Einführung in die statistische Datenanalyse mit R

Lineare Regression II

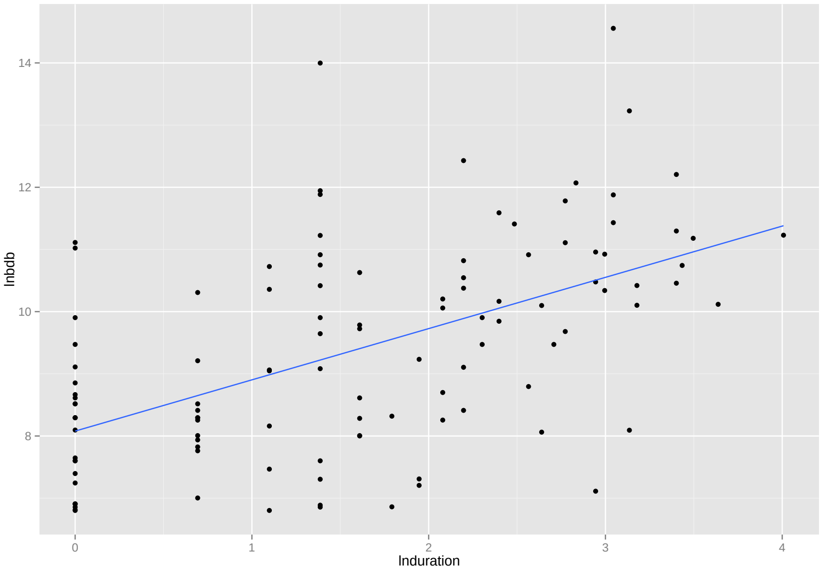
David Benček

Wintersemester 2015/16

Zur Erinnerung: Zweck der Regression

- Formulierung eines plausiblen Modells für die Wirkung von Einflussgrößen (unabhängigen Variablen) auf die abhängige Größe (abhängige Variable),
- Quantifizierung der Wirkung von Einflussgrößen,
- Bestimmung der statistischen Signifikanz von Effekten,
- Vorhersage der abhängigen Größe bei neuen Beobachtungen.

Lineare Einfachregression



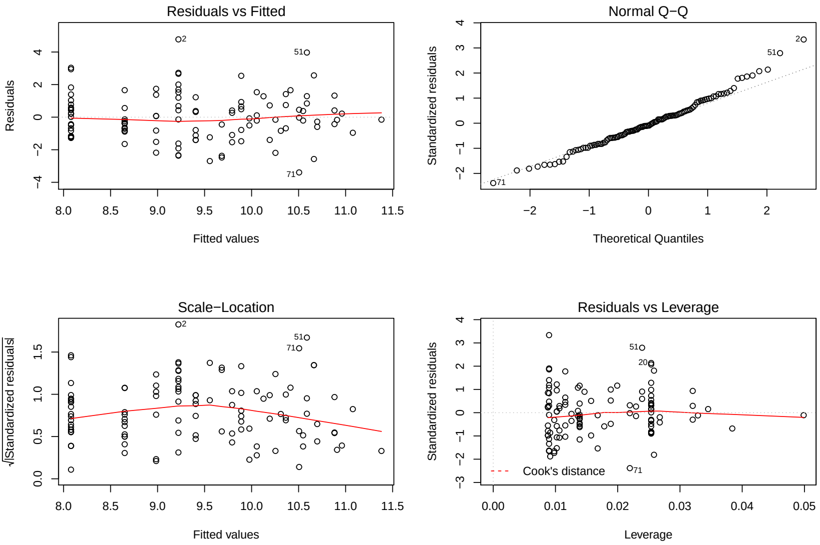
Lineare Einfachregression in R

```r
lacina_one <- lm(lnbdb ~ lnduration, data = lacina)
summary(lacina_one)
```

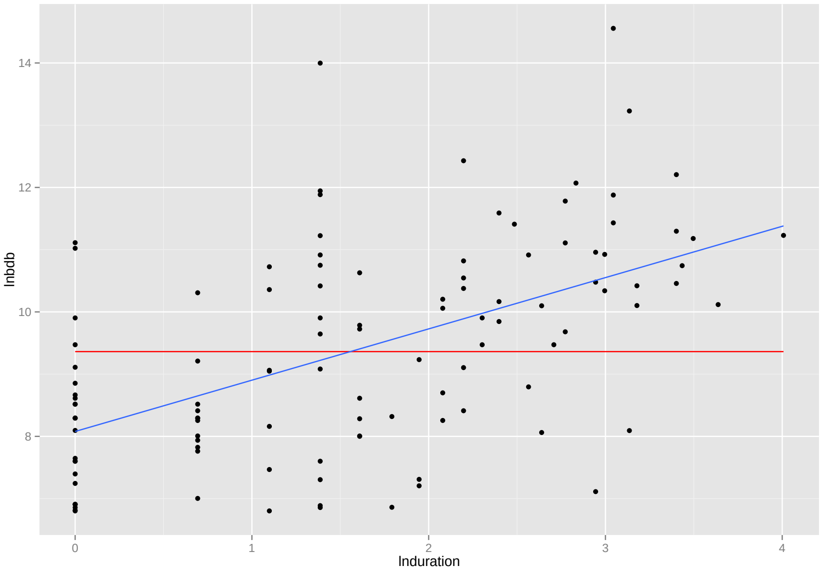
```
##
## Call:
## lm(formula = lnbdb ~ lnduration, data = lacina)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3934 -0.8756 -0.1360  0.7390  4.7763
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.0790     0.2291  35.263  < 2e-16 ***
## lnduration    0.8242     0.1190   6.925 2.89e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.438 on 112 degrees of freedom
## Multiple R-squared:  0.2998, Adjusted R-squared:  0.2936
## F-statistic: 47.96 on 1 and 112 DF,  p-value: 2.892e-10
```

Modellqualität

```r
par(mfrow=c(2,2))
plot(lacina_one)
```



R^2



R^2 – Intuition

- Wieviel Varianz der abhängigen Variable wird durch das Modell erklärt?

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\text{Variation der Regresswerte}}{\text{Variation von Y}}$$

- Manuelle Berechnung:

```
sum((fitted(lacina_one) - mean(lacina$lnbdb))^2) /
  sum((lacina$lnbdb - mean(lacina$lnbdb))^2)
```

```
## [1] 0.29982
```

- Oder einfach:

```
summary(lacina_one)$r.squared
```

```
## [1] 0.29982
```

Vorsicht!

Nicht einfach nur R^2 maximieren! Annahmen beachten!

Multiple Regression

- Bisher:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Jetzt:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} + \epsilon_i$$

Lacina-Modell

TABLE 2

Ordinary Least Squares (OLS) Regressions of Battle Deaths in Civil Conflicts, 1946-2002

Independent Variable	*OLS Regression of ln Total Battle Deaths* Coefficient	(SE)	p-*Value*
Model 1[a]			
ln Duration	0.81	(0.12)	.000
ln Population	–0.044	(0.081)	.580
ln Military quality	0.10	(0.12)	.400
ln Gross domestic product	–0.19	(0.18)	.280
Cold war	0.67	(0.31)	.036
ln Percentage mountainous territory	0.10	(0.12)	.400
Democracy	–0.87	(0.36)	.017
Ethnic polarization	–0.98	(0.34)	.005
Religious polarization	0.12	(0.32)	.710
Intercept	9.50	(2.00)	.000
Model 2[b]			
ln Duration	0.86	(0.11)	.000
Cold war	0.59	(0.27)	.030
Democracy	–0.91	(0.33)	.006
Ethnic polarization	–1.00	(0.30)	.001
Intercept	8.60	(0.35)	.000

a. $n = 105$. Adjusted $R^2 = 0.40$.

b. $n = 114$. Adjusted $R^2 = 0.43$.

Replikation des Lacina-Modells

```
model1 <- lm(lnbdb ~ lnduration + lnpop + lnmilqual +
               lngdp + cw + lnmountain +
               democ + ethnicpolar + relpolar,
             data = lacina)
summary(model1)
```

```
##
## Call:
## lm(formula = lnbdb ~ lnduration + lnpop + lnmilqual + lngdp +
##     cw + lnmountain + democ + ethnicpolar + relpolar, data =
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3879 -0.9286 -0.0551  0.6936  3.5885
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.54259    1.97523   4.831 5.20e-06 ***
## lnduration   0.80722    0.11908   6.778 1.02e-09 ***
## lnpop       -0.04445    0.08065  -0.551  0.58287
```