Multiple Lineare Regression (Teil 2)

Peter von Rohr

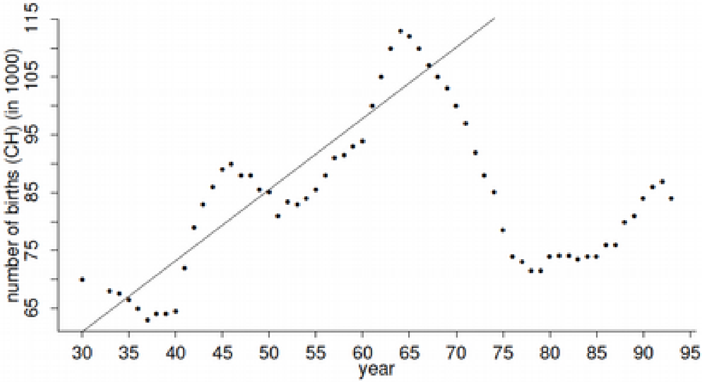
05 März 2018

Massnahmen und Alternativen

Was passiert, wenn Annahmen nicht erfüllt sind?

- 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|)
(Intercept)	1.333	0.8607	1.549	0.1723
x2	0.1111	0.4057	0.2739	0.7934

- 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 ϵ 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 β_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 σ^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 β_j, d.h. falls wir 100 Tests mit Irrtumswahrscheinlchkeit 5% machen, sind automatisch 5 Tests signifikant
- Es kann passieren, dass für kein β_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 β_j quantifizieren nach Abzug des Einflusses aller anderen Variablen auf die Zielgrösse Y

$\rightarrow$ falls z. Bsp. β_i und β_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 β_i und β_j auf Y durch den t-Test nicht gefunden werden.

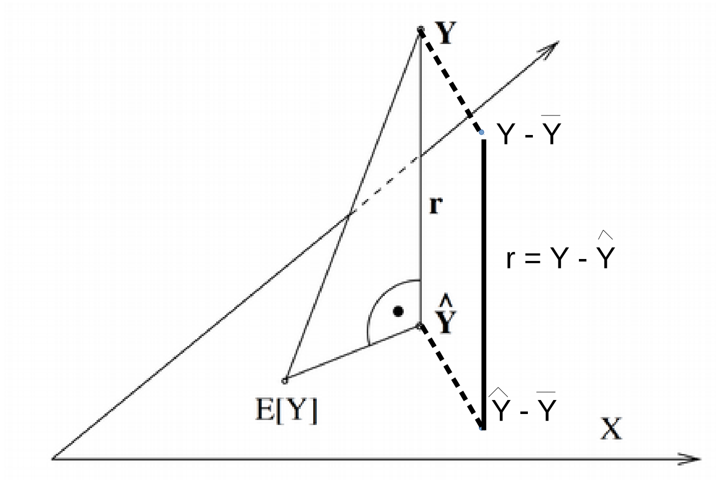
Globaler Test eines Modells

- Beim t-Test hatten wir jede einzelne erklärende Variable getestet.
- Test, ob überhaupt eine der erklärenden Variablen einen Einfluss auf die Zielgrösse hat
- Zerlegung der Länge der totalen quadrierten Abweichungen der Beobachtungswerte $\mathbf{y}$ um deren Mittel $\bar{\mathbf{y}}$ in

$$||\mathbf{y} - \bar{\mathbf{y}}||^2 = ||\hat{\mathbf{y}} - \bar{\mathbf{y}}||^2 + ||\mathbf{y} - \hat{\mathbf{y}}||^2$$

wobei: $||\hat{\mathbf{y}} - \bar{\mathbf{y}}||^2$ der Länge der quadrierten Abweichungen der gefitteten Werte ($\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$) um das globale Mittel ($\bar{\mathbf{y}} = \mathbf{1} * 1/n \sum_{i=1}^{n} y_i$) und $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$ den Residuen entspricht

Geometrische Begründung



Zerlegung als Varianzanalyse (ANOVA)

- ANOVA Tabelle sieht wie folgt aus

	sums of squares	degrees of freedom	mean square
regression	$\|\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|\|^2$	$p-1$	$\|\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|\|^2/(p-1)$
error	$\|\|\mathbf{y} - \hat{\mathbf{y}}\|\|^2$	$n-p$	$\|\|\mathbf{y} - \hat{\mathbf{y}}\|\|^2/(n-p)$
total	$\|\|\mathbf{y} - \bar{\mathbf{y}}\|\|^2$	$n-1$	

- Relevante Teststatistik lautet

$$F = \frac{||\hat{\mathbf{y}} - \bar{\mathbf{y}}||^2/(p-1)}{||\mathbf{y} - \hat{\mathbf{y}}||^2/(n-p)} \sim F_{p-1,n-p}$$

unter der globalen Nullhypothese $H_0 : \beta_j = 0$ für alle j

Bestimmtheitsmass des Modells

- Nützliche Grösse für die Qualität eines Modells ist das Bestimmtheitsmass (coefficient of determination)

$$R^2 = \frac{||\hat{\mathbf{y}} - \bar{\mathbf{y}}||^2}{||\mathbf{y} - \bar{\mathbf{y}}||^2}$$

diese sagt aus, wieviel der totalen Variation von $\mathbf{y}$ um $\bar{\mathbf{y}}$ durch die Regression erkärt wird.

Vertrauensintervall der Schätzung

- Basierend auf der Teststatistik des t-Tests

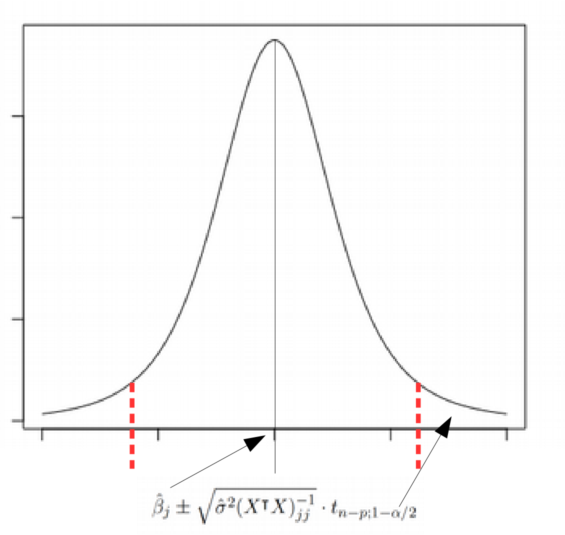
$$T_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1}_{jj}}} \sim t_{n-p}$$

- Vertrauensintervall für den unbekannten Parameter β_j als

$$\hat{\beta}_j \pm \sqrt{\hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1}_{jj}} * t_{n-p,1-\alpha/2}$$

$\rightarrow$ somit beinhaltet das Intervall zwischen den angegebenen Grenzen den wahren Wert mit Wahrscheinlichkeit $1 - \alpha$, wobei $t_{n-p,1-\alpha/2}$ das $1 - \alpha/2$ Quantil der Verteilung t_{n-p} darstellt

Vertrauensintervall im Bild



R Output

```
Call:
lm(formula = LOGRUT ~ ., data = asphalt1)
```

1

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.48348 -0.14374 -0.01198  0.15523  0.39652
```

2

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.781239   2.459179  -2.351 0.027280 *
LOGVISC     -0.513325   0.073056  -7.027 2.90e-07 ***
ASPH         1.146898   0.265572   4.319 0.000235 ***
BASE         0.232809   0.326528   0.713 0.482731
RUN         -0.618893   0.294384  -2.102 0.046199 *
FINES        0.004343   0.007881   0.551 0.586700
VOIDS        0.316648   0.110329   2.870 0.008433 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3

```
Residual standard error: 0.2604 on 24 degrees of freedom
Multiple R-Squared: 0.9722,     Adjusted R-squared: 0.9653
F-statistic: 140.1 on 6 and 24 DF,  p-value: < 2.2e-16
```

4

R Output Bedeutung

1. Funktionsaufruf mit welchem das Resultatobjekt erzeugt wurde. Wichtig, falls Resultate als R-objekt (.rda) gespeichert werden
2. Verteilung der Residuen aufgrund der Quantile
3. Schätzwert und Schätzfehler für die Paramter β_j zu jeder erklärenden Variablen. Werte der t-Teststatistik
4. Schätzung der Rest-Standardabweichung σ. Zusätzliche Modellinformationen, wie F-Teststatistik, R^2 und das um Anzahl erklärende Variablen korrigierte R^2, wobei

$$\bar{R}^2 = R^2 - (1 - R^2)\frac{p - 1}{n - p}$$

Überprüfung der Modellannahmen anhand Analyse der Residuen

- Residuen $r_i = y_i - \hat{y}_i$ als Approximation der unbekannten Fehler ϵ_i bei der Überprüfung der Modellannahmen verwenden
- **Tukey-Anscombe** Plot: zeigt Residuen r_i versus gefittete Werte $\hat{y}_i$. Dieser sollte keine erkennbaren Muster aufweisen

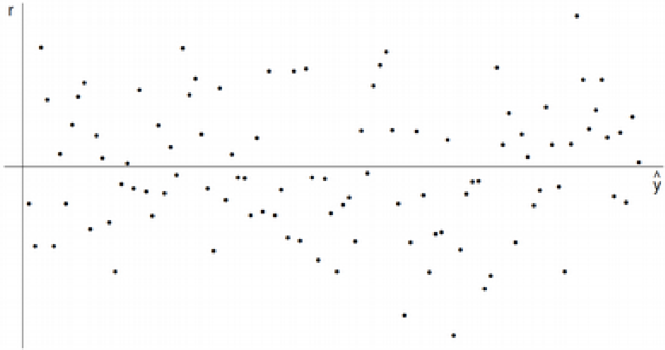


Figure 1.4: Ideal Tukey-Anscombe plot: no violations of model assumptions.

Probleme bei Modellannahmen

Folgende Plots deuten auf Probleme hin

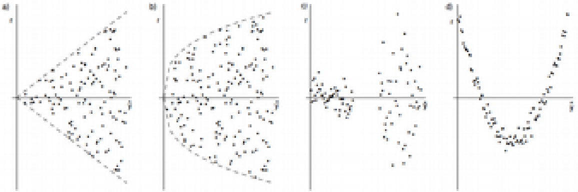
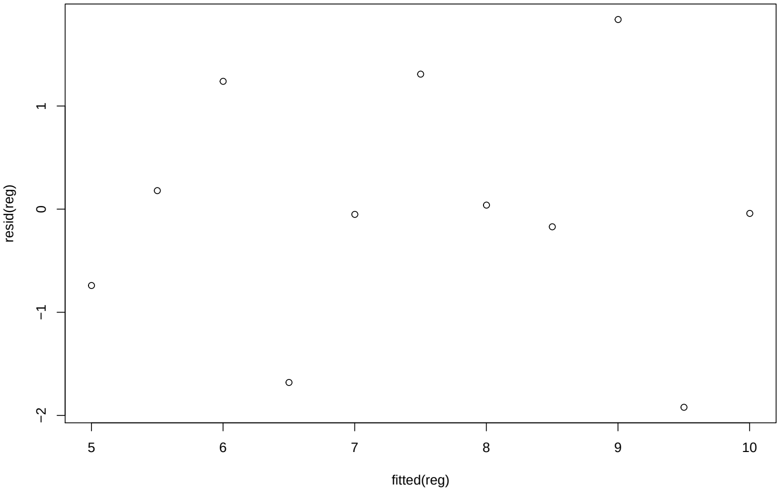


Figure 1.5: a) linear increase of standard deviation, b) nonlinear increase of standard deviation, c) 2 groups with different variances, d) missing quadratic term in the model.

Tukey-Anscombe Plot in R

```r
data(anscombe)
reg <- lm(y1 ~ x1, data = anscombe)
plot(fitted(reg), resid(reg))
```

Tukey-Anscombe Plot - Das Resultat

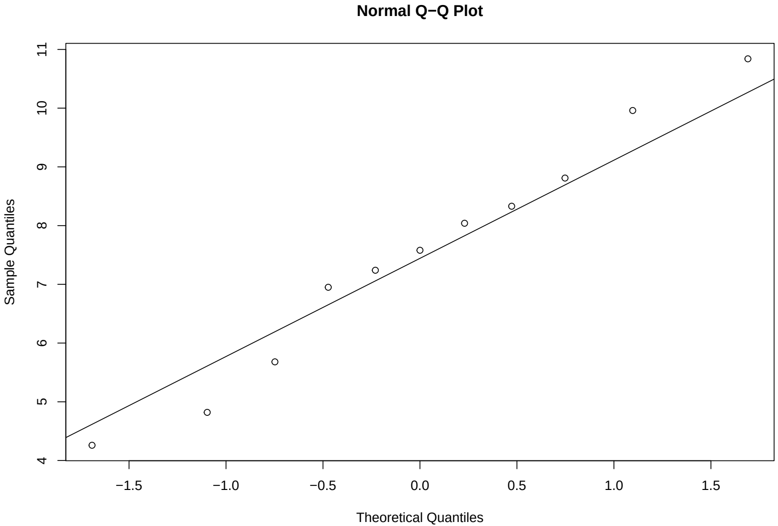


QQ (quantile-quantile) Plot

- Überprüfung der Verteilung der Zufallsvariablen (Zielgrösse und Residuen)
- Empirische Verteilung der Residuen (y-Achse) wird gegen theoretische Quantile der Normalverteilung (x-Achse) aufgezeichnet
- Falls Normalverteilung zutrifft, dann liegen alle Punkte auf einer Linie

In R:

```
qqnorm(anscombe$y1)
qqline(anscombe$y1)
```



Probleme mit Verteilung

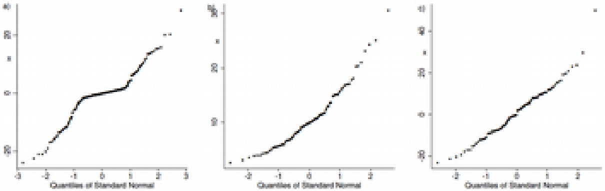


Figure 1.7: QQ-plots for a) long-tailed distribution, b) skewed distribution, c) dataset with outlier.

Quellen

Tukey-Anscombe Plots und QQ-Plots stammen aus dem Skript:

Computational Statistics
Peter Bühlmann and Martin Mächler
Seminar für Statistik ETH Zürich
Version of January 31, 2014