Einführung Lineare Regression mit R

Prof. Dr. Karsten Lübke

SoSe 2017

Modellierung

- *Supervised Learning*: Kann ein Teil der Variation einer abhängigen Variable y durch unabhängige Variable(n) x modelliert werden: $y = f(x) + \epsilon$
- Schätze $\hat{f}$ anhand der Daten
- Annahme: f ist *lineare* Funktion: $y = \beta_0 + \beta_1 \cdot x + \epsilon$ Hier: y numerisch, nur eine unabhängige x
- Kleinste Quadrate Kriterium: $\hat{\beta}$ so, dass $\min \sum \epsilon_i^2$

Vorbereitung: Trinkgeld und Rechnungshöhe

Einlesen der “Tipping”[1] Daten sowie laden des Pakets `mosaic`.
Zufallszahlengenerator setzen.

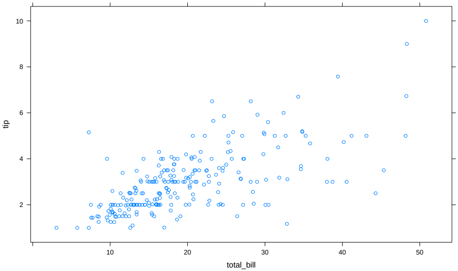
```
download.file("https://goo.gl/whKjnl", destfile = "tips.csv")
tips <- read.csv2("tips.csv")
# Alternativ - heruntergeladene Datei einlesen:
# tips <- read.csv2(file.choose())

library(mosaic) # Paket
set.seed(1896) # Zufallzahlengenerator
```

[1] Bryant, P. G. and Smith, M (1995) Practical Data Analysis: Case Studies in Business Statistics. Homewood, IL: Richard D. Irwin Publishing

Streudiagramm: Trinkgeld und Rechnungshöhe

```
xyplot(tip ~ total_bill, data = tips)
```



Übung 1: Korrelation Trinkgeld und Rechnungshöhe

Welche Aussage stimmt vermutlich für den Korrelationskoeffizient zwischen Trinkgeld und Rechnungshöhe?

- A: Der Korrelationskoeffizient liegt bei $r = -0.68$
- B: Der Korrelationskoeffizient liegt bei $r = -0.34$
- C: Der Korrelationskoeffizient liegt bei $r = 0.68$
- D: Der Korrelationskoeffizient liegt bei $r = 0.34$

Lineare Regression Trinkgeld und Rechnungshöhe

```r
erglm1 <- lm(tip ~ total_bill, data = tips)
summary(erglm1)
```

```
##
## Call:
## lm(formula = tip ~ total_bill, data = tips)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1982 -0.5652 -0.0974  0.4863  3.7434
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.920270   0.159735   5.761 2.53e-08 ***
## total_bill  0.105025   0.007365  14.260  < 2e-16 ***
## ---
```

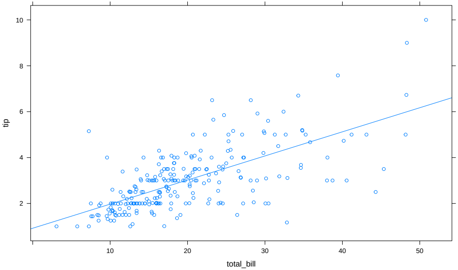
Übung 2: Regression Trinkgeld und Rechnungshöhe

Welche Aussage stimmt?

- A: Im Mittel steigt mit jedem Dollar Trinkgeld die Rechnungshöhe um 0.92
- B: Im Mittel steigt mit jedem Dollar Trinkgeld die Rechnungshöhe um 0.11
- C: Im Mittel steigt mit jedem Dollar Rechnungshöhe das Trinkgeld um 0.92
- D: Im Mittel steigt mit jedem Dollar Rechnungshöhe das Trinkgeld um 0.11

Regressionsgerade

```r
plotModel(erglm1)
```



Geschätzte Regressionsgleichung

Die geschätzte Gleichung lautet:

$$y_i = 0.92 + 0.11 \cdot x_i + \epsilon_i$$

Die Punktprognose lautet dann:

$$\hat{y}_i = 0.92 + 0.11 \cdot x_i$$

Übung 3: Prognose der Trinkgeldhöhe aus Rechnungshöhe

Für $x_0 = 10$ lautet die Prognose $\hat{y}_0 = 0.92 + 0.11 \cdot 10 = 2.02$.

Stimmt die Aussage: Bei einer Rechnungshöhe von 10$ wird das Trinkgeld mit Sicherheit bei 2.02$ liegen?

- Ja.
- Nein.

Prognoseintervalle

```r
predict(erglm1, # Modell
        # Neue Beobachtung mit x=10:
        newdata = data.frame(total_bill = 10),
        # Prognoseintervall:
        interval  = "prediction")
```

```
##         fit         lwr     upr
## 1 1.970515 -0.05184074 3.99287
```

Übung 4: Bestimmheitsmaß

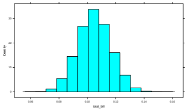
Es gilt: $R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 0.4566$ (`Multiple R-squared`). Welche Aussage stimmt?

- A: Das Modell ist zu 46% korrekt
- B: 46% der Beobachtungen werden richtig modelliert
- C: 46% der Variation der Rechnungshöhe werden modelliert
- D: 46% der Variation der Trinkgeldhöhe werden modelliert

Bootsrap Steigungskoeffizient

```
Bootvtlg <- do(10000) *
  coef(lm(tip ~ total_bill, data = resample(tips)))[2]

histogram( ~ total_bill, data = Bootvtlg)
```



```
quantile( ~ total_bill, data = Bootvtlg,
          probs = c(0.025, 0.975))
```

```
##        2.5%      97.5%
## 0.08235625 0.12797229
```

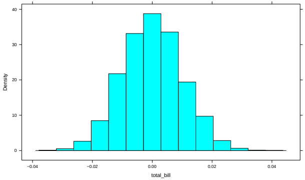
Permutationstest Steigung (I/II)

Wenn $H_0 : \beta_1 = 0$ gilt, so sollte y in keinen (linearen) Zusammenhang zu x stehen:

```
Nullvtlg <- do(10000) *
  coef(lm(tip ~ shuffle(total_bill), data = tips))[2]
```

Permutationstest Steigung (II/II)

```
histogram( ~ total_bill, data = Nullvtlg)
```




```
quantile( ~ total_bill, data = Nullvtlg,
          probs = c(0.025, 0.975))
```

```
##         2.5%       97.5%
## -0.01902087  0.01980632
```

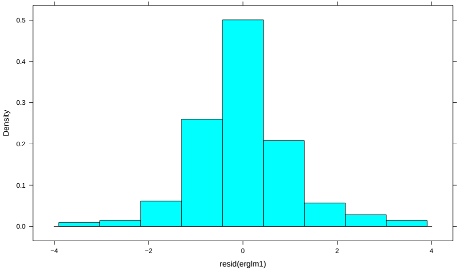
Übung 5: Permutationstest Steigung

Welche Aussage stimmt?

- A: Die Beobachtete Steigung der Stichprobe $\hat{\beta}_1 = 0.11$ ist unter $H_0 : \beta_1 = 0$ ein üblicher Wert.
- B: Die Beobachtete Steigung der Stichprobe $\hat{\beta}_1 = 0.11$ ist unter $H_0 : \beta_1 = 0$ kein üblicher Wert.

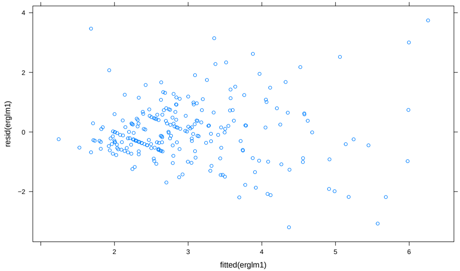
Verteilung Residuen

```
histogram( ~ resid(erglm1))
```



Verteilung Residuen und angepasste Werte

```
xyplot(resid(erglm1) ~ fitted(erglm1))
```



Übung 6: Verteilung Residuen und angepasste Werte

Welche Aussage stimmt?

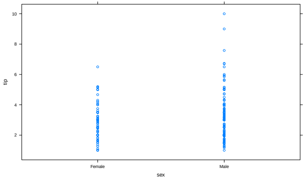
- A: Die Varianz der Residuen scheint unabhängig von der Höhe der angepassten Werte zu sein.
- B: Die Varianz der Residuen scheint mit der der Höhe der angepassten Werte zu steigen.
- C: Die Varianz der Residuen scheint mit der der Höhe der angepassten Werte zu fallen.

Trinkgeld und Geschlecht

```
mean(tip ~ sex, data = tips)
```

```
##   Female     Male
## 2.833448 3.089618
```

```
xyplot(tip ~ sex, data = tips)
```



Regression Trinkgeld und Geschlecht

```r
erglm2 <- lm(tip ~ sex, data = tips)
summary(erglm2)
```

```
##
## Call:
## lm(formula = tip ~ sex, data = tips)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0896 -1.0896 -0.0896  0.6666  6.9104
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.8334     0.1481  19.137   <2e-16 ***
## sexMale       0.2562     0.1846   1.388    0.166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Übung 7: Regression Trinkgeld und Geschlecht

Welche Aussage stimmt?

- A: Im Mittel geben Männer 0.26 mehr Trinkgeld als Frauen
- B: Im Mittel geben Frauen 0.26 mehr Trinkgeld als Männer
- C: Männer geben 0.26 mehr Trinkgeld als Frauen
- D: Frauen geben 0.26 mehr Trinkgeld als Männer

Multiple Regression

Modelliere Trinkgeldhöhe als lineare Funktion von Rechnungshöhe und Geschlecht

```
erglm3 <- lm(tip ~ total_bill + sex, data = tips)
summary(erglm3)
```

```
##
## Call:
## lm(formula = tip ~ total_bill + sex, data = tips)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1914 -0.5596 -0.0875  0.4845  3.7465
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.933278   0.173756   5.371 1.84e-07 ***
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.9332785	0.1737557	5.3712093	0.0000002
total_bill	0.1052324	0.0074582	14.1096681	0.0000000
sexMale	-0.0266087	0.1383340	-0.1923513	0.8476290

Übung 8: Regression Trinkgeld und Geschlecht

Stimmt die Aussage: Gegeben die Rechnungshöhe geben Männer im Mittel mehr Trinkgeld als Frauen.

- Ja.
- Nein.

Permutationstest Steigung Geschlecht

```
Nullvtlg <- do(10000) *
  coef(lm(tip ~ total_bill + shuffle(sex), data = tips))[3]

prop( ~ abs(sexMale) >= abs(coef(erglm3)[3]),
      data = Nullvtlg)
```

```
##   TRUE
## 0.8559
```

Übung 9: Inferenz Regression Trinkgeld und Geschlecht

Gegeben die Rechnungshöhe, kann die Nullhypothese $\beta_2 = \beta_{\text{sex}} = 0$ zum Signifikanzniveau 5% verworfen werden?

- Ja.
- Nein.

Übung 10: Interpretation Regression

Welches ist die korrekteste Interpretation von $\hat{\beta}_1 = \hat{\beta}_{\text{total\_bill}} = 0.11$?

- A: Mit jedem $ Rechnungshöhe steigt das Trinkgeld um 0.11$.
- B: Mit jedem $ Rechnungshöhe steigt das Trinkgeld im Mittel um 0.11$.
- C: Mit jedem $ Rechnungshöhe steigt das Trinkgeld im Mittel um 0.11$, gegeben alle anderen Faktoren bleiben konstant.
- D: In einem linearen Modell steigt mit jedem $ Rechnungshöhe das Trinkgeld im Mittel um 0.11$, gegeben alle anderen Faktoren bleiben konstant.
- E: In der Stichprobe steigt in einem linearen Modell mit jedem $ Rechnungshöhe das Trinkgeld im Mittel um 0.11$, gegeben alle anderen Faktoren bleiben konstant.

Offene Übung: Rechnungshöhe

Modellieren Sie die Rechnungshöhe als Funktion der Anzahl Personen sowie der Tageszeit.