ETH zürich

# Multiple Lineare Regression (Teil 2)


Peter von Rohr

Qualitas AG

# Outline

- Tests und Konfidenzintervalle
- Analyse der Residuen
- Modellwahl

# Annahmen für ein lineares Modell

- Ausser, dass die Matrix $\mathbf{X}$ vollen Rang hat ($p < n$) wurden bis jetzt keine Annahmen gemacht

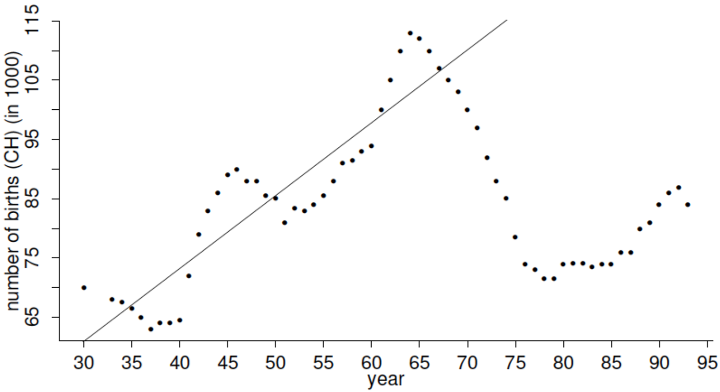
1. Lineares Modell ist korrekt $\rightarrow E(\epsilon) = \mathbf{0}$
2. Die Werte in $\mathbf{X}$ sind exakt
3. Die Varianz der Fehler ist konstant ("Homoskedazidität") für alle Beobachtungen $\rightarrow Var(\epsilon) = \mathbf{I} * \sigma^2$
4. Die Fehler sind unkorreliert
5. Weitere Eigenschaften folgen, falls die Fehler normal verteilt sind

Was passiert, wenn Annahmen nicht erfüllt sind?

# Massnahmen und Alternativen

- Falls Annahme 3 (konstante Varianzen) verletzt ist, verwenden wir weighted least squares
- Falls Annahme 5 der Normalität nicht gilt, verwenden wir robuste Methoden
- Falls Annahme 2 falsch ist, brauchen wir eine Methode namens "errors in variables"
- Falls Annahme 1 nicht zutrifft, brauchen wir ein nicht-lineares Modell

# Annahmen 1 und 4 nicht erfüllt



# Mehrere Regressionen mit einer Variablen

- **Wichtig**: Multiple lineare Regression nicht durch mehrere Regressionen mit einer Variablen ersetzen
- Beispiel: $y = 2 * x1 - x2$

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **x1** | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 |
| **x2** | -1 | 0 | 1 | 2 | 1 | 2 | 3 | 4 |
| **y** | 1 | 2 | 3 | 4 | -1 | 0 | 1 | 2 |

# Einfache Regression mit x2

```
x1 <- c(0, 1, 2, 3, 0, 1, 2, 3)
x2 <- c(-1,0,1,2,1,2,3,4)
y <- 2*x1-x2
dfData <- data.frame(x1=x1, x2=x2, y=y)
lm_simple_x2 <- lm(y ~ x2, data = dfData)
```

# Resultat

Table 2: Fitting linear model: y ~ x2

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **x2** | 0.1111 | 0.4057 | 0.2739 | 0.7934 |
| **(Intercept)** | 1.333 | 0.8607 | 1.549 | 0.1723 |

- Original: $y = 2 * x1 - x2$

# Eigenschaften der Least Squares Schätzer

- Modell: $\mathbf{y} = \mathbf{X}\beta + \epsilon$, mit $E[\epsilon] = \mathbf{0}$, $Cov(\epsilon) = \mathbf{I} * \sigma^2$

1. $E[\hat{\beta}] = \beta \rightarrow$ unverzerrter Schätzer (unbiasedness)
2. $E[\hat{\mathbf{Y}}] = E[\mathbf{Y}] = \mathbf{X}\beta \rightarrow E[\mathbf{r}] = \mathbf{0}$
3. $Cov(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$
4. $Cov(\hat{\mathbf{Y}}) = \sigma^2 P$, $Cov(\mathbf{r}) = \sigma^2(\mathbf{I} - \mathbf{P})$

wobei $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

# Verteilung der Schätzer

Annahme, dass $\epsilon$ normal-verteilt sind, daraus folgt

1. $\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$
2. $\hat{\mathbf{Y}} \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2 P)$
3. $\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p}\chi^2$

# Tests und Vertrauensintervalle

- Angenommen, wir möchten wissen, ob eine bestimmte erklärende Variable $\beta_j$ relevant ist in unserem Modell, dann testen wir die Nullhypothese

$$H_0 : \beta_j = 0$$

gegenüber der Alternativhypothese

$$H_A : \beta_j \neq 0$$

- Bei unbekanntem $\sigma^2$ ergibt sich folgende Teststatistik

$$T_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} \sim t_{n-p}$$

wobei $t_{n-p}$ für die Student-t Verteilung mit $n - p$ Freiheitsgraden steht.

# Probleme bei t-Tests

- Multiples Testen bei vielen $\beta_j$, d.h. falls wir 100 Tests mit Irrtumswahrscheinlchkeit 5% machen, sind automatisch 5 Tests signifikant
- Es kann passieren, dass für kein $\beta_j$ die Nullhypothese verworfen werden kann, aber die erklärende Variable trotzdem einen Einfluss hat. Der Grund dafür sind Korrelationen zwischen erklärenden Variablen
- Individuelle t-tests für $H_0 : \beta_j = 0$ sind so zu interpretieren, dass diese den Effekt von $\beta_j$ quantifizieren nach Abzug des Einflusses aller anderen Variablen auf die Zielgrösse $Y$

$\rightarrow$ falls z. Bsp. $\beta_i$ und $\beta_j$ stark korreliert sind und wir testen die beiden Nullhypothesen $H_{0j} : \beta_j = 0$ und $H_{0i} : \beta_i = 0$, kann durch die Korrektur der anderen Variablen der Effekt von $\beta_i$ und $\beta_j$ auf $Y$ durch den t-Test nicht gefunden werden.