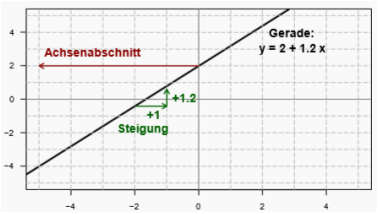
STATISTISCHES MODELLIEREN

LINUS STUHLMANN

FRAGESTELLUNG ZUR MODELLIERUNG

1. Beschreibung
2. Vorhersage, Prognose
3. Schätzung von Parametern (Inferenz)
4. Bestimmung von kausalen Einflussgrössen

EINFACHE LINEARE REGRESSION



$$E(Y|x) = \bar{y} = \alpha + \beta * x$$

- α: Achsenabschnitt
- β: Steigung

ABWEICHUNGEN

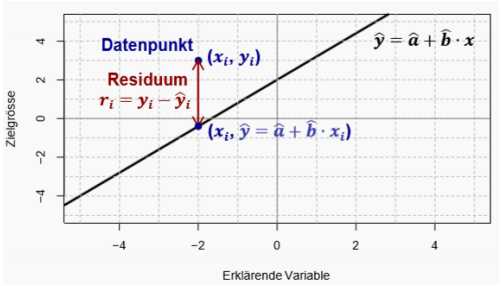
Datenpunkte streuen zufällig um die Regressionsgerade herum, d.h. es gibt (zufällige) Abweichung:

$$Y_i = \alpha + \beta x_i + E_i, \qquad \forall\, i = 1, \dots, n$$

- Y_i: Zielvariable
- x_i: erklärende Grösse
- α, β: Regressionskoeffizienten
- E_i: zufällige Abweichung der Beobachtung zur Geraden.

ANPASSUNG DER GERADEN

Ziel ist es eine Gerade zu finden, dass die Abweichung r_i zwischen den Datenpunkten (x_i, y_i) möglichst klein ist.



- Residuum, $r_i = y_i - \hat{y}_i$
  - Residuen sind Schätzungen für den wahren Fehler ϵ_i

KLEINSTE QUADRATE

Die Regressionsgerade für $E(Y, x)$ wird so durch die Punktwolke gelegt, dass die Summe der **quadrierten** Abweichung r_i minimal ist.

Wir minimieren also:

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (r_i)^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

$$\frac{\partial Q}{\partial \alpha} = 0, \qquad \frac{\partial Q}{\partial \beta} = 0$$

Das daraus resultierende lineare Gleichungssystem mit zwei Unbekannten, nennt sich **Normalengleichung.**

$$\begin{cases} -2 \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i)) = 0 \\ -2 \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i)) x_i = 0 \end{cases}$$

Optimale Lösung gemäss Kleinster Quadrate:

$$\hat{\beta} = \frac{\sum_{m=1}^{M} (x_m - \mu_x)(y_m - \mu_y)}{\sum_{m=1}^{M} (x_m - \mu_x)^2} = \frac{\tilde{s}_{xy}}{\tilde{s}_x^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\, \bar{x}$$

Oder:

$$\theta = (X^T X)^{-1} X^T y$$

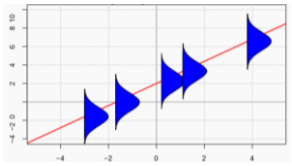
Wobei θ der Parametervektor ist.

EIGENSCHAFTEN

- Die Summe der Residuen addiert sich zu Null

$$\begin{aligned} \sum_{i=1}^{n} r_i &= \sum_{i=1}^{n} [y_i - (\hat{\alpha} + \hat{\beta} \cdot x_i)] = \sum_{i=1}^{n} \left[y_i - \left((\bar{y} - \hat{\beta}\bar{x}) + \hat{\beta} x_i \right) \right] \\ &= (\sum_{i=1}^{n} y_i) - \sum_{i=1}^{n} (\bar{y}) + \sum_{i=1}^{n} (\hat{\beta}\, \bar{x}) - \sum_{i=1}^{n} (\hat{\beta}\, x_i) \\ &= n\bar{y} - n\bar{y} + n\hat{\beta}\bar{x} - n\hat{\beta}\bar{x} = 0 \end{aligned}$$

- Die Gerade geht durch den Daten-Schwerpunkt
  - Die Regressionsgerade geht immer durch den Punkt $(\bar{x}, \bar{y})$.
- Die Abweichungen sollten unabhängig und normalverteilt sein (IID)
  - $E_i \sim \mathcal{N}(0, \sigma^2)$



MIT R

```
fit <- lm(Zeit ~ Menge, data = ga) # Regressionsfit
coef(fit)

(Intercept)          Menge
   3.320780       2.176167
```

FEHLERVARIANZ (RESIDUENVARIANZ)

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

Der Standardfehler der Residuen $\hat{\sigma}$ gibt an, wie stark die Beobachtungen um die angepasste Regressionsgerade streuen. n = Beobachtungen, k = Parameter im Modell.

- Wenn $E_i \sim \mathcal{N}(0, \sigma^2)$ kann angenommen werden, dass 95% der Punkte im Intervall $\pm 2\hat{\sigma}$ liegen.

IN R

```
fit <- lm(Zeit ~ Menge, data = ga)
summary(fit)$sigma

[1] 4.181397
```

INFERENZ UND VORHERSAGE

DETERMINISTISCHE VS. PROBABILISTISCHE REGRESSIONSMODELLE

- deterministische Regressionsmodelle
  - **exakte Beziehung** zwischen unabhängiger und abhängiger Variable
  - keine Abweichung miteinbezogen

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

- probabilistische Regressionsmodelle
  - integrieren Unsicherheiten/ Zufälle im Modell
  - Abhängige Variable → Wahrscheinlichkeitsverteilung

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

PROBABILISTISCHE MODELLE

Probabilistische Modelle können auf zwei Arten beschrieben werden, sagen aber dasselbe aus:

1. $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$
2. $(Y|X = x) \sim \mathcal{N}(\underbrace{\beta_0 + \beta_1 x}_{\mu(x)}, \sigma^2)$

MAXIMUM LIKELIHOOD

RECAP ML

Maximum-Likelihood-Estimation (ML) ist eine Methode zur Schätzung der Parameter eines statistischen Modells, indem die Parameter so angepasst werden, dass die Wahrscheinlichkeit der beobachteten Daten unter dem Modell maximiert wird.

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

IN PROBABILISTISCHEN MODELLEN

- MLE verwendet die beobachteten Daten, um die Parameter des Modells zu finden, die die gemeinsame Wahrscheinlichkeitsverteilung der Daten maximiert.
- Normalverteilung wird angenommen.

SCHÄTZEN VON μ

Bei **konstantem** σ ist die Log-Likelihood für den Parameter μ der Normalverteilung proportional mit dem MSE und kann, als äquivalent betrachtet werden.

$$\mathbf{w}_{\mathrm{ML}} = \underset{w}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mu_{x_i})^2}{2\sigma^2}}$$
$$= \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} -\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{(y_i - \mu_{x_i})^2}{2\sigma^2}$$

Negative Log-Likelihood (NLL)

$$\min \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \Leftrightarrow \min \frac{1}{n}\sum_{i=1}^{n}(\mu_{x_i} - y_i)^2$$

Daraus folgt:

- μ aus $\mathcal{N}(\underbrace{\beta_0 + \beta_1 x}_{\mu(x)}, \sigma^2)$ kann durch Minimieren des MSE bestimmt werden.

SCHÄTZEN VON σ^2

$$\widehat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \mu_{x_i})^2$$

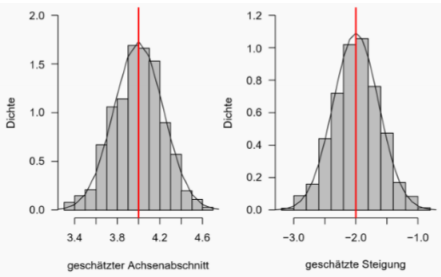
- $n-2$ weil μ auch geschätzt ist (mehr Freiheitsgrade)

oder:

$$\min\left\{\sum_{i=1}^{n} -\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{(\mu_{x_i} - y_i)^2}{2\sigma^2}\right\}$$

EIGENSCHAFTEN DER GESCHÄTZTEN PARAMETER

- Schätzer von Parametern sind erwartungstreu
- Präzise Schätzungen erhält man durch:
  - Grosses n
  - Eine informative erklärende Variable



Die theoretische Verteilung eines linearen Regressionsmodells, sieht wie folgt aus.

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)\right), \qquad \hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{SS_x}\right)$$

- β_0: der wahre y-Achsenabschnitt der Population
- β_1: der wahre Steigungskoeffizient
- σ^2: die wahre Varianz der Residuen
- SS_x: die Summe der quadrierten Abweichung der x_i-Werte von ihrem Mittelwert $\bar{x}$

INFERENZ IN DER REGRESSION

PARAMETERSCHÄTZUNG

- Plausibelste Werte für unbekannte Parameter ist die Punktschätzung.

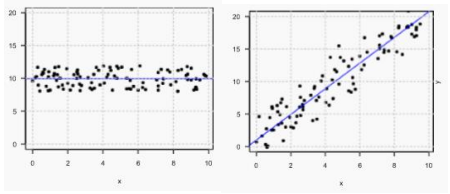
$$\hat{\beta} = \frac{\sum_{m=1}^{M}(x_m - \mu_x)(y_m - \mu_y)}{\sum_{m=1}^{M}(x_m - \mu_x)^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

HYPOTHESENTESTS ÜBER ZUSAMMENHANG

- hat die erklärende Variable einen signifikanten Einfluss auf die Zielgrösse?

$$H_0: \beta = 0, \qquad H_A: \beta \neq 0$$



Der **T-Test** überprüft, ob ein einzelner Regressionskoeffizient signifikant von null verschieden ist.

- Hypothesen $H_0: \beta = 0, H_A: \beta \neq 0$
- Signifikanzniveau: 0.05
- Teststatistik: $T = \frac{\hat{\beta}}{\sqrt{\hat{\sigma}^2/SS_X}}$
- Teststatistik T hat eine t-Verteilung ($n-2$ Freiheitsgrade)
- p-Wert = $P(T > |t|)$
- Testentscheid: Wird H_0 verworfen (p-Wert < 0.05) wird Zusammenhang als statistisch gesichert betrachtet.

Für y-Achsenabschnitt kann derselbe Test angewendet werden, allerdings kann damit nur bestimmt werden, ob der y-Achsenabschnitt durch den Ursprung verläuft (p-Wert > 0.05) oder nicht.

```
summary(fit.ga)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.32078    1.37107   2.422  0.02372 *
Menge        2.17617    0.12403  17.546 8.22e-15 ***
```

KONFIDENZINTERVALL IN REGRESSION

```
predict(fit, input, interval = "confidence", level = 0.99)
```

Die Vorhersage $\hat{y} = \hat{\alpha} + \hat{\beta} * x_0$ ist eine Schätzung. Das Vertrauensintervall für den erwarteten Mittelwert der Regression für den Input x_0 gibt an, wo wir erwarten, dass der **wahre Mittelwert der abhängigen Variablen** liegt.

$$\beta_0 + \beta_1 x_0 \pm qt_{0.975;n-2} * \hat{\sigma} * \sqrt{\frac{1}{n} + \frac{(x_0 - x)^2}{\sum_{i=1}^{n}(x_0 - x_i)^2}}$$

VORHERSAGEINTERVALL

```
predict(fit, input, interval = "prediction", level = 0.99)
```

Das Vorhersageintervall schätzt den Bereich, in dem **zukünftige Einzelbeobachtungen y** mit einer gegebenen Wahrscheinlichkeit liegt.

- Ist breiter, da Unsicherheit bei der Schätzung des Mittelwerts, sondern auch die individuelle Varianz der Daten berücksichtigt wird
- Einzelbeobachtung ist normalverteilt und kann um das Mittel schwanken.

$$\beta_0 + \beta_1 x_0 \pm qt_{0.975;n-2} * \hat{\sigma} * \sqrt{1 + \frac{1}{n} + \frac{(x_0 - x)^2}{\sum_{i=1}^{n}(x_0 - x_i)^2}}$$

RESIDUENANALYSE

BESTIMMTHEITSMASS R^2

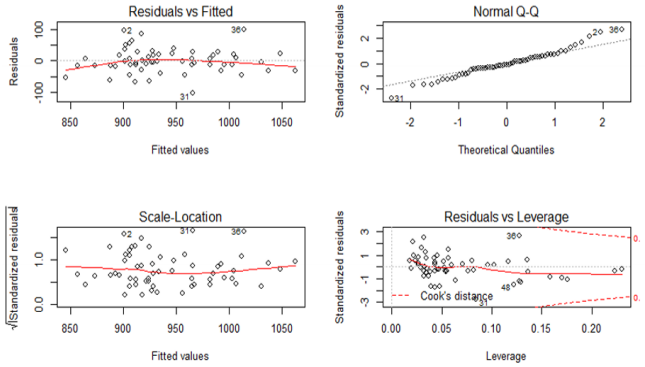
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Vergleicht die Summe der quadrierten Residuen zum Mittelwert mit der Summe der quadrierten Residuals zur Regressionslinie. Gibt die Güte des Modells aus.

RESIDUEN-PLOTS

Es gibt 4 verschiedene Plots

1. Residuals vs. Fitted (Tukey-Anscombe-Plot)
    a. Überprüft die Varianz der Residuen
    b. $E(r) = 0$
    c. Homoskedastizität, nicht-linearität
    d. Je smoother die Linie umso besser
2. Normal Plot (QQ-Plot)
    a. Überprüft ob **Residuen normalverteilt** sind
        i. library(car)
        ii. qqPlot(residuals(fit))
    b. Ausreisser
3. Scale-Location-Plot
    a. Überprüft **konstante Varianz**
    b. Auf x-Achse sind standardisierte Residuen aufgetragen.
    c. Je gerader die horizontale Linie verläuft, umso konstanter die Varianz.
4. Laverage-Plot
    a. Überprüft einflussreiche Punkte mit Cook's Distanz
    b. Gut, wenn Punkte alle innerhalb der Trennlinien liegen.



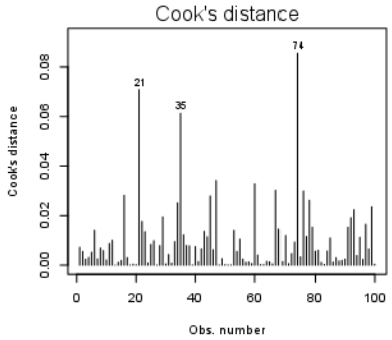
IN R

```
par(mfrow = c(1,3))
plot(fit.ga, which = 1:3) # ohne Simulationen
source("RFn_Plot-lmSim.R")   # mit Simulationen plot.lmSim(fit.ga)
```

COOK'S DISTANZ

Berechnet die potenzielle Veränderung des Modells über alle Werte, wenn ein Datenpunkt x_i ausgelassen würde. **Erkennt Ausreisser**, bzw. Werte die potenziell gefährlich für das Modell sein könnten.

$$D_i = \frac{\sum_{i=1}^{n}\left(\hat{y}^{(-i)}{}_k - \hat{y}_k\right)^2}{(p+1)\sigma_E^2}$$



LÖSUNGSANSÄTZE BEI UNREGELMÄSSIGKEITEN

Wenn die Residuen Analyse Unzulänglichkeiten zeigt:

- **Systematische Abweichung**
    - → Transformationen
- Nicht konstante Varianz
    - Gewichtete Regression
- Ausreisser
    - Robuste Methoden anwenden

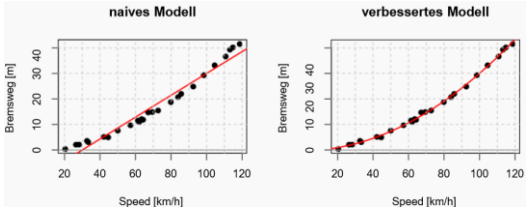
TRANSFORMATIONEN DER ERKLÄRENDEN

Bei nicht-linearen Zusammenhängen kann die erklärende Variable transformiert werden.

BEISPIEL

Bremsweg = $\alpha + \beta * Speed^2 + E_i$

```
bremsweg$speed2 <- (bremsweg$speed)^2
fit.bw2 <- lm(brdist ~ speed2, data = bremsweg)
```



TRANSFORMATIONEN DER ZIELVARIABLE

- Normalität
    - Wenn Residuen des Modells nicht normalverteilt sind
    - Log-Transformation
    - Wurzel-Transformation
- Gegen Heteroskedastizität
    - Logarithmieren der Zielvariable
    - Logarithmieren stabilisiert die Varianz

$$Y' = \log(Y) = \beta_0 + \beta_1 x + E$$

VERBESSERUNG EINES LINEAREN REGRESSIONSMODELL

0) Preprocessing
- learning the meaning of all variables, check for correlations
- give short and informative names
- check for impossible values, errors
- if they exist (missing, error): set them to NA
- consider imputation methods, but be careful

1) First-aid transformations
- bring all variables to a suitable scale (use also field knowledge)
- routinely apply the first-aid transformations

2) Find a good model
- start with a model including important confounders
- perform a residual analysis
- improve model by transformations or adding better predictors
- use your specific knowledge to choose between variables

FIRST AID TRANSFORMATIONEN

Um die Varianz zu stabilisieren, können diese Transformationen auf Y und x_i immer angewandt werden, solange keine praktischen Gründe dagegensprechen (Linearität, keine Verbesserung der Verteilung, …).

- ➔ Für absolute Werte und Konzentrationen
  - $y' = \log(y)$
- ➔ Zähldaten
  - $y' = \sqrt{y}$
- ➔ Proportionen
  - $y' = \arcsin(\sqrt{y})$

Sie bieten folgende Vorteile:

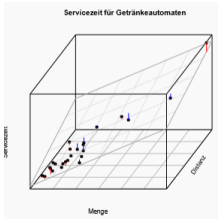
- Verbesserung der Modellannahmen
- Effizientere Schätzungen
- Verringerung der Schiefe
- Umgang mit nicht linearen Beziehungen

MULTIPLE LINEARE REGRESSION

Modell, das mehrere erklärende Variablen beinhaltet, um einen Wert vorherzusagen:

$$Y_i = \beta_0 + \beta_1 X_i^{(1)} + \cdots + \beta_n X_i^{(p)} + E_i$$

- p erklärende Variablen X
- Die Zufallsfehler werden wieder als normalverteilt betrachtet $E_i \sim N(0, \sigma^2)$ in z-Richtung



Ziel → Schätzung der Parameter β_i und σ.

$$\hat{y} = E\left[y \middle| \left(x^{(1)}, \ldots, x^{(n)}\right)\right] = \hat{\beta}_0 + \hat{\beta}_1 x_i^{(1)} + \cdots + \hat{\beta}_n x_i^{(p)}$$

IN R

Geschätze Werte für $\hat{y}_i$:

```
fit.ga <- lm(Zeit ~ Menge + Distanz, data = ga)
fitted(fit.ga)
```

Residuen, $\hat{y}_i - y_i = r_i$: `resid(fit.ga)`

FEHLERVARIANZ (RESIDUENVARIANZ)

Die Fehlervarianz oder auch **Standardfehler** ist Streuung der Residuen. (Erwartungstreue Schätzung)

$$\hat{\sigma}^2 = \frac{1}{n - (p+1)} \sum_{i=1}^{n} r_i^2$$

IN R

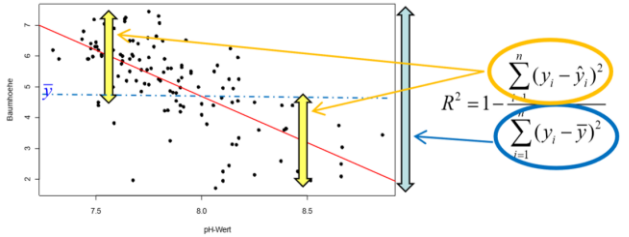
```
fit.ga <- lm(Zeit ~ Menge + Distanz, data = ga)
summary(fit.ga)$sigma
```

```
Residual standard error: 1.008 on 43 degrees of freedom
Multiple R-squared:  0.9269,    Adjusted R-squared:  0.9235
F-statistic: 272.7 on 2 and 43 DF,  p-value: < 2.2e-16
```

BESTIMMTHEITSMASS R^2

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

Vergleicht die Summe der quadrierten Residuen zum Mittelwert mit der Summe der quadrierten Residuals zur Regressionslinie.



ADJUSTIERTES BESTIMMTHEITSMASS R^2_{adj}

R^2 wird durch das Hinzufügen weiterer erklärenden Variablen immer «besser», da durch das Hinzufügen die Residuenquadratsumme SS_E sinkt.

$$R^2_{adj} = 1 - \frac{(n-1)}{(n-(p+1))}(1 - R^2)$$

- n: Stichprobengrösse
- p: Anzahl unabhängiger Variablen
- Besitzt Strafterm für komplexere Modelle
  - Fällt bei kleinem n und grossem p und bei kleinem R^2 hoch aus

F-STATISTIK / F-SCORE

Frage: gibt es einen Zusammenhang zwischen den unabhängigen und der abhängigen Variable?

Hypothesen:

- $H_0: \beta_1 = \beta_1 = \cdots = \beta_p = 0$
- $H_A: \beta_i \neq 0$

F-Statistik:

$$F = \frac{n-(p+1)}{p} \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

TEST GLOBAL

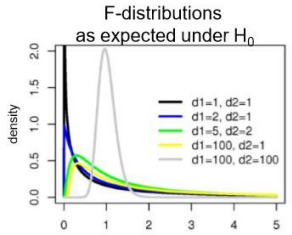
Wenn der F-Wert zu gross ist, wird der p-Wert des F-Tests sehr klein werden und damit wird die Nullhypothese abgelehnt.

→ signifikanter Zusammenhang

- **über alle Variablen**

F-VERTEILUNG

Die F-Verteilung ist abhängig von den Freiheitsgraden $d1$ und $d2$.



- $d1$: Anzahl geschätzte Parameter - 1 bzw. Anzahl unabhängige Variablen – 1.
- $d2$: Anzahl der Beobachtungen x_i minus Anzahl der geschätzten Parameter inkl. β_0.

IN R

```
fit.ga <- lm(Zeit ~ Menge + Distanz, data = ga)
summary(fit.ga)$sigma
```

```
Residual standard error: 1.008 on 43 degrees of freedom
Multiple R-squared:  0.9269,    Adjusted R-squared:  0.9235
F-statistic: 272.7 on 2 and 43 DF,  p-value: < 2.2e-16
```

EINFLUSS DER EINZELNEN VARIABLEN

Sollte über alle unabhängigen Variablen ein signifikanter Zusammenhang zur abhängigen Variable bestehen, kann getestet werden, ob und wie viel die einzelnen Variablen das Modell signifikant verbessern.

- H_0: Zusätzliche Variable hat keinen Einfluss
- H_A: Zusätzliche Variable hat Einfluss

F-Statistik:

$$F = \frac{n-k}{k-j} \frac{R_k^2 - R_j^2}{1 - R_k^2}$$

Kleiner P-Wert und grosse F-Statistik:

➔ Ablehnen von H_0

➔ Signifikante Verbesserung des Modells durch zusätzliche Variable.

IN R

```
fit <- lm(MPI~KPI, data = dat)
fit2 <- lm(MPI~HZ+KPI, data = dat)
anova(fit, fit2)
```

```
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     44 73.881
2     43 43.670  1    30.211 29.748 2.261e-06 ***
```

SUMMARY OUTPUT R

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.70831   10.22588  -6.719 3.29e-08 ***
HZ            2.00691    0.36796   5.454 2.26e-06 ***
KPI           1.61830    0.08921  18.140  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.008 on 43 degrees of freedom
Multiple R-squared:  0.9269,    Adjusted R-squared:  0.9235
F-statistic: 272.7 on 2 and 43 DF,  p-value: < 2.2e-16
```

- **Estimate**: Schätzung der Parameter mittels kleinster Quadrate
- **Std. Error**: Standardabweichung der Residuen Standardfehler
- **t-Value:** Wert der t-Teststatistik
  - **Pr(>|t|)**: p-Wert für Parameter != 0
- **F-statistics**: F-Teststatistik, ob alle Parameter zusammen = 0.

ACHTUNG!

- Je grösser ein Datenset, umso kleiner der p-Wert für gleiche Parameter.
- Ein kleiner p-Wert bedeutet nicht unbedingt, dass Parameter relevant ist.
- Gesamte Steigung des Modells betrachten
- Deswegen muss F-Test / Anova verwendet werden, um die signifikanz einzelner Parameter auf das Modell zu bewerten und nicht nur der t-Test

Wichtiger als p-Werte sind:

- Konfidenzintervalle
  - Wenn KI 0 nicht einschliesst, deutet das auf Signifikanz hin. Muss aber nicht bedeuten, dass Parameter relevant ist.
- Absolute Grösse der Parameter

- (Masseinheiten berücksichtigen)
- Standardisierte unabhängige Variablen benutzen. Somit ist, wenn $\beta_i = -1.5$ der Wert um 1.5 Sd kleiner wenn $x_i + 1$.

MATRIXNOTATION

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1^{(1)} & \cdots & x_1^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^{(1)} & \cdots & x_n^{(p)} \end{pmatrix} * \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} E_1 \\ \vdots \\ E_n \end{pmatrix}$$

$$Y = X\beta + E$$

KLEINSTE QUADRATE SCHÄTZUNG

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} R_i^2$$

$$\sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2$$

$$= \sum_{i=1}^{n} \left(Y_i - \left(\sum_{j=1}^{p} \beta_j * x_{ij}\right)\right)^2$$

$$= (Y - X\beta)^T (Y - X\beta)$$

ABLEITUNG IN MATRIXNOTATION

$$\begin{pmatrix} \frac{\partial}{\partial \beta_0} Q(\beta) \\ \vdots \\ \frac{\partial}{\partial \beta_p} Q(\beta) \end{pmatrix} = \begin{pmatrix} -2 * \left(\sum_{i=1}^{n} 1 * R_i\right) \\ \vdots \\ -2 * \left(\sum_{i=1}^{n} x_i^{(p)} * R_i\right) \end{pmatrix} = -2 * X^T R$$

$$-2X^T R = 0$$

$$\Leftrightarrow X^T\left(Y - X\hat{\beta}\right) = 0$$

$$\Leftrightarrow X^T Y = X^T X\hat{\beta}$$

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

MAXIMUM LIKELIHOOD SCHÄTZER

Es zeigt sich, dass Kleinste Quadrate KQ und Maximum Likelihood ML dieselben Schätzer ergeben.

$$\underline{\hat{\beta}} = \arg\max_{\underline{\beta}} \prod_{i=1}^{n} f(y_i | \underline{\beta}, \sigma^2, \boldsymbol{X})$$

VERTEILUNG VON $\underline{\hat{\beta}}$

$\underline{\hat{\beta}}$ folgt einer $(p+1)$-dimensionalen multivariaten Normalverteilung.

$$\underline{\hat{\beta}} \sim N\left(\underline{\beta}, \sigma^2, (X^T X)^{-1}\right)$$

- Wobei p die Anzahl erklärende Variablen.

Somit gilt für Randdichten:

$$\hat{\beta}_j \sim N\left(\beta, \sigma^2, (X^T X)_{jj}^{-1}\right)$$

INFERENZ

Aus der Verteilung von $\hat{\beta}_j$ lässt sich nun ein Test und Vertrauensintervall ableiten.

- **t-Test:** $T_j = \frac{\widehat{\beta_j} - \beta_{j0}}{\boldsymbol{se}(\widehat{\beta_j})}$
  - $\boldsymbol{se}\left(\widehat{\beta_j}\right) = \hat{\sigma}\sqrt{((X^T X)^{-1})_{jj}}$
- **95% Konfidenzintervall**
  - $\widehat{\beta_j} \pm q * \boldsymbol{se}(\widehat{\beta_j})$
  - $q = qt\left(0.975, n - (p+1)\right)$

TESTS

- Nullhypothese: $H_0: \beta_1 = 2$
- Alternativhypothese: $H_A: \beta_1 \neq 2$

MODELLVIELFALT

Es können vier verschiedenen Typen von erklärenden Variablen verwendet.

- Kontinuierliche Prädikatoren
  - Temperatur, Distanz, …
- Transformierte Grössen
  - $\log(x), \sqrt{x}, \ldots$
- Potenzen
  - $x^{-1}, x^2, \ldots$
- Kategorielle Variablen
  - Geschlecht, Herkunft…

TRANSFORMATIONEN

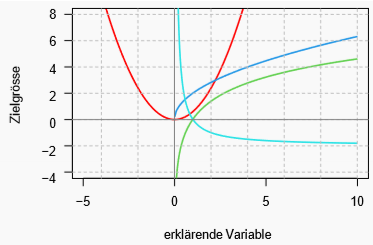
$$\log(Y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + E$$

$$\Rightarrow Y = e^{\beta_0} * e^{\beta_1 \log(x_1)} * e^{\beta_2 \log(x_2)} * e^E$$

$$= e^{\beta_0} * {x_1}^{\beta_1} * {x_2}^{\beta_2} * e^E$$

POLYNOME

Durch Transformation der erklärenden Variablen, kann je nach Grad des Polynoms jede Funktion approximiert werden. $x \to x^2$



REGRESSION MIT KATEGORIELLE VARIABLEN

Kategorielle Variablen sind Variablen, die eine begrenzte Anzahl von diskreten Ausprägungen haben. Für die lineare Regression gilt, es können kategorielle Variablen enthalten sein, solange mindestens eine numerische Variable enthalten ist.

BINÄRE KATEGORIEN

$$X_i^{(1)} = \begin{cases} 0 & \text{falls Kategorie 1} \\ 1 & \text{falls Kategorie 2} \end{cases}$$

Damit kann folgendes Modell aufgestellt werden:

$$Y = \beta_0 + \beta_1 * X_i^{(1)} + \beta_2 * X_i^{(2)} + E_i$$

Daraus ergeben sich folgende Gleichungen:

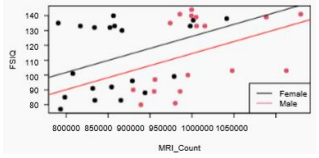
- Für Kategorie 1
  - $Y_i^{(1)} = \beta_0 + \beta_1 * 0 + \beta_2 * X_i^{(2)} + E_i$
- Für Kategorie 2
  - $Y_i^{(2)} = \beta_0 + \beta_1 * 1 + \beta_2 * X_i^{(2)} + E_i$

BEISPIEL

Beispiel IQ zwischen Mann und Frau. $X_i^{(1)}$ repräsentiert die Hirnmasse.

$$X_i^{(1)} = \begin{cases} 0 & \text{falls weiblich} \\ 1 & \text{falls männlich} \end{cases}$$

β_1: ist in diesem Fall «Geschlecht» und hat einen Wert von -11.6. Was bedeutet das der Unterschied unabhängig von der numerischen Variable -11.6 beträgt.



MULTIPLE AUSPRÄGUNGEN

Modellierung Variablen mit multiplen Ausprägungen:

$$d_i^{(j)} = \begin{cases} 1\text{: falls Beobachtung}_i \text{ aus Level}_j \\ 0\text{: sonst} \end{cases}$$

$$Y_i = \beta_0 + \beta_1 * X_i^{(1)} + \beta_2 * d_i^{(2)} + \beta_3 * d_i^{(3)} + \beta_4 * d_i^{(4)} + E_i$$

Daraus entstehen drei Gleichungen, für jede Kategorie eine.

- Kategorie 1:
  - $Y_i^{(1)} = \beta_0 + \beta_1 * X_i^{(1)} + \beta_2 + E_i$
- Kategorie 2:
  - $Y_i^{(2)} = \beta_0 + \beta_1 * X_i^{(1)} + \beta_3 + E_i$
- Kategorie 3
  - $Y_i^{(3)} = \beta_0 + \beta_1 * X_i^{(1)} + \beta_4 + E_i$

Die Koeffizienten $\beta_{i>1}$ der unterschiedlichen Kategorien stellen eine Verschiebung der Zielgrösse dar.

DUMMY-VARIABLE

Dummy-Variable: wird verwendet, um das Vorkommen verschiedener Kategorien zu definieren. In jeder Spalte ist eine Kategorie und in jeder Zeile ist eine Beobachtung.

$$d_i = \overbrace{\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}}^{\text{Kategorien}}$$

- Jedes Modell braucht bei n Kategorien $(n-1)$ Dummy-Variablen

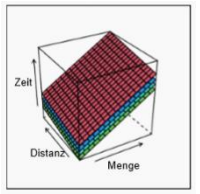
BEISPIEL

```
fit.ga <- lm(Zeit ~ Menge + Distanz + Ort, data = ga)
coef(fit.ga)
```

(Intercept)	Menge	Distanz	OrtBoston	OrtMinneapolis
-0.03638915	1.77027727	0.03611117	4.19027462	0.45263582

OrtSan Diego
2.73773652

Austin, wird nicht im Summary angezeigt, da es in β_0 enthalten ist.



INFERENZ

```
summary(fit.ga)
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.036389	1.754025	-0.0207	0.983665	
Menge	1.770277	0.186790	9.4774	1.241e-08	***
Distanz	0.036111	0.012620	2.8615	0.009987	**
OrtBoston	4.190275	1.749048	2.3957	0.027043	*
OrtMinneapolis	0.452636	2.687420	0.1684	0.868027	
OrtSan Diego	2.737737	1.936269	1.4139	0.173561	

```
Residual standard error: 2.986 on 19 degrees of
 freedom Multiple R-squared: 0.9707 Adjusted R-
 Squared: 0.963
 F-statistic: 125.92 on 5 and 19 DF, p-value:  6.92e-14
```

- Die p-Werte geben lediglich eine Angabe, ob die Orte einen signifikanten Unterschied zum Referenzort aufweisen.
- Deswegen sollte stattdessen getestet werden ob die kategorielle Variable einen signifikanten Einfluss auf das Modell hat.
  - $H_0: \beta_2 = \beta_3 = \beta_4 = 0$
  - Wird H_0 verworfen, hat Variable signifikanten Einfluss auf das Modell

F-TEST

T folgt der F-Verteilung

$$T = \frac{\frac{SS_E^* - SS_E}{q}}{\frac{SS_E}{n-(p+1)}}$$

- $SS_E = \sum_{i=1}^{n} r_i^2$, von Modell mit kat. Variable
- $SS_E^* = \sum_{i=1}^{n} r_i^2$, von Modell ohne kat. Variable

Nullhypothese $H_0: \beta_2 = \beta_3 = \beta_4 = 0$

Alternativhypothese: H_A: min. eine Kategorie hat Einfluss

ANOVA

Vergleicht zwei Modelle miteinander ob es einen signifikanten Unterschied zwischen ihnen gibt, falls eine Variable fehlt.

- `anova(fit1, fit2)`

```
Model 1: y ~ experience
Model 2: y ~ experience + education
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1     28 7496512725
2     26 3379616895  2 4116895830 15.836 3.178e-05 ***
```

DROP1

Überprüft, ob das Modell signifikant schlechter wird, wenn eine Variable entfernt wird.

- `drop1(fit, test="F")`

```
           Df  Sum of Sq        RSS    AIC F value    Pr(>F)
<none>                   3.3796e+09 564.19
experience  1 1.1779e+10 1.5159e+10 607.22  90.618 5.854e-10 ***
education   2 4.1169e+09 7.4965e+09 584.10  15.836 3.178e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

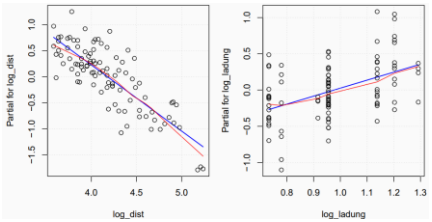
PARTIELLE RESIDUEN

Partielle Residuen-Plots zeigen den **Zusammenhang** zwischen der **erklärenden Variable und der Zielgrösse** unter Berücksichtigung der **anderen Variablen** im Modell.

$$y_i - \sum_{k \neq j} x_i^{(j)} \hat{\beta}_j = \hat{y}_i + r_i - \sum_{k \neq j} x_i^{(j)} \hat{\beta}_j = x_k \hat{\beta}_k + r_i$$

R: `residuals(fit, type = "partial")`

```
termplot(fit.spreng, partial.resid = TRUE, smooth = panel.smooth,
         col.res = "black", col.term = "blue", col.smth = "red")
```



- Ok: angepasste lineare Beziehung (blau), und der rote Glätter keine systematische Differenz haben.
- Falls nicht ok:
  - Transformation der Variablen
    - log , x^2 …
  - Ggf. anderes Modell

INTERAKTION

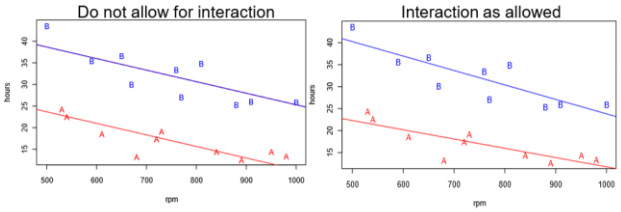
Interaktion von Variablen ist ein mächtiges Tool, das die Interaktion von Variablen zulässt. Das heisst, eine zwei abhängige Variablen sind voneinander abhängig.

KATEGORIELLEN VARIABLEN

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_{kate} + \beta_3 (x_1 * x_{kate})$$

BEISPIEL

Die Anzahl «hours» ist abhängig von «rpm» und «tool». Wir lassen «rpm» und «tool» miteinander interagieren, um nicht nur additive Beziehungen zu modellieren.



`fit1=lm(hours ~ rpm + tool, data=my.dat)`

`fit2=lm(hours ~ rpm * tool, data=my.dat)`

SIGNIFIKANZ

Dadurch bekommen wir eine neue Variable im Summary

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.774760   4.633472   7.073 2.63e-06 ***
rpm         -0.020970   0.006074  -3.452  0.00328 **
toolB       23.970593   6.768973   3.541  0.00272 **
rpm:toolB   -0.011944   0.008842  -1.351  0.19553
```

SIGNIFIKANTE VERBESSERUNG DES MODELLS?

`fit1=lm(hours ~ rpm + tool, data=my.dat)`

`fit2=lm(hours ~ rpm * tool, data=my.dat)`

`anova(fit2, fit1, test="F")`

ZWEI KATEGORIELLEN VARIABLEN

Wenn zwei verschiedenen binäre kategorielle Variablen interagieren, gibt es 4 verschiedene Gleichungen. Die Parameter sind rein additiv.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_{kat_1} + \beta_3 x_{kat_2} + \beta_4 (x_{kat_1} * x_{kat_2})$$

SIGNIFIKANZ

Mit Drop1 kann geprüft werden, ob die zusätzliche Interaktionsvariable einen signifikanten Beitrag zum Modell beiträgt.

```
drop1(fit.ga, test = "F")

Single term deletions

Model:
Zeit - Menge + Distanz + Ort + Generation + Ort:Generation
               Df Sum of Sq   RSS    AIC   F value Pr(>F)
<none>                      1001 270.06
Menge           1     37476 38477 691.29 3966.4681 <2e-16 ***
Distanz         1      3641  4643 445.97  385.3780 <2e-16 ***
Ort:Generation  3        11  1012 265.32    0.3879 0.7619
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

→ Die Interaktion zwischen Generation und Ort ist nicht signifikant

ZWEI NUMERISCHE VARIABLEN

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2)$$

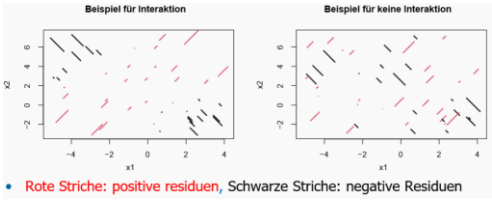
Die Veränderung der Zielgrösse von x_1 hängt von $\beta_1 + \beta_3 x_2$ ab. Das lässt sich auch durch die partielle Ableitung der Funktion nach x_1 zeigen.

$$\frac{\partial \hat{y}}{\partial x_1} = \beta_1 + \beta_3 x_2$$

ÜBERPRÜFUNG INTERAKTION

```
library("sfsmisc")
fit <- lm(y ~ x1 + x2, dat = bsp)
p.res.2x(~ x1 + x2, fit, scol = 2:1) # scol für Farben
```

- Länge des Striches → Absolutbetrag Residuum
- Steigung (+1, -1) → Vorzeichen des Residuums



- Rote Striche: positive residuen, Schwarze Striche: negative Residuen

GEWICHTETE REGRESSION

MOTIVATION

Das multiple Regressionsmodell wird wie folgt beschrieben:

$$Y = X\beta + E$$

Mit **konstanter** Varianz des Fehlers:

$$E \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix}$$

Doch was, wenn die Residuenvarianz nicht konstant ist?
→ **gewichtete Regression**

DEFINITION

$$E \sim N(0, \Sigma), \quad \Sigma = \sigma^2 \begin{pmatrix} 1/w_1 & 0 & 0 \\ 0 & 1/w_2 & 0 \\ 0 & 0 & 1/w_n \end{pmatrix} = \sigma^2 W$$

Die w_i sind Gewichte. Beobachtungen mit grosser Varianz haben kleine Gewichte und mit kleiner Varianz haben grosse Gewichte.

- Unterschiedliche Varianz bei verschiedenen Kategorien.
  - Verschieden präzise Messungen
- Heteroskedastizität
- Zeitreihendaten
  - Unterschiedliche Zeitpunkte können unterschiedlich relevant sein
- Über- und Unterrepräsentation von Gruppen

```
fit <- lm(Fett ~ Art, data = fische, weights=ni)
coef(fit)
```

SIGNIFIKANZ

- Mit drop1 kann geprüft werden, um ein signifikanter Unterschied zwischen verschiedenen Kategorien besteht.
- Vergleich der Modellparameter, mit und ohne Gewichten

FESTLEGEN DER GEWICHTE

Je **relevanter** oder **präziser,** umso **kleiner** soll die **Varianz** sein.

Hohe Gewichte → kleine Varianz $w_i = 1/v_i$

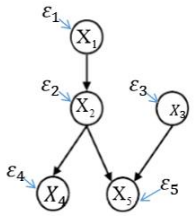
- Optimalfall: Genauigkeit der Beobachtung ist bekannt
  - $var(E_i) = \sigma^2 v_i,\ w_i = 1/v_i$
- Wenn Y_i Mittelwerte von n_i bekannt:
  - $w_i = n_i$
  - Je mehr Beobachtungen umso relevanter, desto höher gewichtet, führt zu tiefer Varianz $(1/v_i)$.
- Aus ungewichteten Regressionen schätzen:

```
# ungewichtete Regression
fit <- lm(y ~ x1 + x2, data = bsp)
# Berechnung der Gewichte
wx <- data.frame(y = abs(resid(fit)), x = fitted(fit))
fit.ei <- lm(y ~ x, data = wx)
bsp$weights <- 1/(fitted(fit.ei)^2)
```

MODELLIERUNG VON KAUSALITÄT

Strukturelle kausale Modelle definieren den datengenerierenden Prozess.

$$X_i \leftarrow h_i(f(X_i), \epsilon_i), \quad \epsilon_i \sim N(\mu, \sigma^2)$$



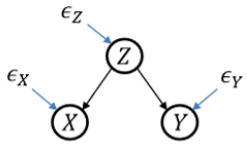
$$X_5 \leftarrow f(X_2) + f(X_3) + \epsilon_5$$

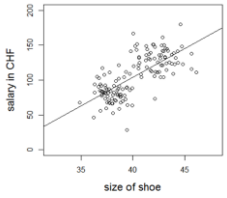
TOTALER KAUSALER EFFEKT

Für die Berechnung des totalen kausalen Effekts von X_1 auf Y, müssen folgende Variablen adjustiert bzw. kontrolliert werden:

CONFOUNDER

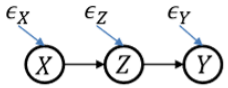
- **Immer** adjustieren, da der Confounder mit der abhängigen und unabhängigen Variable korreliert, muss der Confounder immer adjustiert werden.





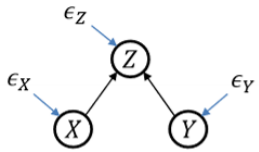
MEDIATOREN

- Mediatoren sollen **nicht adjustiert** werden:
  - Dabei wird der indirekte kausale Effekt eliminiert.



COLLIDER

- Sollten **nie** adjustiert werden, weil dies zu einem Bias führt.



DIREKTER KAUSALIER EFFEKT

Für die Berechnung des direkten kausalen sollte neben dem Confounder auch der Mediator adjustiert werden.

- Mediatoren
  - Dabei wird der indirekte kausale Effekt eliminiert
- Confounder
  - Da der Confounder mit der abhängigen und unabhängigen Variable korreliert, muss der Confounder adjustiert werden.

BACKDOOR KRITERIUM

Um den kausalen Effekt von X auf Y zu erhalten, möchten wir alle Backdoor-Pfade zwischen X und Y schliessen.

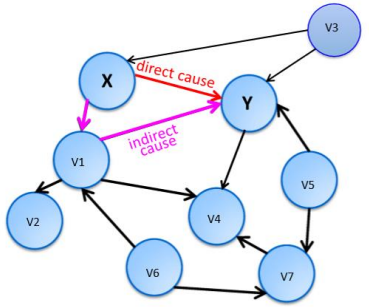
BEISPIEL

Um den **totalen kausalen Effekt** von X auf Y zu bestimmen, müssen wir alle Confounder adjustieren.

- V3
- R: Y~X+V3

Für den direkten kausalen Effekt zu bestimmen, müssen wir Confounder und Mediatoren adjustieren.

- V1, V3
- R: Y~X+V3+V1



PREDICTION PERFORMANCE

TEST/TRAIN SPLIT

In der Statistik ist es üblich die daten in 50% Training und 50% Testdaten zu splitten.

METRIKEN

- MSE
  - Mean Squared Error
  - $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_{xi})^2$
- RMSE
  - Root Mean Squared Error
  - $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_{xi})^2}$
- MAE
  - Mean Absolute Error
  - $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{\mu}_{xi}|$
  -
- MAPE
  - Mean Absolute Percentage Error

TEST NEGATIVE LOG LIKELIHOOD – NLL

RMSE und MAE fangen nicht alle relevanten Informationen von probabilistischen Modellen ein.

Der MSE ist der Log-Likelihood der Normalverteilung gegeben einer konstanten Varianz.

$$NLL(\theta) = -\log(\prod_{i=1}^{n} N(\mu, \sigma^2))$$

$$\arg\min\{\sum_{i=1}^{n} -\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} + e^{-\frac{(y_i-\mu_{xi})^2}{2\sigma^2}}\right)\}$$

$$\arg\min\{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_{xi})^2\}$$

IN R

```
get_nll = function(dat, fit) {
  s = summary(fit)$sigma
  -mean(log(dnorm(dat$y, mean=dat$y_hat, sd=s)))
}
```

Um das Konfidenzintervall der Prediction zu bekommen, kann mit Boot Strap eine künstliche Verteilung simuliert werden. Für ein 0.95 CI wird das 0.025 Quantil und das 0.975 Quantil als Intervallsgrenzen verwendet.

IN R

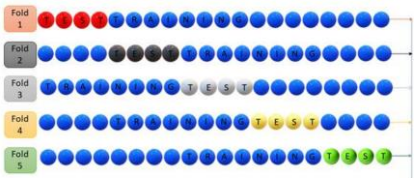
```
# Function to compute nll of a bootstrap sample of observations
nll_fun = function(data, fit, indices) {
  nll_bs = get_nll(data[indices,], fit)
  return(nll_bs)
}
nll_fun(data=iris, fit=fit_iris, indices=1:nrow(iris))
#0.5912817
# Bootstrap to compute confidence interval for NLL
library(boot)
nll_boot = boot(data = iris, fit=fit_iris,
                statistic = nll_fun, R = 5000)

boot.ci(nll_boot, type = "bca")    # (0.5033,  0.7152)
```

KREUZVALIDIERUNG

Wenn zu wenig Daten für einen Train/Test Split vorliegen, kann Kreuzvalidierung helfen.

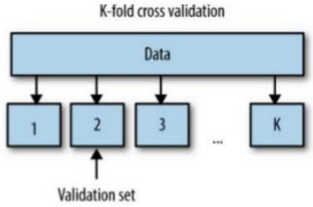
5-FOLD CROSS VALIDATION

Das Modell wird 5-mal gefittet und dabei jedes Mal 1/5 als Testset verwendet. Dabei wird der Mittelwert der Performance über die 5 Durchgänge verwendet, um die Qualität des Modells zu bestimmen.



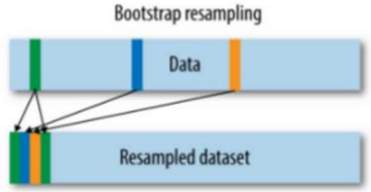
LEAVE-ONE-OUT (N-FOLD)

Dabei wird jede einzelne Beobachtung als Testdatensatz verwendet. Also immer eine Beobachtung ausgelassen und der Rest $(n-1)$ für das Training verwendet. Dabei müssen n Modelle trainiert werden.



BOOTSTRAP RESAMPLING

Es wird immer ein Trainingsset mit Zurücklegen gezogen und das Testset (Out-of-Bag) besteht aus den Samples die nicht gezogen wurden. Das hat den Vorteil zu den vorherigen Methoden, dass weniger «ungünstigen» Daten Splits auftreten.



AIC & BIC

Unter Statistikern ist es nicht üblich Modelle auf Testdaten zu evaluieren. Stattdessen wird die Modellgüte auf AIC/BIC und dem Likelihood auf den Testdaten gemessen.

Informationskriterium von Akaike	$AIC = -2\max(\text{Log-Likelihood}) + 2q$ $= n\log\left(\frac{1}{n}\sum_{i=1}^{n} R_i^2\right) + 2q + \text{konst}$
Bayes Informationskriterium	$BIC = -2\max(\text{Log-Likelihood}) + \log(n)\,q$ $= n\log\left(\frac{1}{n}\sum_{i=1}^{n} R_i^2\right) + \log(n)\,q + \text{konst}$

RÜCKWÄRTSSELEKTION MIT AIC/BIC

Man beginnt mit dem «vollen» Modell und lässt schrittweise jene erklärende Variable weg, welche zur grössten Verbesserung des AIC/BIC-Wertes führt.

Stoppregel: Wird der AIC/BIC nicht mehr kleiner, ist keine Verbesserung mehr möglich. Dann wird die Selektion abgebrochen.

```
fit.full <-    lm(hipcenter ~ . , data = seatpos)
mbackAIC <- step(fit.full, direction="backward") # AIC
n <- nrow(seatpos)
mbackBIC <- step(fit.full, direction="backward", k=log(n)) # BIC
```

VORWÄRTSSELEKTION MIT AIC/BIC

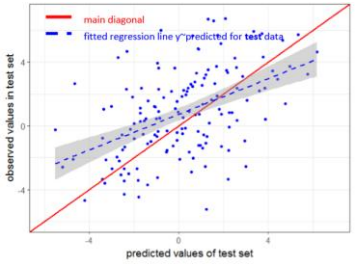
Die Vorwärtsselektion mit AIC/BIC ist das Gegenteil der Rückwärtsselektion. Statt mit dem vollen Modell zu beginnen und Variablen zu entfernen, beginnt man mit einem leeren Modell und fügt schrittweise diejenige erklärende Variable hinzu, welche die größte Verbesserung des AIC/BIC-Wertes bewirkt.

```
fit.empty <- lm(hipcenter ~ 1 , data = seatpos)
fit.full <-  lm(hipcenter ~ . , data = seatpos)

# AIC
mforwardAIC <- step(fit.empty, direction = "forward",
            scope= list(upper= fit.full))
# BIC
mforwardBIC <- step(fit.empty, direction = "forward",
            scope= list(upper= fit.full), k = log(n))
```

KOEFFIZIENTEN SHRINKAGE UM PREDICTION ZU OPTIMIEREN

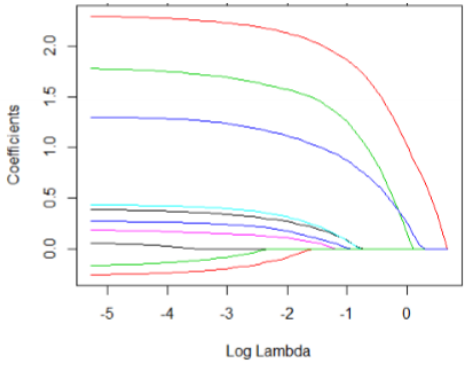
LASSO REGRESSION

Wenn ein Modell nicht kalibriert ist, kann es zu extremen Vorhersagen führen, was bedeutet, dass die vorhergesagten Wahrscheinlichkeiten oder Werte möglicherweise nicht gut mit den tatsächlichen Ergebnissen übereinstimmen.



Lasso (Least Absolute Shrinkage and Selection Operator) ist eine Regularisierungstechnik, sie fügt eine Penalisierung zur Verlustfunktion hinzu, die auf der Summe der absoluten Werte der Koeffizienten basiert. Dies führt dazu, dass einige Koeffizienten auf genau null gesetzt werden, wodurch das Modell vereinfacht wird.

$$\hat{\beta}^{lasso} \min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$



Lasso-Regression kann dazu beitragen, die Vorhersage von extremen Werten zu minimieren und stabilere Vorhersagen zu liefern, indem es die

- Modellkomplexität wird reduziert
- Overfitting wird verhindert