Einführung in die statistische Datenanalyse mit R

Lineare Regression

David Benček

Wintersemester 2015/16

Wann verwenden wir eine lineare Regression?

Eine Regressionsanalyse hilft uns, die Beziehung einer

- **abhängigen Variable** Y und
- einer oder mehreren **unabhängigen Variablen** $X_1, X_2, ..., X_p$

zu erklären.

- Wenn $p = 1$, sprechen wir von einer *einfachen Regression*,
- bei $p > 1$ von einer *multivariaten Regression*.

Variable

Abhängige Variable

Die abhängige Variable Y muss eine stetige Variable sein.

Unabhängige Variable

Die unabhängigen Variablen $X_1, ..., X_p$ können stetige, diskrete oder kategoriale Variablen sein.

Erste Schritte

Vor jeder formalen Analyse sollten die Daten näher begutachtet werden:

- Fehler
- fehlende Werte
- Ausreißer
- unerwartete Verteilung einzelner Variablen
- unerwartete Muster

Erste Schritte II

Sehen die Daten aus wie wir es erwarten?

- Numerische Begutachtung:

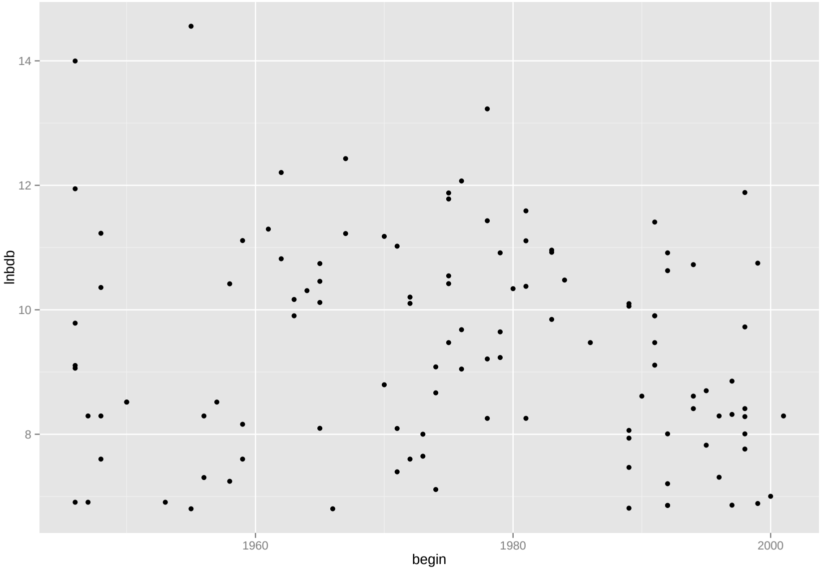
```
head(x)
summary(x)
cor(x, y)
```

- Grafische Begutachtung:

```
plot(x, y)
hist(x)
boxplot(x)
```

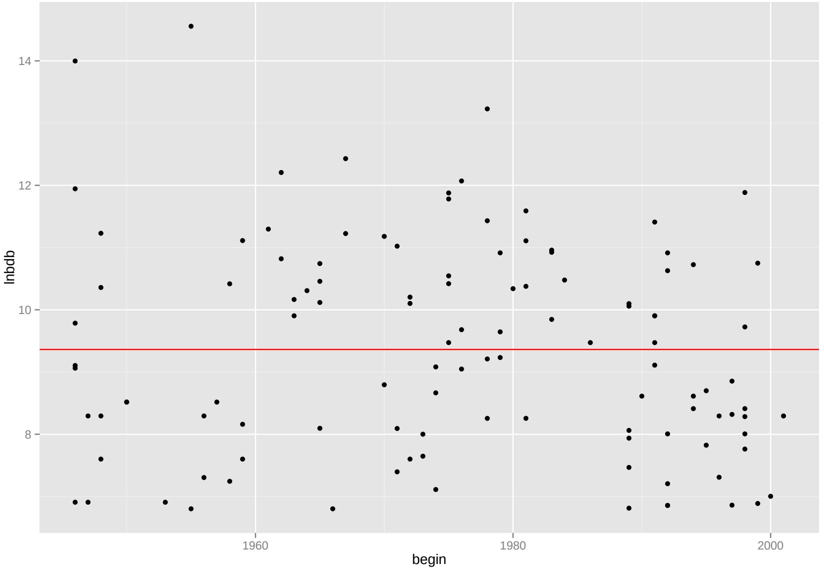
Statistisches Modell

Mittelwert



Statistisches Modell

Mittelwert



Statistisches Modell

Mittelwert

Wie gut erklärt der Mittelwert $\bar{x}$ die Beobachtungen?

$\rightarrow$ Gütemaß notwendig!

Varianz

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Lineare Regression mit einer Variable

Ziel: Abhängige Variable durch Zusammenhang mit einer unabhängigen Variable erklären.

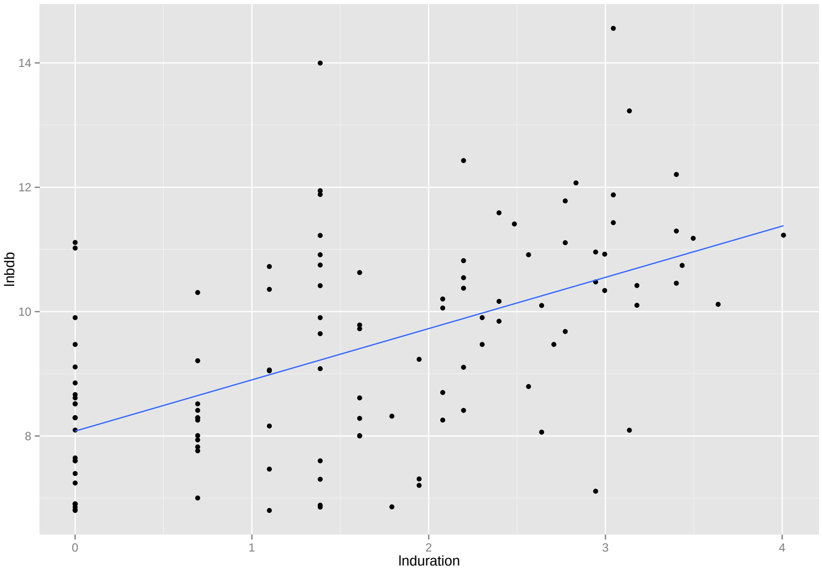
Variable

- Y: Konfliktintensität, gemessen als Anzahl der Toten.
- X: Dauer des Konflikts, gemessen in Jahren.

Im Datensatz haben wir Paare von Beobachtungen:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_{114}, y_{114})$$

Lineare Regression mit einer Variable (Plot)



Lineare Regression mit einer Variable (Modell)

Regressionsgleichung:

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

dabei gilt:

- Fehlerterm ϵ_i, normalverteilt mit Erwartungswert 0 und unabhängig.
- Lineare Funktion: $\beta_0 + \beta_1 x_i = E(Y|X = x_i)$

Unbekannte Paramter

- β_0 (Achsenabschnitt)
- β_1 (Steigung)

Schätzung der unbekannten Parameter

Ziel: Finde eine lineare Gleichung, die möglichst gut zu den Daten passt.

$\Rightarrow$ "Fitted Values" von y_i, gegeben durch

$$\hat{y}_i = b_0 + b_1 x_i$$

sollen so nah wie möglich an den beobachteten Werten y_i liegen.

Residuen

Residuen zeigen an, wie groß der Unterschied zwischen Beobachtung und "fitted value" ist:

$$e_i = y_i - \hat{y}_i$$

Schätzung der unbekannten Parameter

Methode der kleinsten Quadrate

Schätzung der Parameter b_0 und b_1 durch Minimierung der Summe quadrierter Residuen (residual sum of squares (RSS)):

$$\begin{aligned} RSS &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \end{aligned}$$

Schätzung in R

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

Output

```
##
## 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 ' '
##
## 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
```