

Kapitel 4

Regression

4.1 Allgemeines Regressionsmodell

Oft werden in der Statistik die Abhängigkeiten einer Variable Y (genannt **Zielgröße** oder **Variable**) von Einflüssen X_1, \dots, X_d (genannt **unabhängigen Variablen** oder **Einflussgrößen**) untersucht.

Modell 1 (Allgemeines Regressionsmodell) Ein (unvollständiges) statistisches Modell der Form:

$$P^Y(\cdot | X_1 = x_1, \dots, X_d = x_d) = \text{Funktion von } x_i \text{ und } \beta_i$$

welches die bedingte Verteilung von Y gegeben die Einflussgrößen als bekannte Funktion der Einflussgrößen und der Parameter β_i beschreibt, heißt ein **Regression**.

Einfache Beispiele sind:

- Einfache **lineare Regression** (x ist reell)

$$P^Y(\cdot | X = x) = N(a + bx, \sigma^2)$$

$$\beta = (a, b, \sigma^2)$$

- Einfaches **Varianzanalysemodell** (X ist kategoriell)

$$P^Y(\cdot | X = x) = N(a_x, \sigma^2)$$

$$\beta = (a_1, \dots, a_k, \sigma^2)$$

- Poisson Mehrgruppenmodell (Y ist ganzzahlig, X ist kategoriell)

$$P^Y(\cdot | X = x) = Po(\lambda_x)$$

$$\beta = (\lambda_1, \dots, \lambda_d)$$

4.1.1 Überblick über die Regressionsmodelle

- **Lineare Modelle** $Y \sim N(\sum_{k=0}^p f_k(X)\beta_k, \sigma^2)$
 - **Lineare Regression** ($X \in \mathbb{R}$, $Y \sim N(a + bX, \sigma^2)$)
 - **Einfache Varianzanalyse** (X kategoriell, $Y \sim N(a_X, \sigma^2)$) (engl. **ANOVA** = ANalysis Of VArianz)
 - **Multiple Regression** ($X \in \mathbb{R}^d$, $Y \sim N(a + b_1X_1 + \dots + b_dX_d, \sigma^2)$)

- **Multiple Varianzanalyse** (X_1, \dots, X_d kategoriell, $Y \sim N(a_{X_1} + \dots + c_{X_d} + d_{X_1 X_2} + \dots, \sigma^2)$)
- **Polynomiale Regression** $Y \sim N(\text{Polynom}(X), \sigma^2)$
- **Lineares Modell** X_1, \dots, X_d irgendwas, $Y \sim N(\dots, \sigma^2)$
- **Random Effekts Modell** (X nominal)
- **Mixed Effects Modell** (Mischung aus Random Effects und linearem Modell)

- **Tree Regression**

In der Tree Regression wird für unterschiedliche Bereiche von X ein unterschiedliches Lineares Modell verwendet.

- **Generalisierte lineare Modelle**

In generalisierten linearen Modellen wird eine lineare Beziehung zwischen Parametern und einer Funktion des Erwartungswertes vorausgesetzt: X beliebig, $Y \sim P_\theta$ als reelle Zahl interpretierbar (z.B. Anzahl, dichotom, reell) mit einer (fast beliebigen Verteilungsfamilie) P_θ . Das allgemeinste Modell lautet dann:

$$\text{Linkfunktion}(E[Y]) = \sum_{k=0}^p f_p(X) \beta_k$$

Beispiele für generalisierte lineare Modelle sind:

- **Logistische Regression:** $Y \sim B(p)$ dichotom, Linkfunktion(p) = $\ln \frac{p}{1-p}$
- **Loglineares Modell:** Y = Anzahl der Beobachtungen in den Zellen einer Kontingenztafel, Linkfunktion(λ) = $\log(\lambda)$
- **Poisson Regression:** z.B. $Y \sim \text{Po}(\lambda(X))$
- **Gamma Regression:** z.B. $Y \sim \text{Gamma}(\lambda(X), d)$

- **Generalisierte additive Modelle**

- **Nichtparametrische Regression** $X \in \mathbb{R}$, $Y \sim N(f(X), \sigma^2)$ mit $f(X)$ einer hinreichend einfachen Funktion (z.B. stetig diffbar, wenige Sprungstellen)
- **Multiple nichtparametrische Regression** $X \in \mathbb{R}^d$, $Y \sim N(\sum_{i=1}^d f(X_i), \sigma^2)$

4.2 Allgemeines lineares Modell

4.2.1 Definition

Das lineare Modell vereinfacht das allgemeine Regressionsmodell:

Modell 2 (Lineares Modell) Ein Regressionmodell der Form:

$$P^Y(\cdot | X_1 = x_1, \dots, X_d = x_d) = N\left(\beta_0 + \sum_k = 1^p f_k(x_1, \dots, x_d) b_k, \sigma^2\right) \text{ Funktion von } x_i \text{ und } \beta_i$$

welches die bedingte Verteilung von Y gegeben die Einflussgrößen als Normalverteilung mit unbekannter aber fester Varianz σ^2 und einem Erwartungswert der sich als eine Funktion schreiben lässt, die linear in den übrigen Parametern ist, heißt lineares Modell. Außerdem fordert man meistens, dass der Linearitätskoeffizient zum ersten Parameter b_0 die Konstante 1 ist. Das vereinfacht einiges.

Die Parameter sind dann $\beta = (b_0, b_1, \dots, b_p, \sigma^2)$.

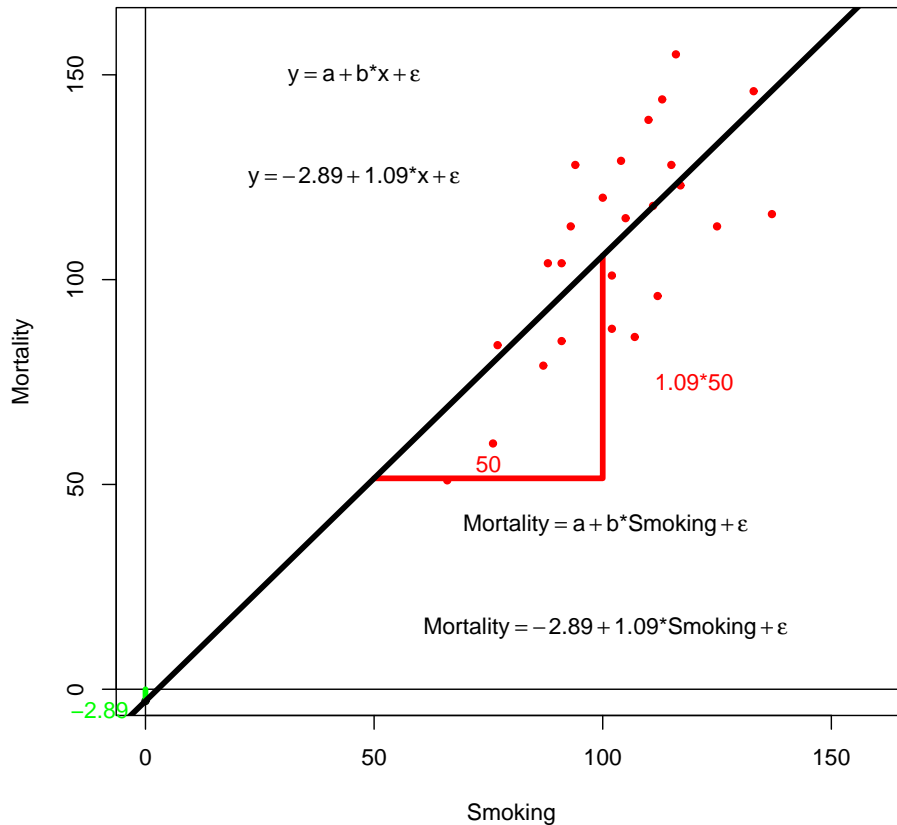


Abbildung 4.1: Geradengleichung der linearen Regression mit Achsenabschnitt a und Steigung b am Beispiel des Datensatzes <http://lib.stat.cmu.edu/DASL/Datafiles/SmokingandCancer.html> (bei Statlib)

Lineare Modelle werden meist in einer an die folgende Form angelehnten Schreibweise notiert:

$$Y_i = \beta_0 + \beta_1 f_1(\mathbf{x}_i) + \beta_2 f_2(\mathbf{x}_i) + \dots + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Dabei werden die Zielgröße Y und der **Fehler** ϵ als die einzigen zufälligen Einflüsse angesehen, auch wenn die x möglicherweise selbst zufällig sind. Diese Schreibweise und Bezeichnungsweise bezieht sich auf die Vorstellung, dass Y fast eine durch

$$Y \approx \beta_0 + \beta_1 f_1(x) + \beta_2 f_2(x) + \dots$$

gegebene Funktion der \mathbf{x} ist, aber jede einzelne Beobachtung durch einen zufälligen "Fehler" ϵ von diesem theoretischen Wert abweicht.

4.2.2 Beispiel lineare Regression

Ein wichtiger Spezialfall des linearen Modells ist die lineare Regression. Dabei ist die Einflussgröße ein einzelne reelle Größe und die Abhängigkeit wird als Gerade $a + bx$ mit Achsenabschnitt a und Steigung b modelliert.

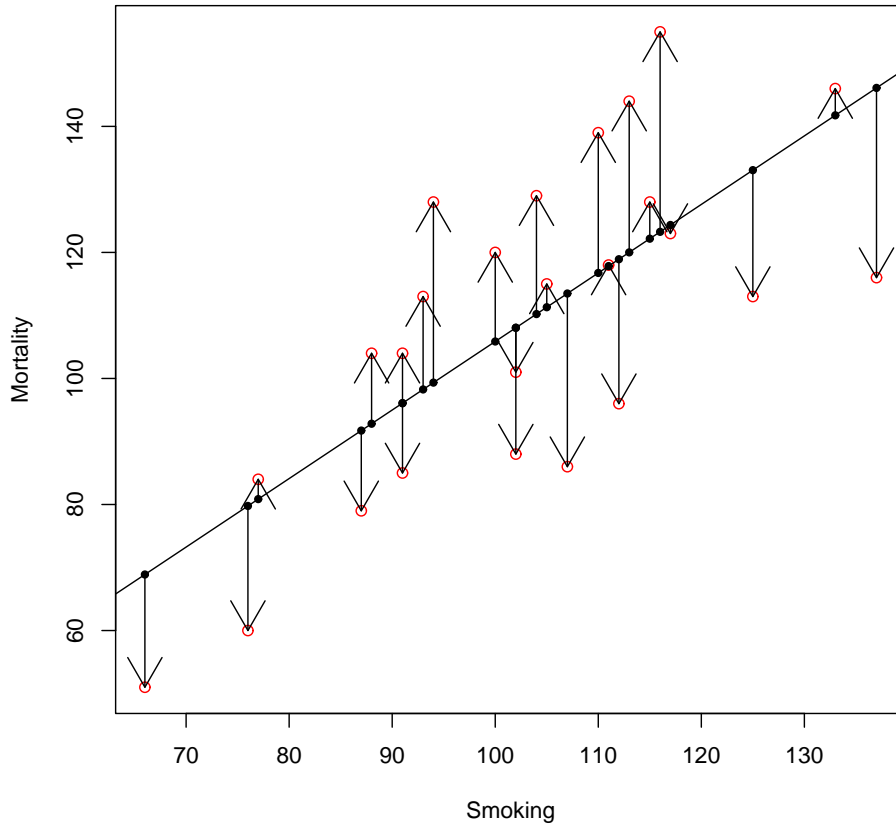


Abbildung 4.2: Residuen der linearen Regression am Beispiel des Datensatzes <http://lib.stat.cmu.edu/DASL/Datafiles/SmokingandCancer.html> (bei Statlib)

Modell 3 (Lineare Regression)

$$Y = a + bx + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

4.3 Statistik linearer Modelle

4.3.1 Ziele I

4.3.1.1 Beispiel: Transmissivität eines Grundwasserleiters

.	logT	Teufe	Type
1	-3.5755508	78.64	Poren
2	-2.6172958	49.00	Poren
3	-2.2072749	47.00	Poren
4	-1.9379420	43.67	Poren
5	-1.7719568	37.00	Poren
6	-0.8209806	23.50	Poren
7	0.4700036	9.00	Poren
8	0.5877867	80.50	Kluft

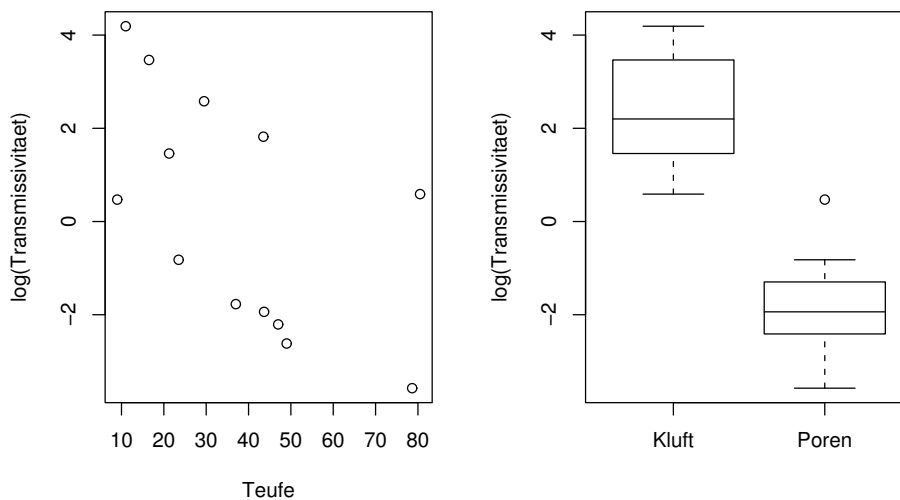
```

9   1.4586150 21.25 Kluft
10  1.8196988 43.50 Kluft
11  2.5802168 29.50 Kluft
12  3.4657359 16.50 Kluft
13  4.1896547 11.00 Kluft

```

In einem Waldgebiet wurden in verschiedenen Bohrlöchern in unterschiedlicher Tiefe die Transmissivität der Grundwasserleiter gemessen. Die Transmissivität ist ein Maß für die Wasserleitfähigkeit des Grundwasserleiters. Man unterscheidet drei grundsätzliche Typen: Porenleiter, Kluftleiter und Nichtleiter.

$\log T$ = $\log(\text{Transmissivität})$
 $Teufe$ = Tiefe in Metern unter der Erdoberfläche
 $Type$ = Typ des Leiters (Poren oder Kluft)



- Abhängigkeit in den Daten durch ein lineares Modell beschreiben.

Was muss man dazu können?

- Geeignete Modelle formulieren. z.B. $\log T = a + b \text{Teufe} + \varepsilon$
- Das richtige Modell auswählen. z.B. $\log T = a + b_{\text{Type}}$
- Überprüfen ob dieses Modell die Daten richtig beschreibt.
- Voraussetzungen der dazu benötigten Tests überprüfen.
- Parameter schätzen und Konfidenzintervalle angeben (a, b, σ^2).
- Die Güte der Beschreibung quantifizieren und bewerten (z.B. Tiefeneffekt, klein).

Was hat man davon?

- Das Modell beschreibt die Zusammenhänge.
- Die Art der Zusammenhänge lässt oft Rückschlüsse auf die zugrundeliegenden Wirkmechanismen zu.

- Man kann für weitere unbeobachtete Fälle, den vermuteten y-Wert ungefähr angeben, wenn man die x-Werte kennt.
- Man kann die Wichtigkeit von Einflüssen quantifizieren.

4.3.2 Design linearer Modelle

Was gibt es für Modelle und was bedeuten sie?

4.3.2.1 Aufsteigende Modellsequenzen

- Ein lineares Modell wird aus Bausteinen aufgebaut.

$$y = \underbrace{a}_{B_0} + \underbrace{bx}_{B_1} + \underbrace{c_k}_{B_2} + \dots + \underbrace{\epsilon}_{\text{Residuen}}$$

- Daraus ergibt sich eine aufsteigende Folge von Teilmodellen

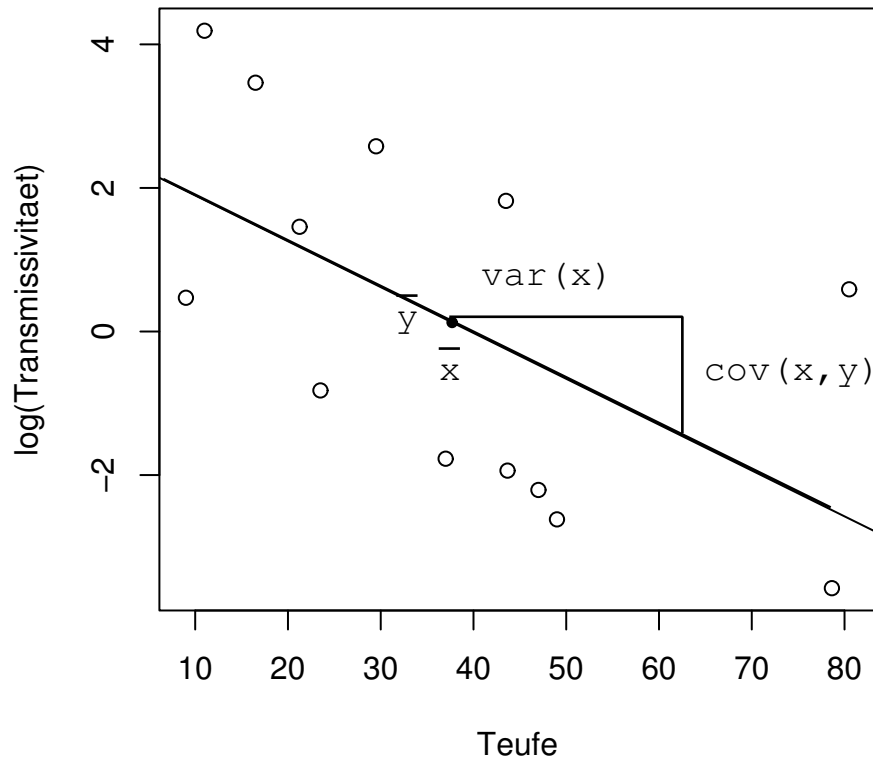
$$\begin{aligned} \text{Modell}_1 : y &= \underbrace{a}_{B_0} + \epsilon \\ \text{Modell}_2 : y &= \underbrace{a}_{B_0} + \underbrace{bx}_{B_1} + \epsilon \\ \vdots & \\ &: y = \underbrace{a}_{B_0} + \underbrace{bx}_{B_1} + \dots + \epsilon \end{aligned}$$

Wobei jeweils mit Modell_i darstellbaren Abhängigkeiten auch mit allen höheren Modellen_j, $j > i$ darstellbar sind, wenn man einfach gewisse Parameter auf 0 setzt.

4.3.2.2 Problem: Auswahl des richtigen Modells

- So einfach wie möglich.
- So kompliziert wie nötig.
 - Nach logischer Analyse des Untersuchungsgegenstandes.
 - Nach der Datenlage.

4.3.2.3 Lineare Regression



$$y = a + b\text{Teufe} + \epsilon$$

$$y = 2.53 + -0.06 \frac{1}{m} \text{Teufe} + \epsilon,$$

$$\epsilon \sim N(\mu = 0, \sigma = 2.096)$$

a = Achsenabschnitt
 = $\log T$ an der Oberfläche
 b = Anstieg
 = Änderung von $\log T$ je Meter Tiefe

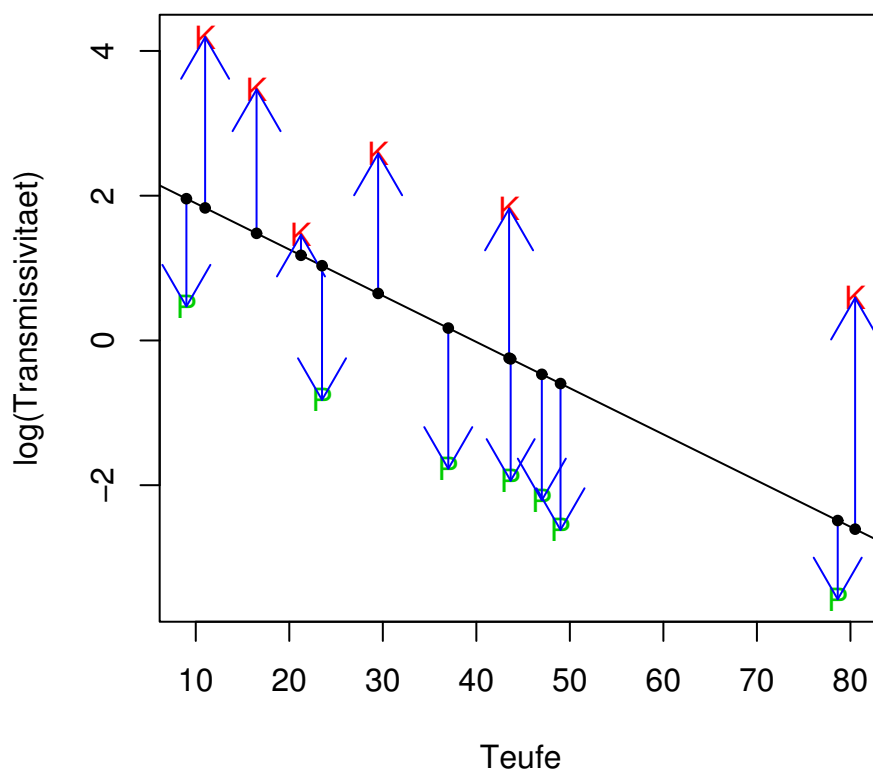
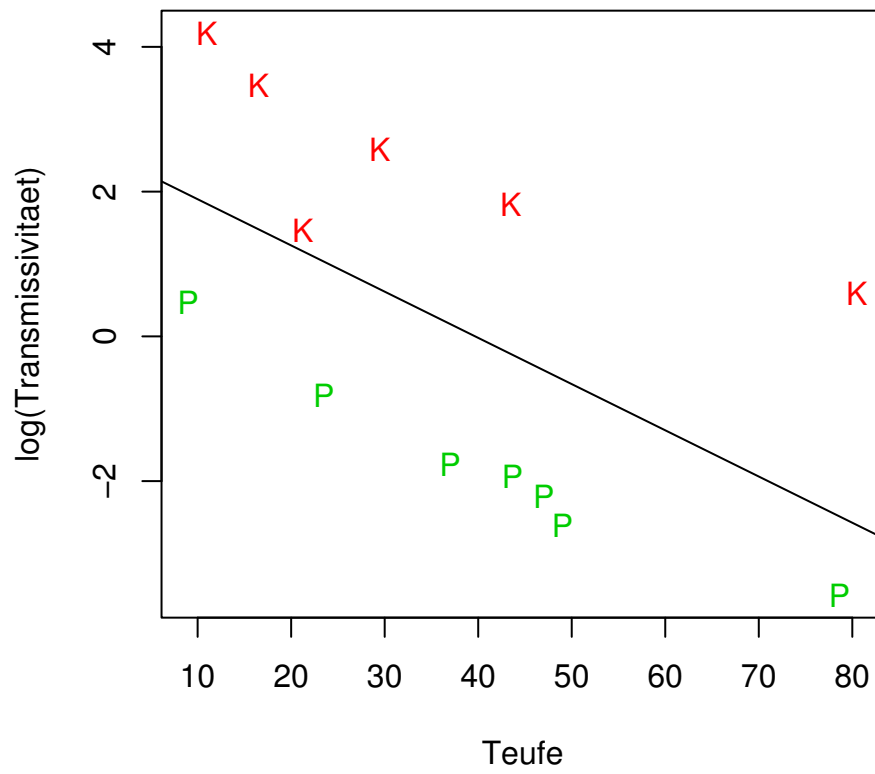
Die Linie geht immer durch \bar{x}, \bar{y} und hat den Anstieg $\frac{\widehat{\text{cov}}(x,y)}{\widehat{\text{var}}x}$.

Lineare Modelle werden oft im sogenannten **Wilkinson-Roger-Syntax** dargestellt. Diese Schreibweise lässt die Konstanten weg und ersetzt das Gleichheitszeichen durch ein \sim :

Wilkinson-Roger-Syntax: $y \sim x$

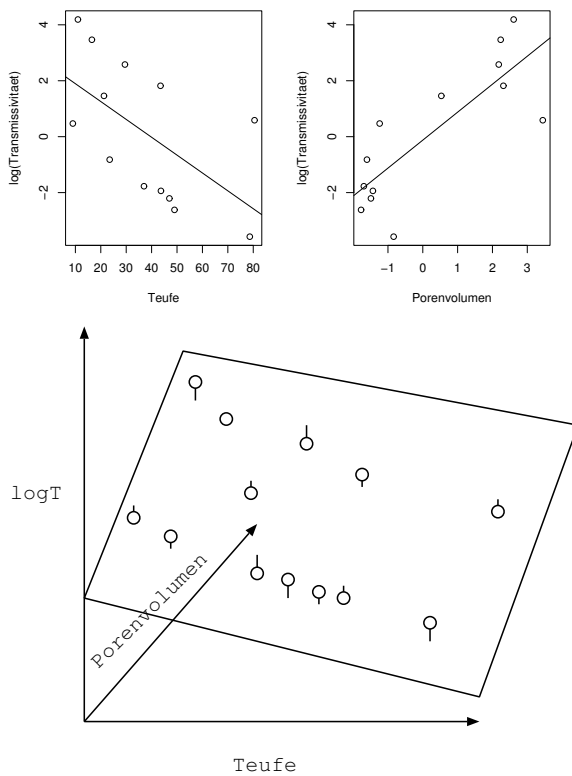
Das \sim könnte man als ein "wird modelliert als abhängig von" lesen.

```
> coef(lm(logT~Teufe,data=Aqui))
(Intercept)      Teufe
 2.53348382 -0.06385867
```



4.3.2.4 Multiple lineare Regression

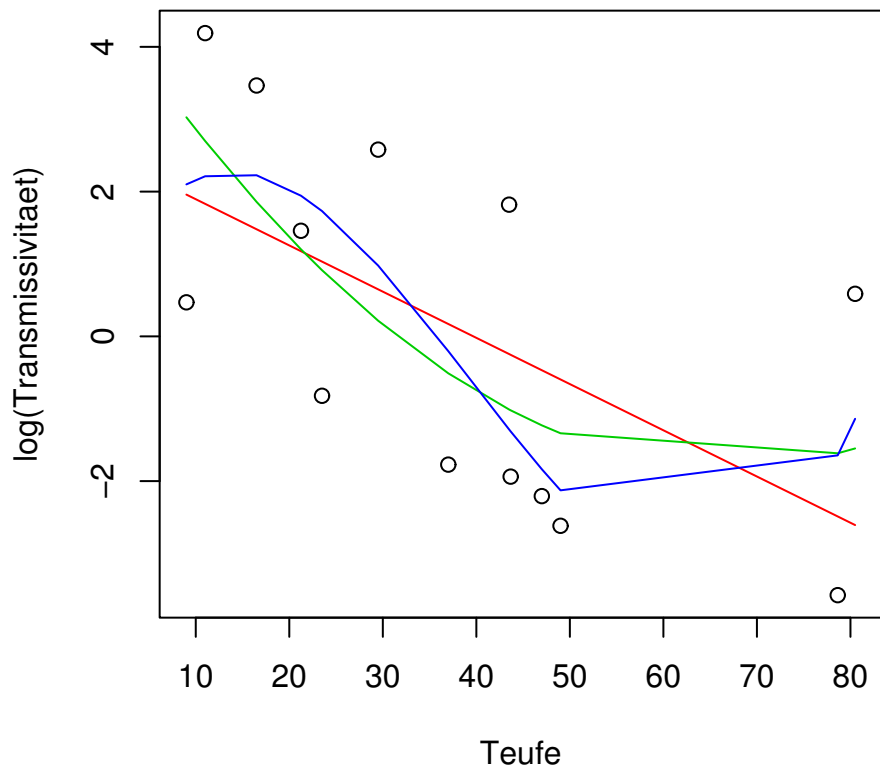
$$y = a + b_1 \text{Teufe} + b_2 \text{Spezifisches Porenvolumen} + \epsilon$$



Wilkinson-Roger-Syntax: $y \sim \text{Teufe} + \text{Spezifisches Porenvolumen}$
 Das + kann als "und" gelesen werden.

4.3.2.5 Polynomiale Regression

$$y = a + \underbrace{b_1x + b_2x^2 + b_3x^3}_{\text{niederer Monome in } x} + \epsilon$$



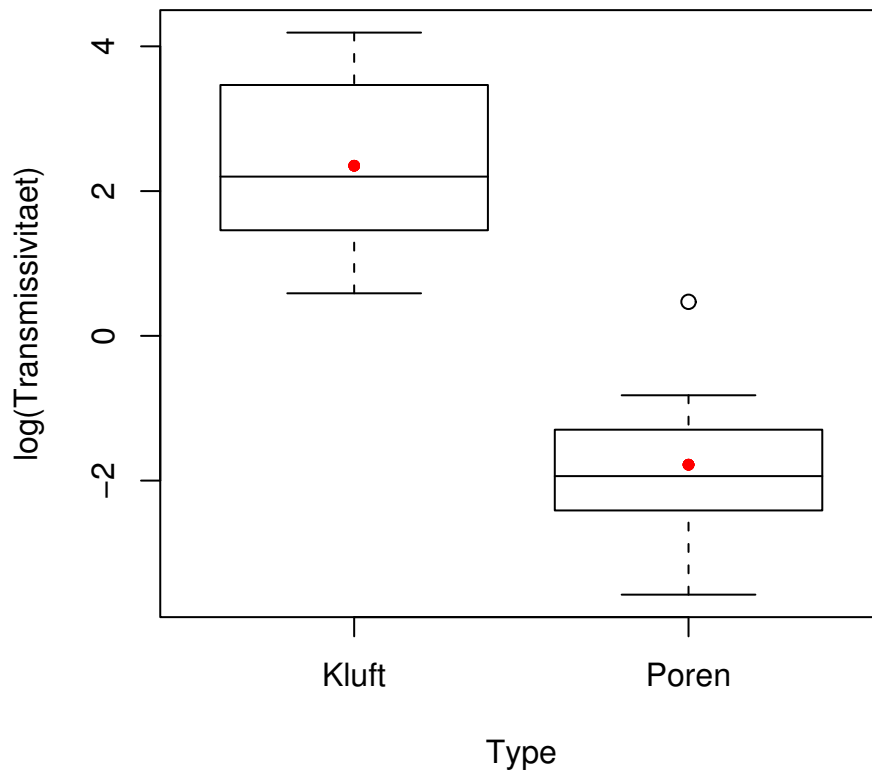
Hohe Polynomgerade führen praktisch immer zu unsinnigen Interpolationen.

Wilkinson-Roger-Syntax:

$$y \sim Teufe + I(Teufe^2) + I(Teufe^3)$$

4.3.2.6 Varianzanalyse/ANOVA

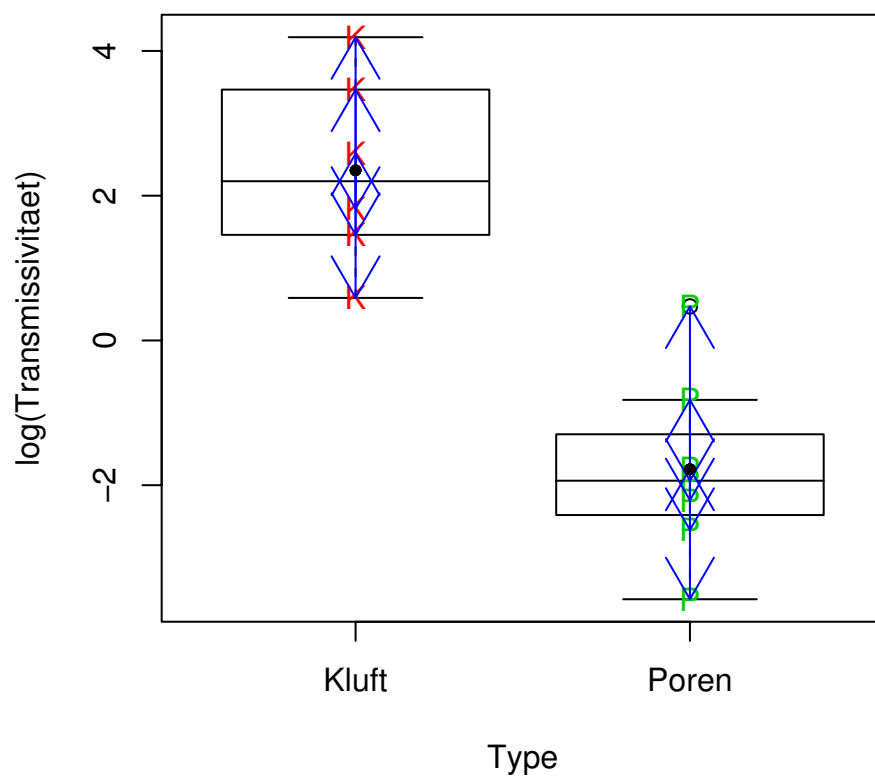
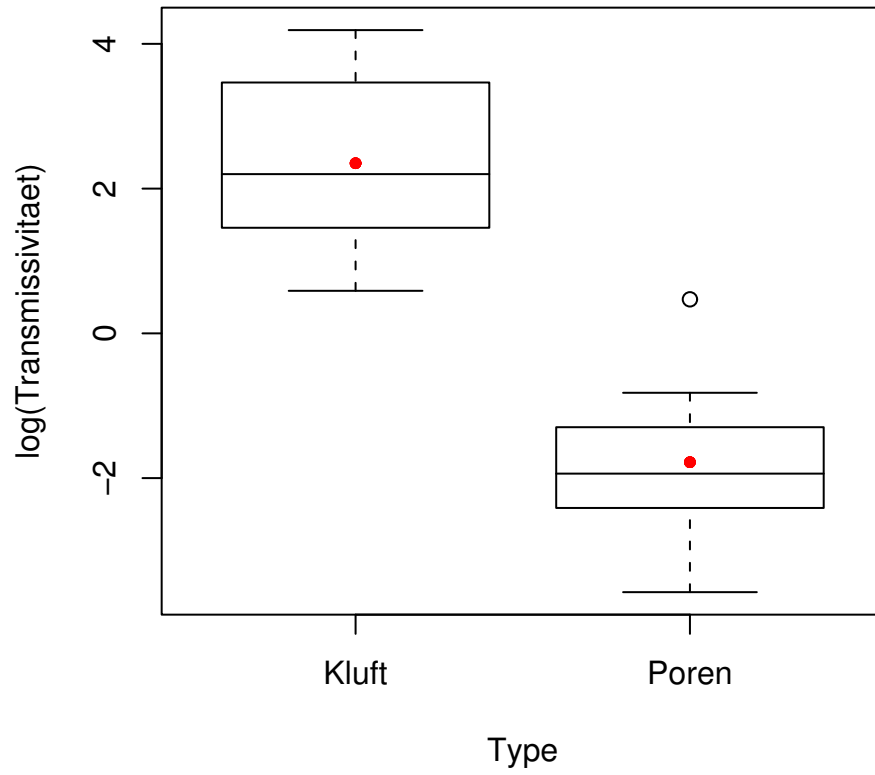
ANOVA (ANalysis Of VARIanz)



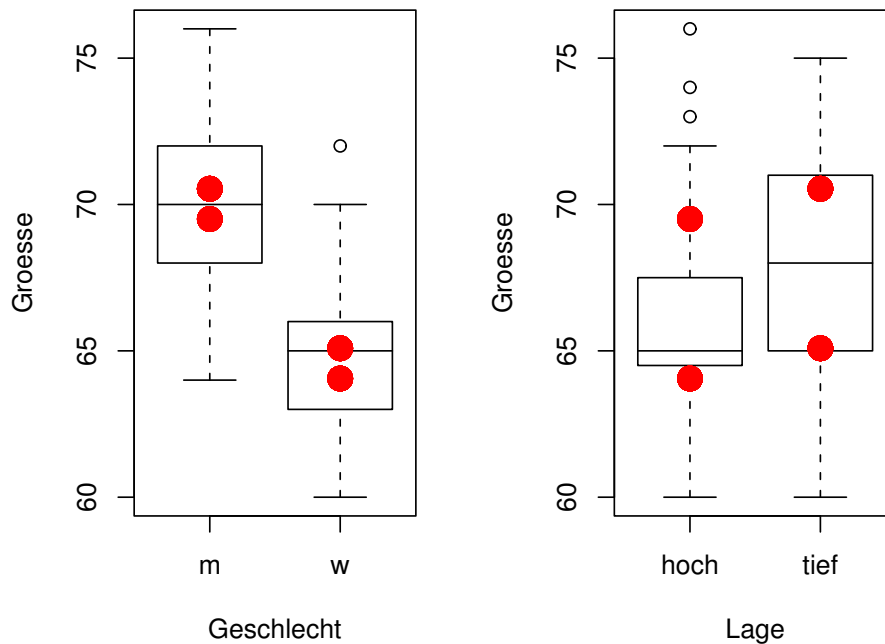
$$y = a + c_k + \varepsilon$$

Wilkinson-Roger-Syntax:
 $\log T \sim \text{Type}$

(Intercept)	TypePoren
2.350	-4.130



4.3.2.7 Multifaktorielle Varianzanalyse



$$y = a + c_k + d_x + \dots + \varepsilon$$

Wilkinson-Roger-Syntax:

$Groesse \sim Geschlecht + Lage$

Männer sind im Schnitt 5in größer als Frauen

Leute mit tiefer Stimme im Schnitt 1in größer als solche mit hoher Stimme.

(Intercept)	Geschlechtw	Lagetief
69.503	-5.449	1.034

4.3.2.8 Interaktion

Drei äquivalente Ideen führen zu Interaktionen

- Idee: Der Größenunterschied zwischen hohen und tiefen Stimmen könnte bei Männern und Frauen unterschiedlich stark ausgeprägt sein.
- Idee: Der Größenunterschied zwischen Männern und Frauen könnte bei unterschiedlicher Stimmlage unterschiedlich stark ausgeprägt sein.
- Idee: Für jede der Gruppen hoch-w, tief-w, hoch-m, tief-m gibt es einen verschiedenen Mittelwert.

4.3.2.9 Interaktion von Faktoren

$$y = a + b_k + c_x + d_{kx} + \varepsilon$$

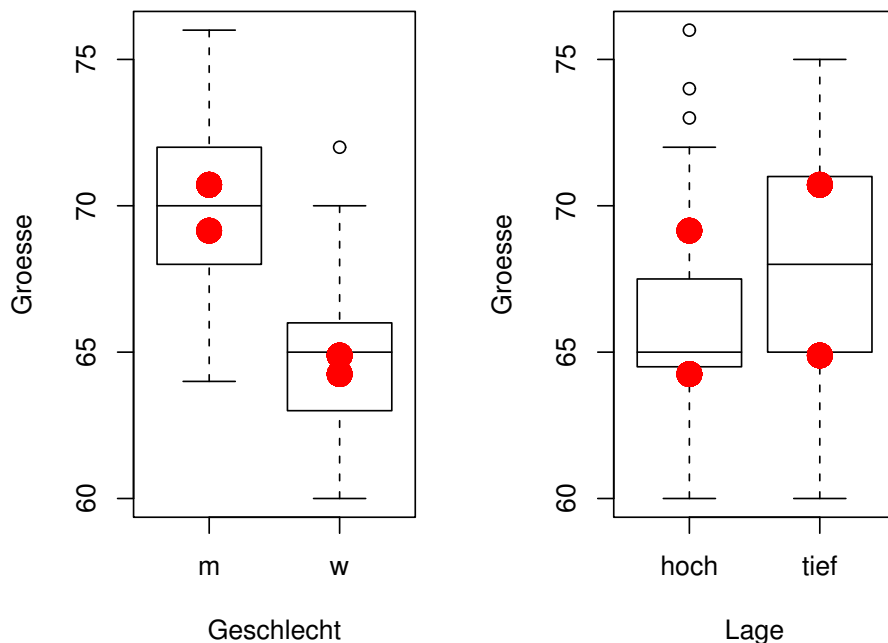
Wilkinson-Roger-Syntax:

$Groesse \sim Geschlecht + Lage + Geschlecht * Lage$

bzw.

$Groesse \sim Geschlecht * Lage$

Die kleineren Terme gelten jeweils als implizit mitnotiert. Bei den Effekten, die nur von einer Einflussgröße abhängen, spricht man auch von **Haupteffekten**. Effekte, die von mehreren Zufallsgrößen abhängen, heißen auch **Interaktionen**.



$Groesse \sim Geschlecht * Lage$

Die Stimmlage hat bei Männern einen größeren Einfluss als bei Frauen.

(Intercept)	Geschlechtw	Lagetief	Geschlechtw:Lagetief
69.1500	-4.9000	1.5679	-0.9322

4.3.2.10 Höhere Faktorinteraktionen

Höhere Faktorinteraktionen sind Interaktionen mehrerer Faktoren.

Beispiel: Ein Interaktion zweiter Stufe (mit 3 Faktoren)

- Interpretation: z.B. Je nach Kombination von a und b hat c eine andere Wirkung.

Wilkinson-Roger-Syntax:

$Groesse \sim Geschlecht + Lage + Schuhgröße$
 $+ Geschlecht * Lage + Geschlecht * Schuhgröße + Lage * Schuhgröße$
 $+ Geschlecht * Lage * Schuhgröße$

bzw.

$$Groesse \sim \text{Geschlecht} * \text{Lage} * \text{Schuhgrösse}$$

Die Faktoren oder Regressoren selbst heißen übrigens auch Haupteffekte.

4.3.2.11 Geschachtelte Faktoren/nested Factors

Machmal macht das Modell $y \sim \text{Geschlecht} + \text{Stimme}$ keinen Sinn, weil die gleichen Level in Stimme für verschiedene a nicht identifiziert werden können, hier z.B. weil hohe Frauenstimme viel höher sind als hohe Männerstimmen.

Wilkinson-Roger-Syntax:

$$Groesse \sim \text{Geschlecht} + \text{Lage} \% \text{in} \% \text{Geschlecht}$$

(Intercept)	Geschlechtw	Geschlechtm:Lagetief	Geschlechtw:Lagetief
69.1500	-4.9000	1.5679	0.6357

Das klassische Beispiel: Round-Robin-Test

$$Laborwert \sim \text{Patient} + \text{Labor} + \text{Laborant} \% \text{in} \% \text{Labor}$$

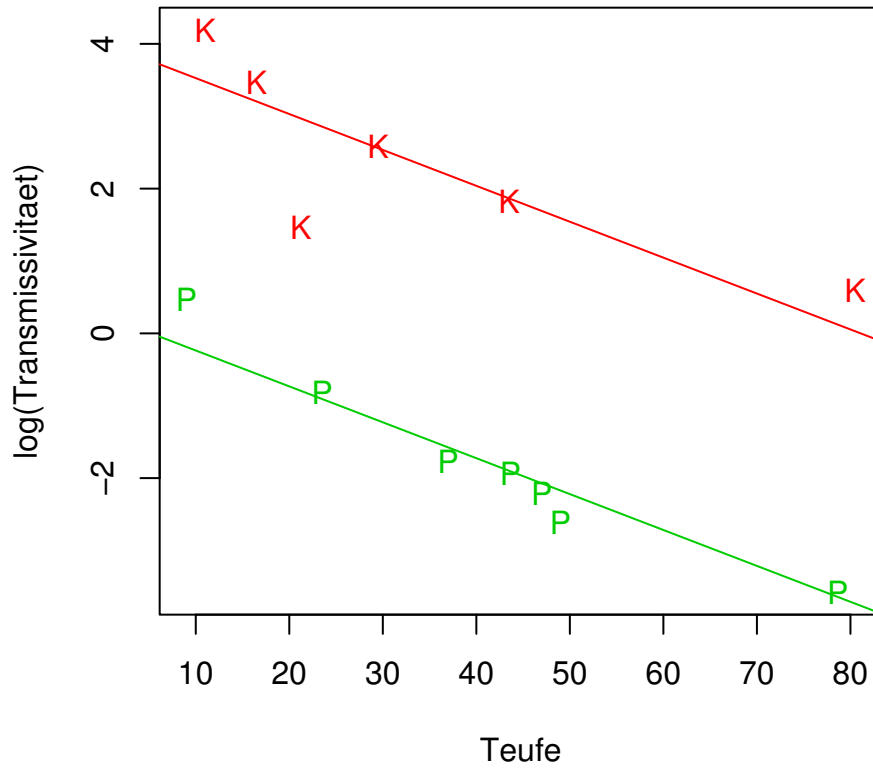
Der Hauptunterschied zu normalen Interaktionen liegt in der Veränderung der Sequenzfolge und nicht im resultierenden Modell.

4.3.2.12 Lineare Modelle mit Regressoren und Faktoren

$$\log T \sim \text{Type} + \text{Teufe}$$

$$\log T = a + b_{\text{Type}} + c \cdot \text{Teufe} + \varepsilon$$

Interpretation: die Leitfähigkeit beider Grundwasserleitertypen unterscheidet sich bereits bei 0m Tiefe. Sie ändert sich linear mit der Tiefe.



(Intercept)	TypePoren	Teufe
4.0223	-3.7630	-0.0496

Die Änderung pro Tiefenmeter ist für beide Leitertypen gleich.

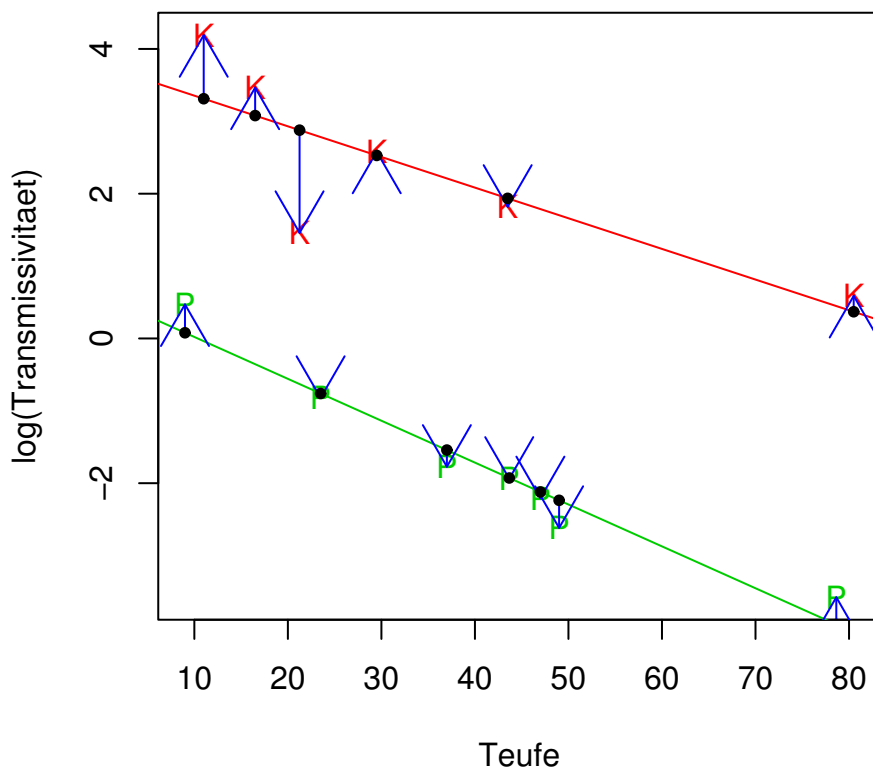
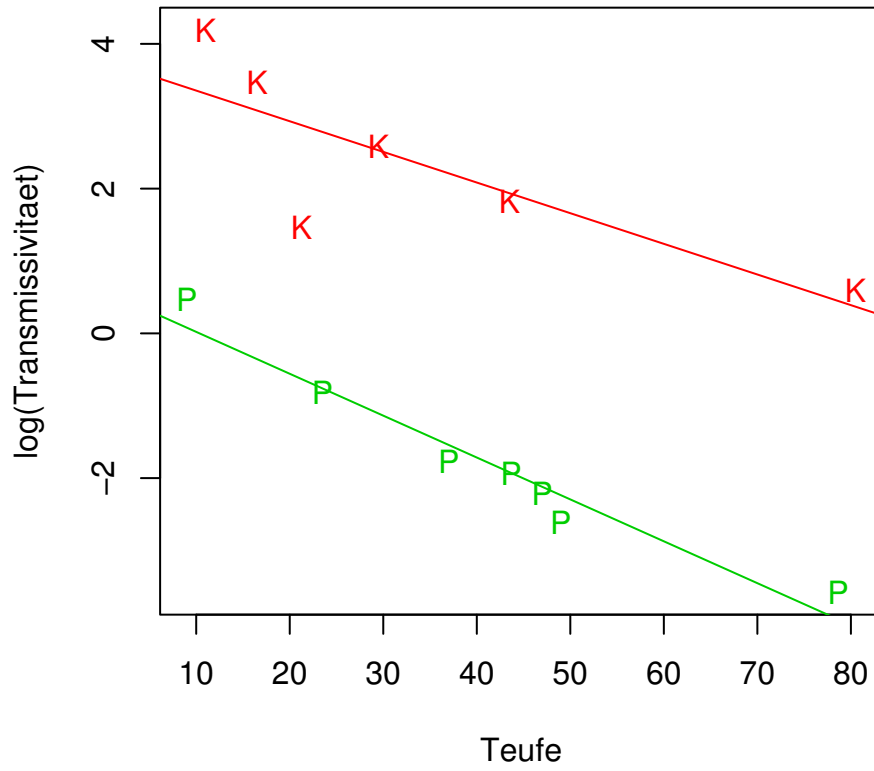
4.3.2.13 Faktor-Regressorinteraktion

Folgende äquivalente Ideen führen zu Interaktion von Faktoren und Regressoren

- Der Anstieg ist in den verschiedenen Gruppen unterschiedlich.
- Der Einfluss des Faktors ändert sich als lineare Funktion des Regressors.

$$y = a + bx + c_k + d_k x + \varepsilon$$

(Intercept)	TypePoren	Teufe	TypePoren:Teufe
3.77785	-3.17872	-0.04235	-0.01552



blau: Residuen r , schwarz: Vorhersagen \hat{y}
 Dieses Modell erklärt die Daten sehr genau (mit wenig Zufall).

4.3.2.14 Regressor-Regressorinteraktion

- Idee: Der Einfluss eines Regressors x verändert sich proportional zu dem Wert von z .
- Idee: Der Einfluss eines Regressors z verändert sich proportional zu dem Wert von x .

$$y = a + bx + cz + dxz + \varepsilon$$

z.B.

Wachstum \sim Nährstoffmenge + Temperatur + Nährstoffmenge * Temperatur

4.3.2.15 Ausblick: Zufallseffekte/random-effect-models

Beispiel:

- Wir wollen den Einfluss eines Wirkstoffs M auf die Blutgerinnung untersuchen.
- Dazu haben wir 5 Versuchspersonen Blut abgenommen in kleine Unterproben aufgeteilt, von denen jeweils 2 mit $0\mu\text{g}$, $10\mu\text{g}$ und $20\mu\text{g}$ des Wirkstoffs versetzt werden. Nach zwei Stunden im Wärmeschrank wird die Gerinnungsgeschwindigkeit gemessen G . Insgesamt haben wir also einen Datensatz mit 30 Messungen.
- Wir gehen davon aus, dass die Blutgerinnung zwischen verschiedenen Personen und je nach Tagesform ohnehin schwankt.

$$G_i = a + b_{\text{Dosis}(i)} + c_{\text{Person}(i)} + \varepsilon_i$$

Dieses Modell hat ein paar Nachteile:

- Es erlaubt keine Aussage über die Gerinnung bei einer eventuellen weiteren Person (z.B. Fr. Mayer), deren c_{Mayer} ja nicht bekannt ist.
- Es modelliert nicht, dass die Patienten zufällig ausgewählt wurden und somit die $c_{\text{Person}(i)}$ einer Verteilung mit Mittelwert und Varianz genügen.

Lösung:

Einführung eines zufälligen Effekts

$$\begin{aligned} G_i &= a + b_{\text{Dosis}(i)} + \varepsilon_{\text{Person}(i)} + \varepsilon_i \\ \varepsilon_{\text{Person}(i)} &\sim N(0, \sigma_p^2) \\ \varepsilon_i &\sim N(0, \sigma_r^2) \end{aligned}$$

Vorteile:

- σ_p^2 kann aus den Daten geschätzt werden.
- Mit $\hat{G}_{n+1} := \hat{a} + \hat{b}_{\text{Dosis}(n+1)}$ kann eine Vorhersage für einen neuen Patienten gemacht werden. Die Genauigkeit ist dann allerdings nur $\sigma_p^2 + \sigma_r^2$ plus den Schätzfehler der Parameter.

Treten im gleichen Modell **Zufallseffekte** (random effects) und gewöhnliche **feste Effekte** (fixed effects) auf, so spricht man von einem Modell mit gemischten Effekten (mixed effects model).

4.3.2.16 Aufsteigende Modellsequenzen

Ein lineares Modell wird aus mehreren Termen der Modellgleichung aufgebaut, die schrittweise hinzugefügt werden. Das gibt eine aufsteigende Folge von Modellen:

$$\begin{aligned}
 M_0 : & & y &= a + \varepsilon \\
 M_1 : & & y &= a + b\text{Teufe} + \varepsilon \\
 M_2 : & & y &= a + b\text{Teufe} + c_{\text{Type}} + \varepsilon \\
 M_3 : & & y &= a + b\text{Teufe} + c_{\text{Type}} + d_{\text{Type}}\text{Teufe} + \varepsilon \\
 & \vdots & & \\
 M_\infty & & y &= a_i
 \end{aligned}$$

Aufsteigend in dem Sinne: Das Modell M_m ist einfacher als das Modell M_l weil es weniger Parameter hat.

Frage: Ist es nötig das komplizierte Modell anzunehmen oder genügt das einfachere.

4.3.2.17 Anova-Tabellen I

Früher wurden diese Berechnungen in ANOVA-Tabellen durchgeführt:

Term	SS Sum of Squares	df Freiheitsgrade	MSS Mean SS	F	p
				F	Quantil
Teufe	$\ (H_1 - H_0)y\ ^2$	$\text{rang } H_1 - \text{rang } H_0$	SS/df	$MSS/MRSS$	$1 - F_{F,df,df_r}^{-1}(F)$
Type	$\ (H_2 - H_1)y\ ^2$	$\text{rang } H_2 - \text{rang } H_1$	SS/df	$MSS/MRSS$	$1 - F_{F,d,df_r}^{-1}(F)$
Type*Teufe	$\ (H_3 - H_2)y\ ^2$	$\text{rang } H_3 - \text{rang } H_2$	SS/df	$MSS/MRSS$	$1 - F_{F,d,df_r}^{-1}(F)$
Residuen	$\underbrace{\ (H_\infty - H_2)y\ ^2}_{RSS}$	$\underbrace{\text{rang } n - \text{rang } H_3}_{df_r}$	$\underbrace{RSS/df_r}_{MRSS}$		

hierbei bezeichnet H_i die Matrix, welche y auf die vom Modell i für y angepassten Werte abbildet. Praktisch z.B.:

	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		

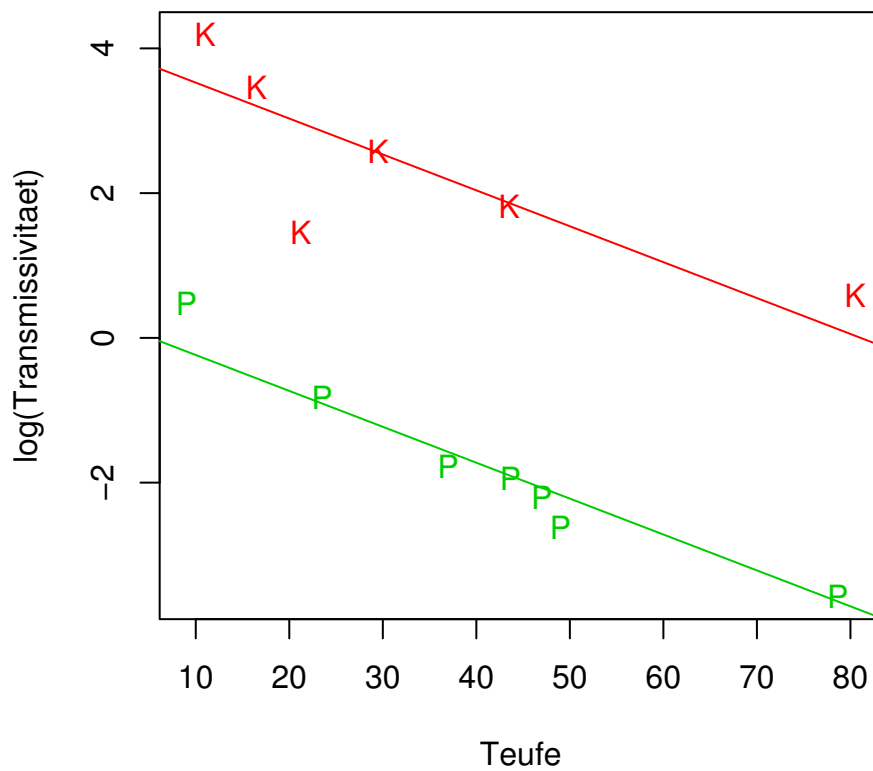
4.3.2.18 Auswertung

	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		

- Die Haupteffekte sind signifikant.
- Die Interaktion ist nicht signifikant. Sie verkompliziert das Modell also unnötig.
- Wir wählen ein neues Modell ohne die Interaktion.

	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		

- Warum verändern sich die Signifikanzen? (Anderes MRSS, Varianz der Residuen genauer geschätzt)
- Der Einfluss der anderen beiden Parameter ist statistisch signifikant nachgewiesen.



```
(Intercept)      Teufe  TypePoren
4.02234151 -0.04960366 -3.76299388
```

$$\log T \sim \text{Teufe} + \text{Type}$$

$$\log T = 4.02234151 - 0.04960366 * \text{Teufe} - 3.76299388 \delta_{\text{Poren}}(\text{Type}) \pm 0.387$$

4.3.2.19 Beispiel: Körpergrösse

Response: Groesse

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Geschlecht	1	1018.86	1018.86	161.1262	< 2e-16 ***
Lage	1	33.09	33.09	5.2329	0.02383 *
Geschlecht:Lage	1	6.58	6.58	1.0413	0.30948

Weglassen des nichtsignifikanten Parameters

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Geschlecht	1	1018.86	1018.86	161.0739	< 2e-16 ***
Lage	1	33.09	33.09	5.2312	0.02384 *
Residuals	127	803.32	6.33		

--- bzw.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Lage	1	122.39	122.39	19.349	2.28e-05 ***
Geschlecht	1	929.55	929.55	146.956	< 2.2e-16 ***
Residuals	127	803.32	6.33		

- Problem: Signifikanz ist von der Reihenfolge abhängig (Geometrische Interpretation an der Tafel)
- Lösungsmöglichkeit: Partielle Tests: Parameter für Test immer als letzten zufügen.
- Nachweis der Lageabhängigkeit auf 1%-Niveau nicht erbracht.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Geschlecht	1	1018.86	1018.86	155.92	< 2.2e-16 ***
Residuals	128	836.41	6.53		

4.3.2.20 Beispiel: Körpergrösse

Response: Groesse

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Geschlecht	1	1018.86	1018.86	161.1262	< 2e-16 ***
Lage	1	33.09	33.09	5.2329	0.02383 *
Geschlecht:Lage	1	6.58	6.58	1.0413	0.30948

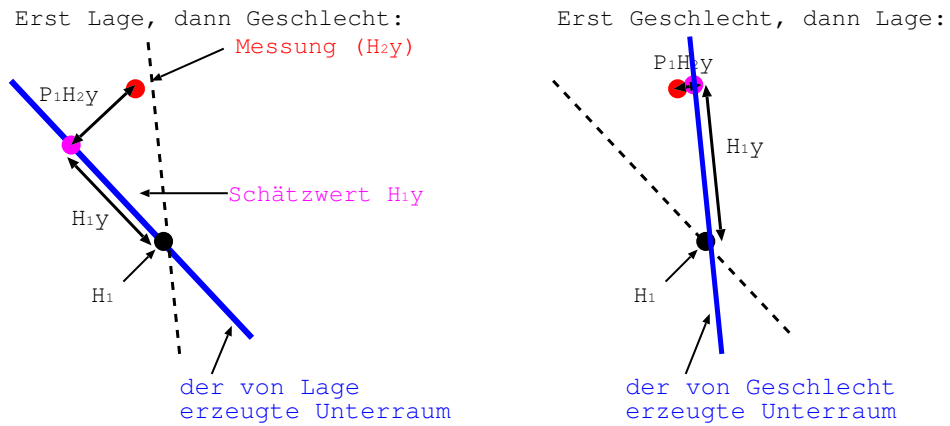
Weglassen des nichtsignifikanten Parameters

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Geschlecht	1	1018.86	1018.86	161.0739	< 2e-16 ***
Lage	1	33.09	33.09	5.2312	0.02384 *
Residuals	127	803.32	6.33		

--- bzw.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Lage	1	122.39	122.39	19.349	2.28e-05 ***
Geschlecht	1	929.55	929.55	146.956	< 2.2e-16 ***
Residuals	127	803.32	6.33		

- Problem: Signifikanz ist von der Reihenfolge abhängig.



Wären die beiden Räume senkrecht, dann wäre das kein Problem!
Ein solches orthogonales Design heißt auch balanciert.

- Lösungsmöglichkeit: Partielle Tests: Parameter für Test immer als letzten zu-fügen (d.h. die kürzeste Strecke wählen).
- Nachweis der Lageabhängigkeit auf 1%-Niveau nicht erbracht.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Geschlecht	1	1018.86	1018.86	155.92	< 2.2e-16 ***
Residuals	128	836.41	6.53		

4.3.3 Wiederholung: Modellvergleich

```
> Aq.lm <- lm(logT~Teufe+Type+Teufe*Type,data=Aqui)
> coef(Aq.lm)
Call:
lm(formula = logT ~ Teufe + Type + Teufe * Type, data = Aqui)
```

Coefficients:

(Intercept)	Teufe	TypePoren	Teufe:TypePoren
3.77785	-0.04235	-3.17872	-0.01552

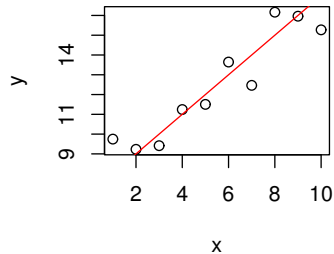
```
> anova(Aq.lm)
anova(Aq.lm)
Analysis of Variance Table
```

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		

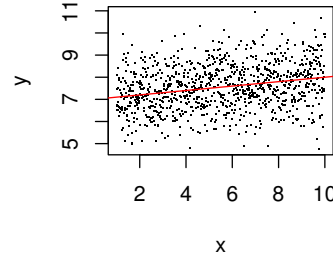
4.3.4 Erklärungskraft des Modells

Was unterscheidet die beiden folgenden Regressionsmodelle?



p-value= 8.1e-05 R^2= 0.87

Zusammenhang signifikant nachgewiesen
Einfluß sehr bedeutend



p-value= 1.24e-15 R^2= 0.062

Zusammenhang signifikant nachgewiesen
Einfluß unbedeutend

Gesucht: Eine Größe, welche die Bedeutung des Einflusses beschreibt (z.B. Correlation)

4.3.4.1 Das Bestimmtheitsmaß R^2

$$R^2 := \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{SS}{TSS} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2} = \frac{SS}{SS + RSS} \quad (= \rho^2)$$

$$SS := \sum_i (\hat{y}_i - \bar{y})^2 = \|H(y - \bar{y})\|^2 \quad \text{Sums of Squares}$$

$$RSS := \sum_i (y_i - \hat{y}_i)^2 = \|P(y - \bar{y})\|^2 \quad \text{Residual Sums of Squares}$$

$$TSS := \sum_i (y_i - \bar{y})^2 = \|y - \bar{y}\|^2 \quad \text{Total Sums of Squares}$$

$$TSS = RSS + SS$$

$$R^2 = \frac{\hat{\text{var}}_{ML}(\hat{y})}{\hat{\text{var}}_{ML}(y)} = \text{Anteil der erklärten Varianz} \in [0, 1] \quad (\text{Betaverteilt unter } H_0)$$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\hat{\text{var}}_{ML}(r)}{\hat{\text{var}}_{ML}(y)} = 1 - \text{Anteil der residuellen Varianz}$$

$$F = \frac{\frac{1}{p-1} SS}{\frac{1}{n-p} RSS}$$

- Die Streuung der Vorhersagen und die Streuung der Residuen ergänzen sich zur Streuung des Datensatzes.
- R^2 beschreibt den Anteil der Streuung, die durch das Modell nun nicht mehr durch Zufall sondern durch ein Abhängigkeitsgesetz erklärt wird.
- Die F-Statistik für das Gesamtmodell hängt 1-1 mit R^2 zusammen, wenn man die Anzahl der Parameter und der Daten kennt.
- Wenn man die Anzahl der Daten und Parameter nicht berücksichtigt, ist keine Umrechnung möglich.

4.3.4.2 Das wahre R^2

Angenommen das Modell stimmt und x hat eine Streuung könnten wir definieren:

$$R_w^2 = 1 - \frac{\text{var}\varepsilon}{\text{var}y} = 1 - \frac{\text{var}\varepsilon}{\beta^t \text{var}(x)\beta + \text{var}(\varepsilon)}$$

$$\text{var}(y) = \text{var}(\beta^t x + \varepsilon) = \text{var}(\beta^t x) + \text{var}(\varepsilon) = \beta^t \text{var}(x)\beta + \text{var}(\varepsilon),$$

da nach Voraussetzung $x \perp \varepsilon$ (unabhängig).

Problem: R^2 schätzt das nicht erwartungstreu:

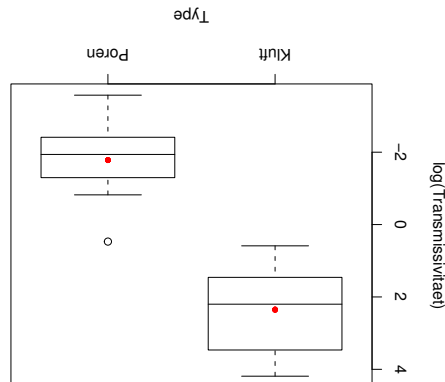
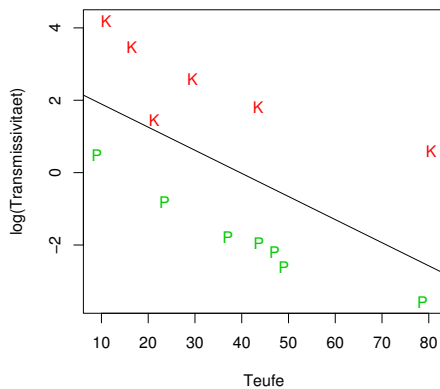
$$E[1 - R^2] = E[]$$

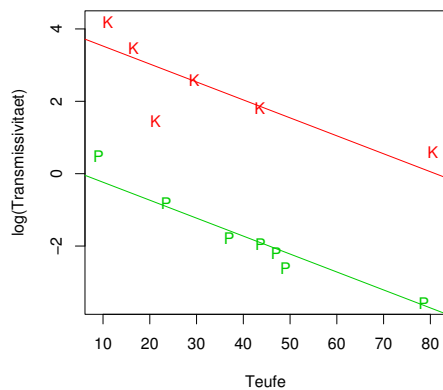
4.3.4.3 R^2 im Einsatz

```
> R2(lm( logT~Teufe           ,data=Aqui))
[1] 0.347679
```

```
> R2(lm( logT~Type           ,data=Aqui))
[1] 0.7438713
```

```
> R2(lm( logT~Teufe+Type     ,data=Aqui))
[1] 0.9477658
```





4.3.4.4 Relatives R^2

$$R_{rel}^2(M_2, M_1) := \frac{\sum_i (\hat{y}_i^{(2)} - \hat{y}_i^{(1)})^2}{\sum_i (y_i - \hat{y}_i^{(1)})^2}$$

```
> R2(lm( logT~Type          ,data=Aqui))
[1] 0.7438713

> R2(lm( logT~Teufe+Type    ,data=Aqui))
[1] 0.9477658

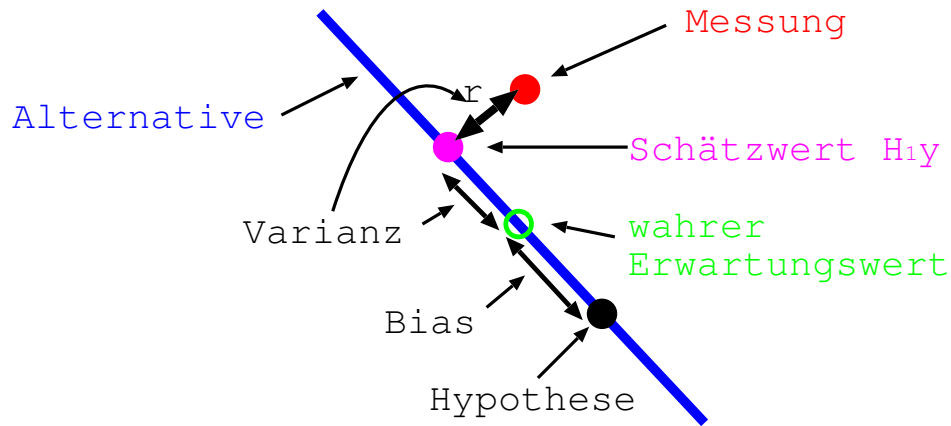
> R2rel(lm( logT~Teufe+Type ,data=Aqui),lm( logT~Teufe ,data=Aqui))
[1] 0.6000868
```

Welcher Anteil der nach Modell M_1 noch übrigen Varianz wird von Modell M_2 erklärt?

4.3.4.5 Probleme mit R^2

```
> zufall1 <- rnorm(length(Aqui$Type))
> zufall2 <- rnorm(length(Aqui$Type))
> zufall3 <- rnorm(length(Aqui$Type))
> zufall4 <- rnorm(length(Aqui$Type))
>
> R2(lm( logT~Teufe+Type      ,data=Aqui))
[1] 0.9477658
> R2(lm( logT~Teufe+Type+zufall1+zufall2+zufall3+zufall4,data=Aqui))
[1] 0.9675746
>
> R2adj(lm( logT~Teufe+Type    ,data=Aqui))
[1] 0.9320955
> R2adj(lm( logT~Teufe+Type+zufall1+zufall2+zufall3+zufall4 ,data=Aqui))
[1] 0.929745
```

Warum R-Quadrat überschätzt wird:



4.3.4.6 Verbesserung durch R_{adj}^2

$$R_{adj}^2 = 1 - \frac{\frac{1}{n-p-1} \sum_i (y_i - \hat{y}_i)^2}{\frac{1}{n-1} \sum_i (y_i - \bar{y})^2}$$

```
> R2( lm( logT~Teufe+Type ,data=Aqui))
[1] 0.9477658
> R2adj(lm( logT~Teufe+Type ,data=Aqui))
[1] 0.9320955
```

R_{adj}^2 ist nicht mehr strikt ansteigend und kann auch negativ werden. Unter der Hypothese ist der Erwartungswert 0.

4.3.4.7 Vergleich: p -Wert und R^2

	signifikant	nicht signifikant
R^2 groß	Einfluss nachgewiesen Bedeutender Einfluss wichtiges Ergebnis	Einfluss nicht nachgewiesen großes R^2 ist Zufall n wahrscheinlich sehr klein
R^2 klein	Einfluss nachgewiesen Einfluss unbedeutend n wahrscheinlich sehr groß	nix Einfluss

4.3.4.8 Konfidenzintervalle für R^2

$$F = \frac{\frac{1}{n-p-1} \text{Summe der Quadrate im Differenzraum}}{\frac{1}{n-1} \text{Summe der Quadrate im Residuenraum}}$$

$$R^2 = 1 - \frac{\text{Summe der Quadrate im Differenzraum}}{\text{Summe der Quadrate im Residuenraum} + \text{Summe der Quadrate im Differenzraum}}$$

Bei bekannten Parameteranzahlen können F und R^2 ineinander umgerechnet werden. Darüber kann man auch Konfidenzintervalle konstruieren.

4.3.5 Modellauswahl

Response: Groesse

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Geschlecht	1	1018.86	1018.86	161.1262	< 2e-16 ***
Lage	1	33.09	33.09	5.2329	0.02383 *
Geschlecht:Lage	1	6.58	6.58	1.0413	0.30948

Problem: Welches Teilmodell ist das richtige?

4.3.5.1 Probleme des sequenziellen Testens in der Modellauswahl

- Die F-Tests der verschiedene Parameter sind positiv korreliert, da die F-Statistiken zwar unabhängige Zähler, aber immer den gleichen Nenner haben.
- Die F-Tests auf den gleichen Einfluss bei verschiedenen Gesamtmodelle können verschieden Aussage treffen, da sich sowohl Nenner als auch Zähler unterscheiden
- Der Nenner-Unterschied ist nicht bedeutend, solange die zusätzlichen Einflüsse nicht vorhanden sind, da der Nenner dann die gleiche Varianz schätzt. Vergessene Parameter können allerdings leicht auch weitere maskieren. Deshalb: alle wichtigen Einflüsse identifizieren und aufnehmen.
- Der Zähler unterscheidet sich nicht bei balancierten Designs (Einflüsse stehen senkrecht), andernfalls kann ein nicht vorher ins Modell aufgenommenen wichtiger Einfluss den Zähler bedeutend vergrößern. Deshalb: Parameter am besten als letzten aufnehmen.

4.3.5.2 Optimalselektion

- **Idee:** Auswahl des Modells mit dem größten R_{adj}^2 , welches nur signifikante Parameter enthält.
- **Problem:** Mit k Einflüssen gibt es 2^{2^k} Modelle.
- **Problem:** Auch ungeschickt aufgeblasene Modelle können durch Weglassen, des entscheidenden Einfluss ein großes R^2 bekommen.

4.3.5.3 Vorwärtsselektion

Algorithmus:

- Beginne mit $y \sim$
- Berechne alle Modelle mit einem zusätzlichen Effekt und wähle das mit dem kleinsten p -Wert aus.
- Wiederhole den letzten Schritt, bis kein signifikanter Effekt mehr gefunden werden kann.

4.3.5.4 Rückwärtsselektion

Algorithmus:

- Beginne mit $y \sim \dots$ Alles...
- Berechne alle Modelle mit einem Effekt weniger und wähle das mit dem größten p -Wert aus.
- Wiederhole den letzten Schritt, bis kein nicht signifikanter Effekt mehr gefunden werden kann.

- Diese Linearkombinationen heißen auch Kontraste.
- Einige Kontraste haben eine wichtige Bedeutung: z.B. Unterschied der Gruppenmittelwerte, Gruppenmittelwerte, Änderung des Anstiegs bei Wechsel von Gruppe a nach Gruppe b usw.
- Es gibt mehr Kontraste als Parameter (unendlich viele)
- Es gibt mehr relevante Kontraste als Parameter (etliche)

4.3.6.3 Problem des multiplen Testens: Notwendigkeit von Post-Hoc-Tests

Frage: Welche Gruppenmittelwerte sind unterschiedlich?

- **Naive Idee:** Verwende paarweise Tests (z.B. two sample t-test)
 - Verwendet ungenaue Schätzer für Varianz der Residuen \Rightarrow die Power des Tests lässt nach.
 - Es werden viele Tests durchgeführt (z.B. $\frac{k}{2}$). Es wäre zu erwarten, dass einige davon zufällig ablehnen.
 - Ergebnisse der Tests sind stochastisch abhängig.
 - Unter Annahme der Unabhängigkeit und keinem Einfluss der Gruppen:

$$P(\text{Mindestens ein Test lehnt (falsch) ab}) = 1 - (1 - \alpha)^{\frac{k}{2}} \approx \frac{k}{2} \alpha$$

- Also würden einige Paarvergleiche fälschlich für signifikant gehalten werden.
- **Idee 1:** Ändere die p-Werte so, dass insgesamt der richtige p-Wert herauskommt.
- **Idee 2:** Definiere eine gemeinsame Konfidenzmenge für Kontraste und betrachte einen Kontrast immer dann als signifikant, wenn 0 nicht im entsprechenden Konfidenzbereich des Parameters liegt.

4.3.6.4 Das Bonferroni-Prinzip zur Korrektur von p-Werten und α -Niveaus beim multiplen Testen

$$\alpha_{\text{einzel}} = \frac{1}{\text{Anzahl Tests}} \alpha_{\text{gesamt}}$$

Begründung:

- Falls die k Tests perfekt negativ korreliert sind, gilt unter der Hypothese:

$$1 - \alpha = P(\text{„Kein Test lehnt fälschlich ab“}) = 1 - \sum_{i=1}^k \alpha_i = 1 - \alpha_{\text{einzel}}$$

- Falls die k Tests unabhängig sind, gilt unter der Hypothese:

$$1 - \alpha = P(\text{„Kein Test lehnt fälschlich ab“}) = \prod_{i=0}^k (1 - \alpha_i) \geq 1 - \underbrace{\sum_{i=1}^k p_i}_{\alpha_{gesamt} \in [1 - \alpha_{gesamt}, 1]} \underbrace{\prod_{j \neq i} (1 - \alpha_j)}_{\in [(1 - \alpha_{gesamt})\alpha_{gesamt}, \alpha_{gesamt}]}$$

Die Worst-Case-Abschätzung ist also auch im üblichen Fall eine gute Abschätzung.

- Falls die k Tests perfekt positiv korreliert sind:

$$1 - \alpha = P(\text{„Kein Test lehnt fälschlich ab“}) = 1 - \alpha_1 = 1 - \frac{1}{n} \alpha_{gesamt}$$

Für positiv korrelierte Tests kann die Abschätzung also sehr konservativ werden.

4.3.6.5 Die Problemstellung der Post-Hoc-Tests

```
> anova(aov(breaks ~ tension, data = warpbreaks))
          Df Sum Sq Mean Sq F value    Pr(>F)
tension    2 2034.3  1017.1   7.2061 0.001753 **
Residuals  51 7198.6   141.1
> model.tables(...)
tension
      L      M      H
8.241 -1.759 -6.481
```

Nachdem sich ein Gruppeneinfluss als signifikant herausgestellt hat (hier die Fadenspannung *tension* eingeteilt in die Gruppen Low, Medium und High als Einfluss auf die Fadenbruchhäufigkeit), möchte man wissen welche Gruppen sich signifikant unterscheiden und in welcher Art.

- Beispiel Idee: Brüche treten nur bei sehr hoher Fadenspannung gehäuft auf. Bei kleinen und mittleren Fadenspannungen wird die Reißfestigkeit des Garns nicht überschritten.

4.3.7 Post-Hoc-Tests

- Diese Tests finden statt, wenn ein anderer Test schon signifikant war (daher post-hoc).
- Es sind immer mehrere.
- Sie sollten aufzeigen, was zur Signifikanz des ersten Tests geführt hat.
- Sie teilen den Unterschied zwischen Hypothese und Alternative in überlappende Teilttestprobleme auf.

4.3.7.1 Einfacher Post-Hoc-Test: Tukeys HSD (Honest Significant Difference)

Bei einem ANOVA-Design mit d Freiheitsgraden, in dem alle Parameterwerte die gleiche Varianz haben (bilanziertes Design), ist unter der Nullhypothese der standardisierte Unterschied zwischen dem kleinsten und größten Wert gegeben durch

die Verteilung der Spannweite von

$$\frac{z_1}{s}, \frac{z_2}{s}, \dots, \frac{z_k}{s}$$

mit

$$z_i \sim N(0, 1), \quad s \sim \sqrt{\chi_d^2}$$

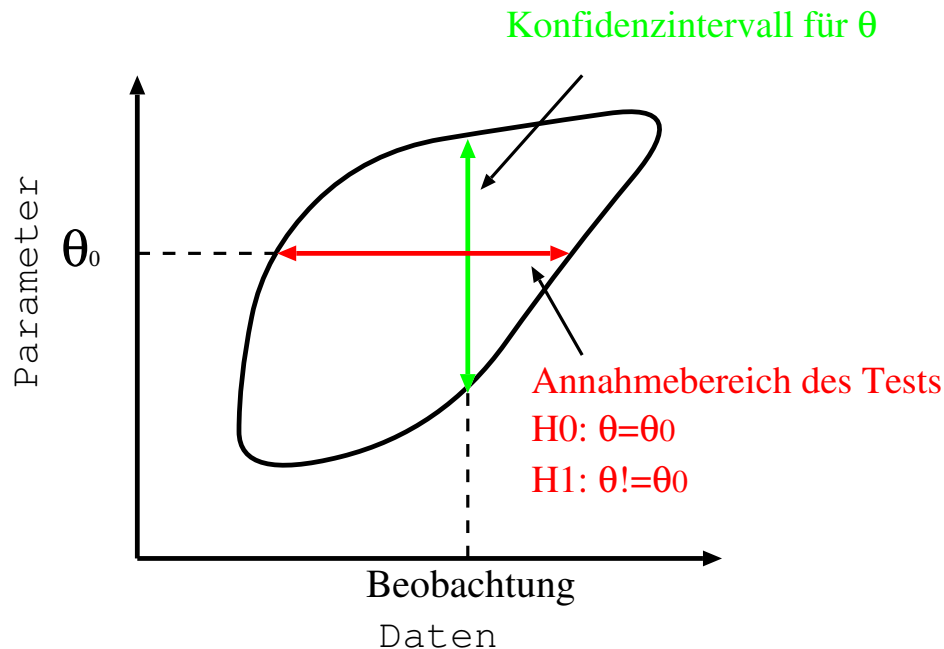
Unterschiede zwischen Klassenmittelwerten sind also erst signifikant, wenn er größer ist, als das $1 - \alpha$ Quantil dieser Verteilung.

Beispiel: Garnbrüche

- Problem Tukey's HSD funktioniert nur, wenn in allen Gruppen gleich viele Beobachtungen vorliegen.

4.3.7.2 Äquivalenz zwischen Konfidenzintervallen und Tests

Äquivalenz von Konfidenzintervallen und Tests



Zu jedem $(1 - \alpha)$ Konfidenzintervall gehört eine Familie von α -Niveau Tests und umgekehrt.

4.3.7.3 Problem der multiplen Konfidenzintervalle

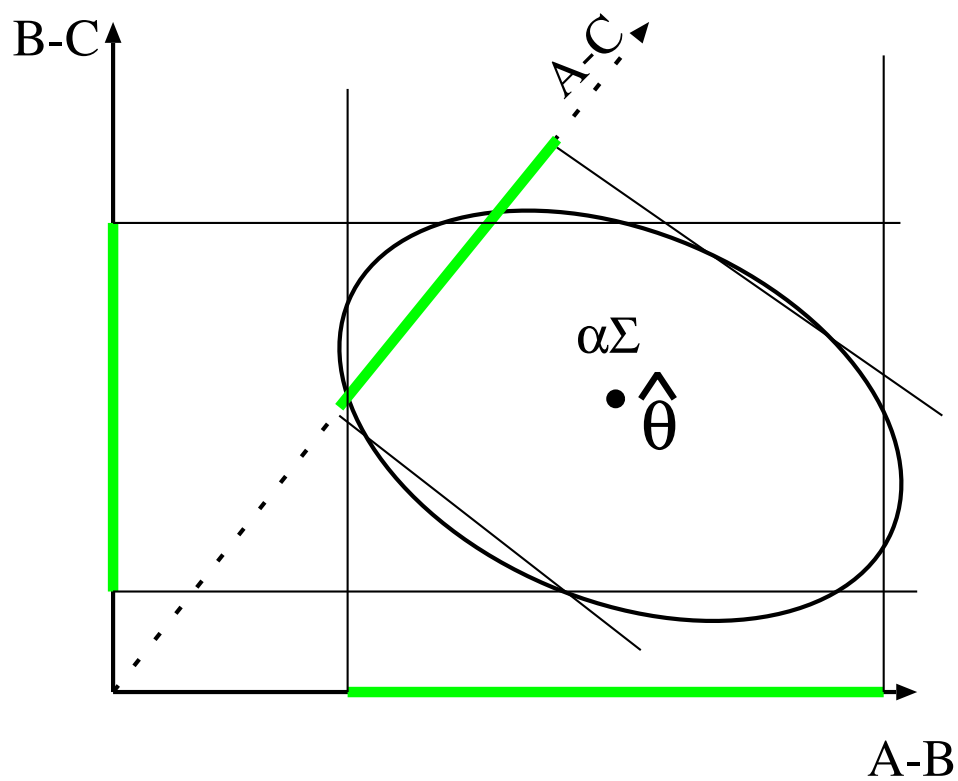
Das Problem des multiplen Testens spiegelt sich bei multiplen Konfidenzintervallen folgendermaßen wieder:

- Werden mehrere Konfidenzintervalle für verschiedene Kontraste angegeben, so erhöht sich die Wahrscheinlichkeit, daß nicht alle Parameter in ihrem jeweiligen Konfidenzintervall liegen.

4.3.7.4 Konfidenzintervalle nach Bonferroni

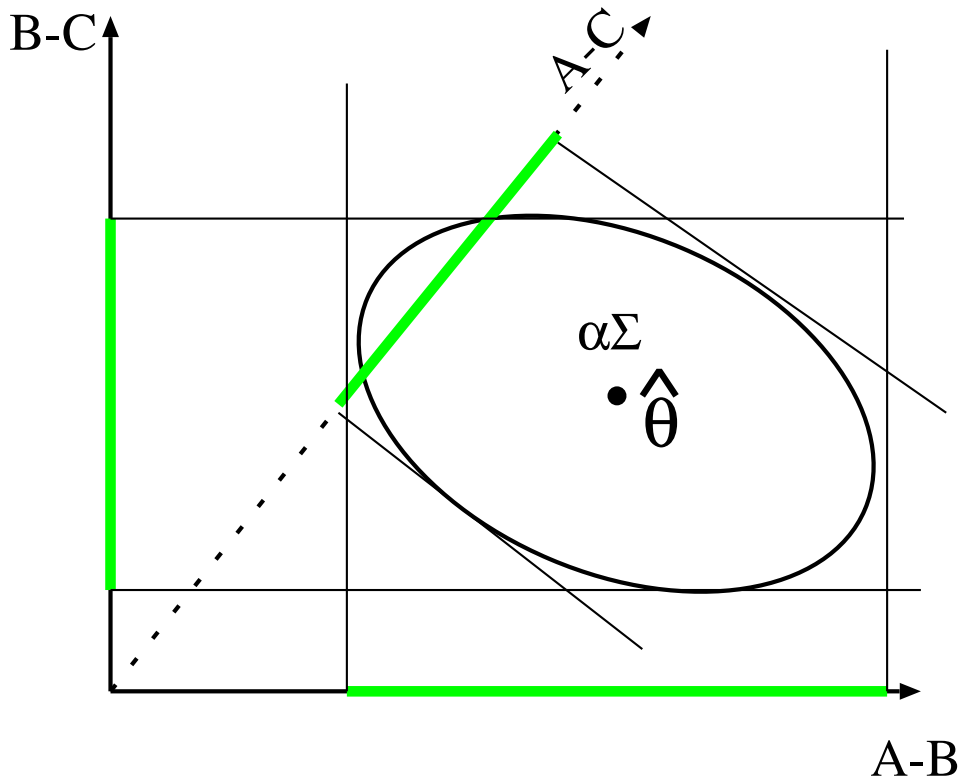
- Idee: Nutze die Äquivalenz von Tests und Konfidenzintervallen aus und verwende einfach $1 - \frac{\alpha}{\text{Anzahl Konfidenzintervalle}}$ Parameter.

4.3.7.5 Konfidenzintervalle im Least Signifikant Difference Schema



Für jeden Kontrast $c^t\beta$ wird als Konfidenzintervall $(c\hat{\beta} \pm a\sqrt{c^t\hat{\Sigma}c})$ angegeben mit $a \in \mathbb{R}^+$ so dass, das gemeinsame Konfidenzlimit für alle durchgeführten Kontrastschätzungen genau eingehalten wird.

4.3.7.6 Simultane Konfidenzintervalle nach Scheffé



Für jeden Kontrast $c^t \beta$ wird als Konfidenzintervall $(c\hat{\beta} \pm a \sqrt{c^t \hat{\Sigma} c})$ angegeben mit $a \in \mathbb{R}^+$ so dass, das gemeinsame Konfidenzlimit für alle möglichen Kontrastschätzungen $\{c^t \beta : c \in \text{im } X\}$ genau eingehalten wird.

Beispiel: Kuckuckseier

4.4 Regressionsdiagnostik

4.4.1 Hebelwirkungen und Cook-Distanzen

Problem: Einzelne Beobachtungen mit extremen Faktor- und Regressorkombinationen können die Ergebnisse eines linearen Modells sehr stark beeinflussen, aber nur in einer idealen Welt stimmen die Fehlergrößenordnungen für alle Beobachtungen. Dazu benötigt man diagnostische Werkzeuge um Irreführung durch einzelne Beobachtungen zu vermeiden.

4.4.1.1 Hebelwirkung/leverage

Frage: Wie stark beeinflusst diese Beobachtung bei dieser Faktor-Regressorkombination potentiell ihre eigene Vorhersage:

$$\hat{y} = Hy = \begin{pmatrix} \ddots & & \ddots & & \ddots \\ h_{l1} & \cdots & h_{li} & \cdots & h_{ln} \\ \ddots & & \ddots & & \ddots \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}$$

$h_{ii} \in [0, 1]$ zeigt mit welchem Faktor der Wert selbst in die Vorhersage eingeht. Werte mit größer Heblewirkungen können die Ergebnisse stark beeinflussen.

4.4.1.2 Cook-Distance/Einfluss

Selbst wenn ein Wert eine große Hebelwirkung hat, muss er die nicht einsetzen, z.B. weil er nahe an den ohnehin aus den anderen Werten vorhergesagten Werten liegt. Durch wechselseitiges Weglassen der Punkte kann man den effektiven Einfluss ermitteln. In einer Approximation erster Näherung kann man das auch über die unmittelbar zu brechende Cook-Distance ausdrücken:

$$c_i = \frac{h_{ii}}{(n-p)} \frac{(y_i - \hat{y}_i)^2}{\hat{\sigma}^2(1-h_{ii})^2}$$

4.4.2 Robuste Regression

Problem: Die Parameter des Modells können durch Ausreißer stark verfälscht werden. Oft ist es dann sogar schwierig anhand der Residuen die Ausreißer zu erkennen. Idee: Man sucht eine Parameterkombination, so dass ein Anteil von $1-p$ der Daten einen möglichst kleinen quadratischen Abstand von den vorhergesagten Werten hat. Auf diese Weise kann ein Anteil bis p von falschen Daten die geschätzte Gerade nicht beliebig verfälschen. Eine Vorgehensweise mit dieser Eigenschaft nennt man robuste Regression. Robuste Regression ist deutlich rechenaufwendiger und komplizierter. Sie sollte aber zumindest eingesetzt werden, um mögliche Ausreißer zu erkennen.