Aspekte des Allgemeinen Linearen Modells

Kolloquium für Statistik-Interessierte

Departement of Health Professions
Bern University of Applied Sciences

5. September 2023

General Linear Model (LM)

- Erklären oder vorhersagen
- einer quantitativen abhängigen Variablen
- durch eine Menge von unabhängigen Variablen, die kategoriell oder kontinuierlich sein können.
  - *t*-Tests
  - die einfache Regression
  - klassische Varianzanalysen
  - Kovarianzanalysen
  - multiple Regressionen

- Abgrenzung zu
  - Generalized Linear Model (GLM): Generalisierung auf diskrete Zielgrössen (Poisson-Regression, logistische Regression).
  - Linear Mixed Models (LMM): Korrelierte Fehler (Wiederholte Messungen).
  - Generalized Linear Mixed Models (GLMM): Korrelierte Fehler mit diskreten Zielgrössen.

Modell mit p Eingangsgrössen

$$\underbrace{Y_i}_{\text{Zielgrösse}} = \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip}}_{\text{linearer Prädiktor}} + \underbrace{\epsilon_i}_{\text{Messfehler}} \qquad i = 1, \ldots, n.$$

- Fehler unabhängig und gleichverteilt: ϵ_i i.i.d.
- Fehler haben Erwartungswert 0: $\mathrm{E}(\epsilon_i) = 0$
- Konstante Varianz: $\mathrm{Var}(\epsilon_i) = \sigma^2$

Matrixnotation

$$\boxed{\underset{n\times 1}{\boldsymbol{Y}} = \underset{n\times p}{X}\,\underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}}$$

Man nennt die Matrix X die **Design**-Matrix.
Ausgeschrieben ist dies

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \\ x_{i1} & x_{i2} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

- Die p Kolonnen von X sind linear unabhängig mit $n > p$.
- Die erste Eingangsgrösse ist meistens eine Konstante, $x_{i1} \equiv 1$. β_1 ist dann das Intercept (das a aus "$a + bx$" aus der Schulzeit).

Stochastisches Modell

- Die Fehler $\epsilon_i, \ldots, \epsilon_n$ bilden eine i.i.d. Zufallstichprobe.
- Alle Arten von Messfehlern oder Unmöglichkeit der Erfassung von nicht-systematischen Effekten werden in dieser Zufallsvariable mit Erwartungswert 0 subsumiert.
- Die beobachteten Zielgrössen in den Daten werden als Realisierungen von Zufallsvariablen $Y_1, \ldots, Y_n$ betrachtet.

Einfache Regression

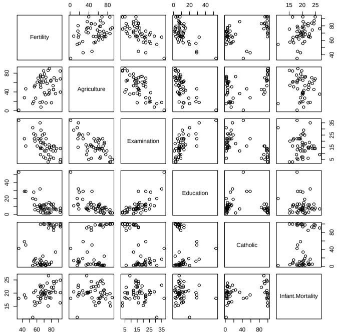
Dieses Modell hat zwei Parameter.

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

lm(): Punktschätzungen der $\beta_j, \quad j = 1, \ldots, p$

```
pairs(swiss)
```




```
mod <- lm(Fertility ~ ., swiss)
mod

##
## Call:
## lm(formula = Fertility ~ ., data = swiss)
##
## Coefficients:
##      (Intercept)       Agriculture       Examination         Education          Catholic  Infant.Mortality
##           66.915            -0.172            -0.258            -0.871             0.104             1.077
```

Ziele

- Gute Anpassung des Modells an die Daten.
- Gute Schätzungen der Parameter des Modells.
- Vorhersage der abhängigen Variablen bei neuen Daten als Eingangsgrössen.
- Unsicherheit und Signifikanz quantifizieren.
- Entwicklung eines guten Modells (In einem interaktiven Prozess werden Teile des Modells verändert um zu einem besseren Modell zu gelangen).

Kleinste-Quadrate-Schätzer*

- Optimierungsproblem: $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{Y} - X\beta||^2$

- Lösung: $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}$
- Mit den Residuen $r_i = Y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$: $\quad \hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n r_i^2$

- Erwartungswert und Varianz der Schätzer:
  - $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ (Erwartungstreue)
  - $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (X^T X)^{-1}$. (Daraus werden die Standardfehler berechnet).

Verteilung der Schätzer bei Normalverteilung*

- Wenn zusätzlich ϵ_i i.i.d. $\sim \mathcal{N}(0, \sigma^2)$. (Normalverteilung)
- Dann gilt
  1. Parameterschätzungen: $\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(X^T X)^{-1})$ (multivariat normal)
  2. Geschätzte Residualvarianz: $\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p}\chi^2_{n-p}$
- Wenn nicht normalverteilt: Für grosse n sind diese Aussagen dann trotzdem annähernd wahr (Zentraler Grenzwertsatz).
- erlaubt Konfidenzbereiche mit z.B. ± 1.96-Regel

Simpson-Paradox

Multiple Regressionen sagen viel mehr aus als mehrere einfache Regressionen.

- SATM: Average score of graduating high-school students in the state on the math component of the Scholastic Aptitude Test (a standard university admission exam).
- percent: Percentage of graduating high-school students in the state who took the SAT exam.
- pay: Average teacher's salary in the state, in 1000s.

```
psych::headTail(States)

##     region  pop SATV SATM percent dollars pay
## AL     ESC 4041  470  514       8    3.65  27
## AK     PAC  550  438  476      42    7.89  43
## AZ     MTN 3665  445  497      25    4.23  30
## AR     WSC 2351  470  511       6    3.33  23
## ...   <NA>  ...  ...  ...     ...     ... ...
## WA     PAC 4867  437  486      44    5.04  33
## WV      SA 1793  443  490      15    5.05  26
## WI     ENC 4892  476  543      11    5.95  33
## WY     MTN  454  458  519      13    5.26  29
```

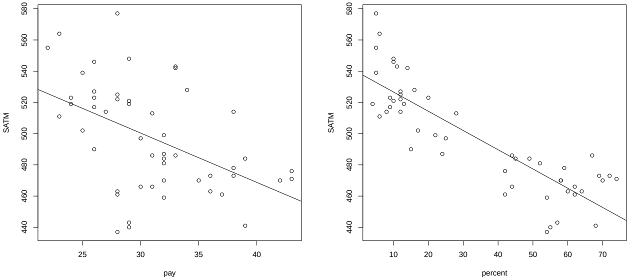
Simpson-Paradox

```
lm(SATM ~ pay, data = States)$coef  ## regression on pay

## (Intercept)         pay
##      595.19       -3.16

lm(SATM ~ percent, data = States)$coef  ## regression on percent

## (Intercept)     percent
##      538.97       -1.23
```



Simpson-Paradox

```
lm(SATM ~ pay + percent, data = States)$coef  ## regression on pay and percent

## (Intercept)         pay     percent
##     513.699       0.972      -1.374
```

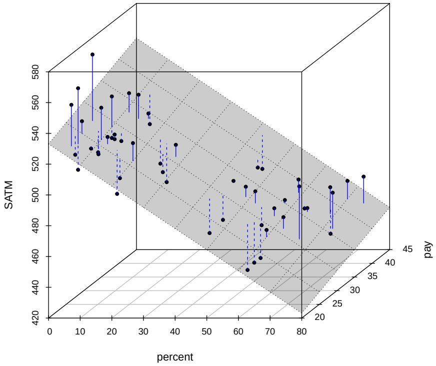


Abbildung: Regression von SATM auf percent und pay

t-Tests von $H_0 : \beta_j = 0, \quad j = 1, \ldots, p$

```
round(cbind(summary(mod)$coef, confint(mod)), 3)

##                  Estimate Std. Error t value Pr(>|t|)  2.5 % 97.5 %
## (Intercept)        66.915     10.706    6.25    0.000 45.294 88.536
## Agriculture        -0.172      0.070   -2.45    0.019 -0.314 -0.030
## Examination        -0.258      0.254   -1.02    0.315 -0.771  0.255
## Education          -0.871      0.183   -4.76    0.000 -1.241 -0.501
## Catholic            0.104      0.035    2.95    0.005  0.033  0.175
## Infant.Mortality    1.077      0.382    2.82    0.007  0.306  1.848
```

F-Tests von $H_0 : \beta_j = 0 \quad j = 1, \ldots, p$

F-tests with Type III Sum of Squares ist identisch mit t-tests.

```
car::Anova(mod, type = 3)  #F-Test, TypeIII SS, marginal

## Anova Table (Type III tests)
##
## Response: Fertility
##                  Sum Sq Df F value  Pr(>F)
## (Intercept)        2006  1   39.07 1.9e-07
## Agriculture         308  1    5.99  0.0187
## Examination          53  1    1.03  0.3155
## Education          1163  1   22.64 2.4e-05
## Catholic            448  1    8.72  0.0052
## Infant.Mortality    409  1    7.96  0.0073
## Residuals          2105 41
```

Achtung: `anova()` macht sequentielle Tests (Type I Sum of Squares), dort werden andere (sequentielle) Hypothesen getestet. Reihenfolge der Eingangsgrössen wichtig.

```
anova(mod)  #F-Test, TypeI SS, sequential

## Analysis of Variance Table
##
## Response: Fertility
##                  Df Sum Sq Mean Sq F value  Pr(>F)
## Agriculture       1    895     895   17.43 0.00015
## Examination       1   2210    2210   43.05 6.9e-08
## Education         1    892     892   17.37 0.00015
## Catholic          1    667     667   12.99 0.00084
## Infant.Mortality  1    409     409    7.96 0.00734
## Residuals        41   2105      51
```