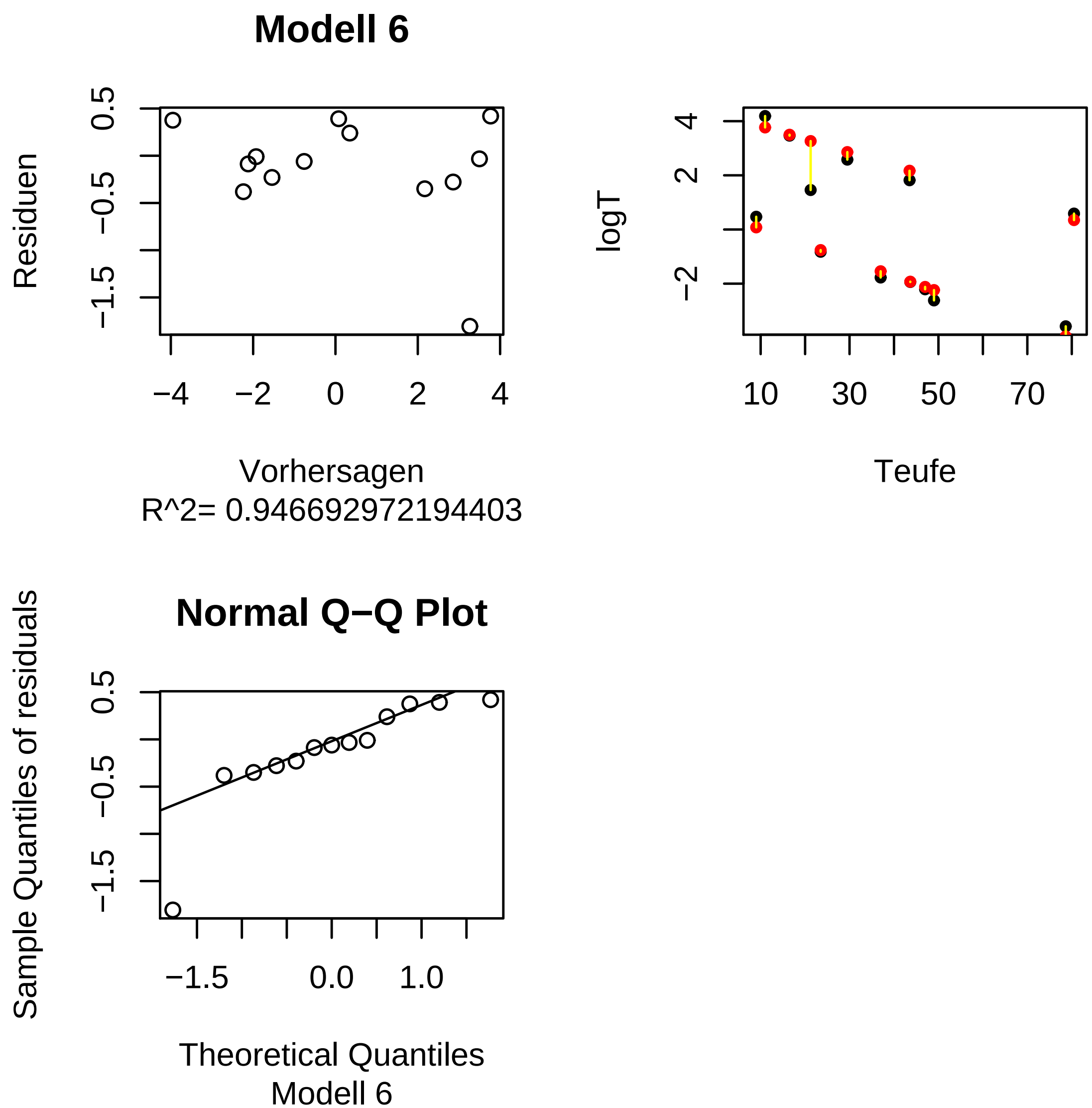
Robuste Schätzung

Die Robuste Methodik kann auch hier angewendet werden.

…und mit robuster Schätzung

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

R^2= 0.9467


Terms added sequentially (first to last)


            Chisq Df RobustF   Pr(F)
(Intercept)        1
Teufe              1     3.9  0.0430 *
Type               1    51.0 3.3e-13 ***
Teufe:Type         1     6.9  0.0076 **
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ
Estimated Parameters of the model:
    (Intercept)            Teufe        TypePoren
       4.310786        -0.049231        -3.711659
```

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