

# Multiple Lineare Regression

Peter von Rohr

27 Februar 2017

# Outline

- ▶ Das Modell
- ▶ Stochastische Komponente: Zufälliger Rest
- ▶ Methode der kleinsten Quadrate - Least Squares
- ▶ Annahmen und Eigenschaften
- ▶ Tests und Konfidenzintervalle
- ▶ Analyse der Residuen
- ▶ Modellwahl

# Das lineare Modell

- ▶ Gegeben: **Zielgrösse** (response variable)  $y_i$  für Individuum  $i$ .  
Entspricht einer Beobachtung oder Messung, welche zu  $i$  gehört.
- ▶ Gegeben: mehrere **erklärende Variablen** (predictors or covariables)  $x_{i,1}, x_{i,2}, \dots, x_{i,p}$ , welche Eigenschaften von  $i$  beschreiben
- ▶ Zusammengefasst wissen wir von Individuum  $i$ :  
 $(x_{i,1}, x_{i,2}, \dots, x_{i,p}, y_i)$
  
- ▶ Multiple lineare Regression sagt:

**Zielgrösse = lineare Funktion der erklärenden Variablen + Rest**

# Das Modell als Formel

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_p x_{i,p} + \epsilon_i$$

- ▶ In einer Population mit  $n$  Individuen, können wir  $n$  solche Gleichungen aufstellen, wobei die  $i$  von 1 bis  $n$  laufen
- ▶ Zur Vereinfachung wird eine Matrix-Vektor Schreibweise verwendet

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (1)$$

wobei:

- $\mathbf{y}$  Vektor mit Beobachtungen (Länge  $n$ )
- $\boldsymbol{\beta}$  Vektor mit Parametern (Länge  $p$ )
- $\mathbf{X}$  Matrix mit erklärenden Variablen (Dimension  $n \times p$ )
- $\boldsymbol{\epsilon}$  Vektor mit zufälligen Resten (Länge  $n$ )

# Stochastisches Modell

- ▶ Zufällige Komponente  $\epsilon$  im Lineares Modell (siehe Gleichung (1))
- ▶ Somit ist die Zielgröße ( $\mathbf{y}$ ) auch zufällig
- ▶ In unserem Beispiel sind die erklärenden Variablen als fix angenommen
- ▶ In gewissen Anwendungen können auch erklärende Variablen als zufällig angenommen werden (BLUP-Zuchtwertschätzung)
- ▶ Verschiedene Einflüsse auf  $\epsilon$ : Messfehler, unbekannte Einflussfaktoren
- ▶ Annahme: unbekannte Faktoren haben sich im “Mittel” auf  $\rightarrow E(\epsilon) = \mathbf{0}$
- ▶ Streuung wird quantifiziert mit:  $var(\epsilon) = \mathbf{I} * \sigma^2$

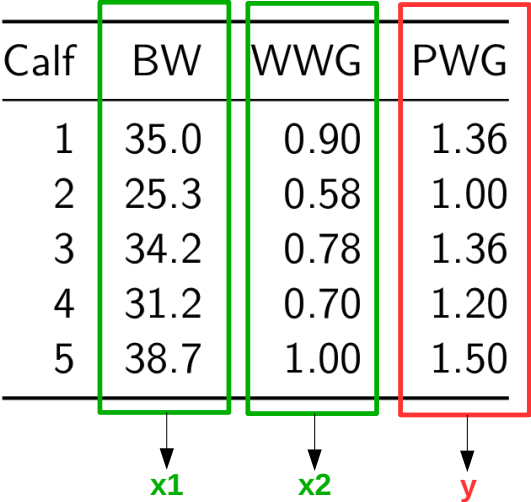
## Beispiel

Einfluss des Geburtsgewichts (BW) und Zunahme vor dem Absetzen (WWG) auf Zunahme nach dem Absetzen (PWG)

Calf	BW	WWG	PWG
1	35.0	0.90	1.36
2	25.3	0.58	1.00
3	34.2	0.78	1.36
4	31.2	0.70	1.20
5	38.7	1.00	1.50

# Identifikation der Komponenten des Modells

Calf	BW	WWG	PWG
1	35.0	0.90	1.36
2	25.3	0.58	1.00
3	34.2	0.78	1.36
4	31.2	0.70	1.20
5	38.7	1.00	1.50



## Komponenten als Formeln

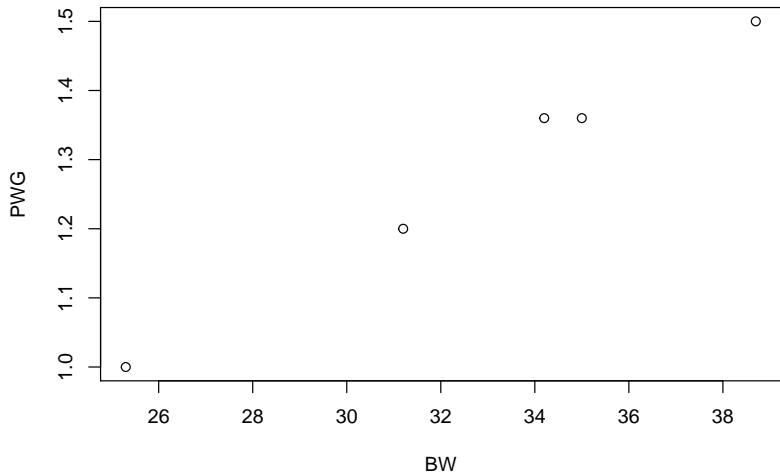
$$\mathbf{y} = \begin{bmatrix} 1.36 \\ 1.00 \\ 1.36 \\ 1.20 \\ 1.50 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 35.00 & 0.90 \\ 25.30 & 0.58 \\ 34.20 & 0.78 \\ 31.20 & 0.70 \\ 38.70 & 1.00 \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix} \quad (2)$$

- Achsenabschnitt:  $\mathbf{y}$  und  $\epsilon$  ändern sich nicht

$$\mathbf{X} = \begin{bmatrix} 1 & 35.00 & 0.90 \\ 1 & 25.30 & 0.58 \\ 1 & 34.20 & 0.78 \\ 1 & 31.20 & 0.70 \\ 1 & 38.70 & 1.00 \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \quad (3)$$



## Weshalb ein Achsenabschnitt



# Quadratische Regression und Transformationen

- ▶ **Wichtig:** Auch das sind lineare Regressionen

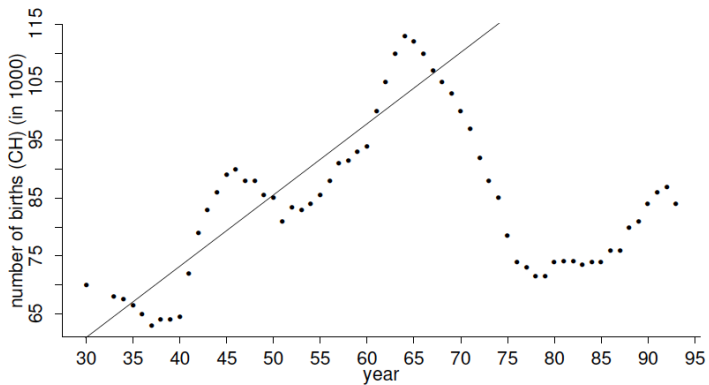
$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \\ 1 & x_5 & x_5^2 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & \log(x_{12}) & \sin(x_{13}) \\ 1 & \log(x_{22}) & \sin(x_{23}) \\ 1 & \log(x_{32}) & \sin(x_{33}) \\ 1 & \log(x_{42}) & \sin(x_{43}) \\ 1 & \log(x_{52}) & \sin(x_{53}) \end{bmatrix} \quad (4)$$

→ Modell  $\mathbf{y} = \mathbf{X}\beta + \epsilon$  heißt **lineares** Modell, weil es linear in den Koeffizienten  $\beta$  ist.

## Ziele einer Regression

- ▶ **Anpassung**: Möglichst kleine Abweichungen der angepassten Ebenen und der Zielgrösse
- ▶ Gute **Schätzung** der unbekannt Parameter: Sollen Änderungen der Zielgrösse in Abhängigkeit der Änderung der erklärenden Variablen darstellen
- ▶ Gute **Voraussage**: Soll neue Zielgrößen als Funktion von neuen Werten der erklärenden Variablen voraussagen können.  
**Achtung**: keine Extrapolation
- ▶ **Unsicherheit** bei der Schätzung: Vertrauensintervallen und statistische Tests als gute Werkzeuge

## Keine Extrapolation



# Schätzung der Parameter

- ▶ Methode der kleinsten Quadrate (Least Squares)
- ▶ Suche einer guten Schätzung für  $\beta$ , so dass die Abweichungen oder Reste möglichst klein sind
- ▶ Mathematische Formulierung: Abweichungen entsprechen

$$L = \|\mathbf{y} - \mathbf{X}\beta\|^2 = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

- ▶ Möglichst kleine Abweichung: Minimierung durch Ableiten und die Ableitung 0 setzen
- ▶ Somit ist der Least Squares Schätzer  $\hat{\beta}$  definiert als

$$\hat{\beta} = \operatorname{argmin}_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2$$

# Normalgleichungen

- ▶ Der Schätzer  $\hat{\beta}$  berechnet sich als  $p$  dimensionaler Gradient

$$\frac{\partial L}{\partial \beta} = (-2)\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\beta})$$

- ▶ Daraus folgen die **Normalgleichungen**

$$(\mathbf{X}^T\mathbf{X})\hat{\beta} = \mathbf{X}^T\mathbf{y}$$

- ▶ Auflösung nach  $\hat{\beta}$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

## Restvarianz

- ▶ Least Squares liefert eigentlich keine Schätzung für die Restvarianz  $\sigma^2$
- ▶ Aufgrund der Residuen  $r_i = y_i - \mathbf{x}_i^T \hat{\beta}$ , ergibt sich die Schätzung

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

- ▶ Schätzung ist plausible, da sie auf der Momentenmethode basiert
- ▶ Ungewöhnlicher Faktor  $1/(n-p)$  so gewählt, dass:

$$E(\hat{\sigma}^2) = \sigma^2$$

- ▶ Dieser Schätzer wird oft als Least-Squares Schätzer für  $\sigma^2$  bezeichnet

## Annahmen für ein lineares Modell

- ▶ Ausser, dass die Matrix  $\mathbf{X}$  vollen Rang hat ( $p < n$ ) wurden bis jetzt keine Annahmen gemacht
- ▶ Lineares Modell ist korrekt  $\rightarrow E(\epsilon) = \mathbf{0}$
- ▶ Die Werte in  $\mathbf{X}$  sind exakt
- ▶ Die Varianz der Fehler ist konstant (“Homoskedazidität”) für alle Beobachtungen  $\rightarrow \text{Var}(\epsilon) = \mathbf{I} * \sigma^2$
- ▶ Die Fehler sind unkorreliert
- ▶ Weitere Eigenschaften folgen, falls die Fehler normal verteilt sind