Rückblick: Lineare Regression

Prof. Dr. Martin Elff
E-mail: `martin.elff@zu.de`

Fall Semester 2015

1 Von einfacher zu multipler Regression

Zweck der Regressionsanalyse

- Zentrale Konzepte
  - *Abhängige Variable* – auch *Regressand* oder in Englisch auch *response* oder *response variable*
  - *Independent variable* – manchmal auch als *Regressor* bezeichnet
- Verwendung
  - Beschreibung: Hat sich das durchschnittliche Einkommen erwerbstätiger Frauen erhöht?
    * Abhängige Variable: Einkommen
    * Unabhängige Variable: Zeit
  - Erklärung: Warum hat sich das durchschnittliche Einkommen erwerbstätiger Frauen erhöht?
    * Abhängige Variable: Einkommen
    * Unabhängige Variable: Bildung (z.B. gemessen durch Ausbildungsdauer)
  - Vorhersage: Um wie viel erhöht sich das zu erwartende Einkommen wenn man seinen Master oder Doktor macht?

Regression als Wahrscheinlichkeitsmodell

- Wir betrachten *bedingte* Verteilungen, ausgedrückt in der Dichte- bzw. Wahrscheinlichkeitsfunktion:
$$f(y|x_1, x_2, \ldots)$$

- Die Werte der bedingenden oder unabhängigen Variablen $(x_1, x_2, \ldots)$ können Werte von Zufallsvariablen sein
- müssen es aber nicht, da die Verteilung der unabhängigen Variablen *irrelevant* für die Modellkonstruktion ist.

Lineare Regressionsmodelle

- Modellierung des bedingten *Erwartungswerts* (oder „Durchschnitts")

  $$E(y|x_1, x_2, \ldots, x_p)$$

  wobei dieser bedingte Erwartungswert eine lineare Funktion der unabhängigen Variablen ist:

  $$\hat{y} := E(y|x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- Das ist äquivalent zu:

  $$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \qquad E(\epsilon) = 0$$

- Da die Werte der abhängigen und der unabhängigen Variablen von Beobachtung zu Beobachtung variieren, ist es angemessener, das so zu schreiben:

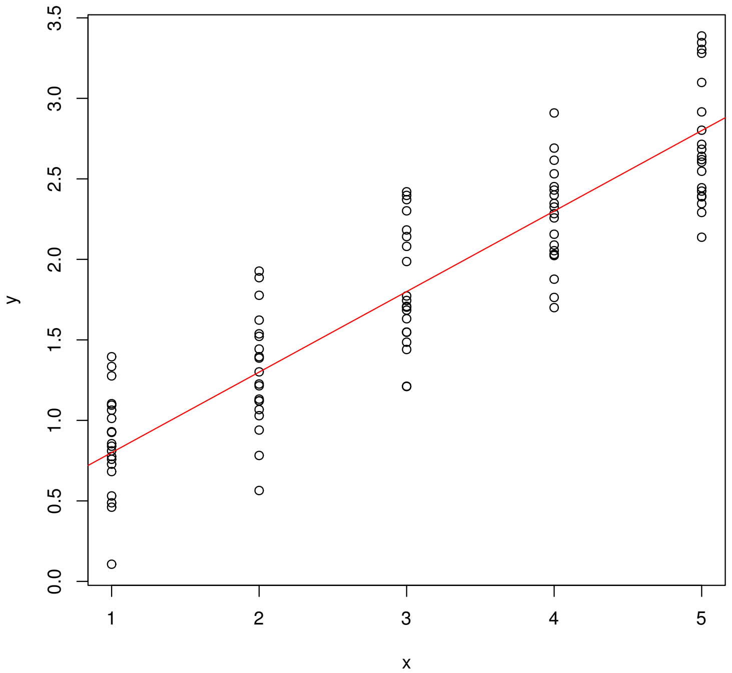
  $$Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i \qquad E(\epsilon_i) = 0$$

  wobei i die Beobachtungsnummer ist $(i = 1, \ldots, n)$

„Wahre" lineare Regressionen

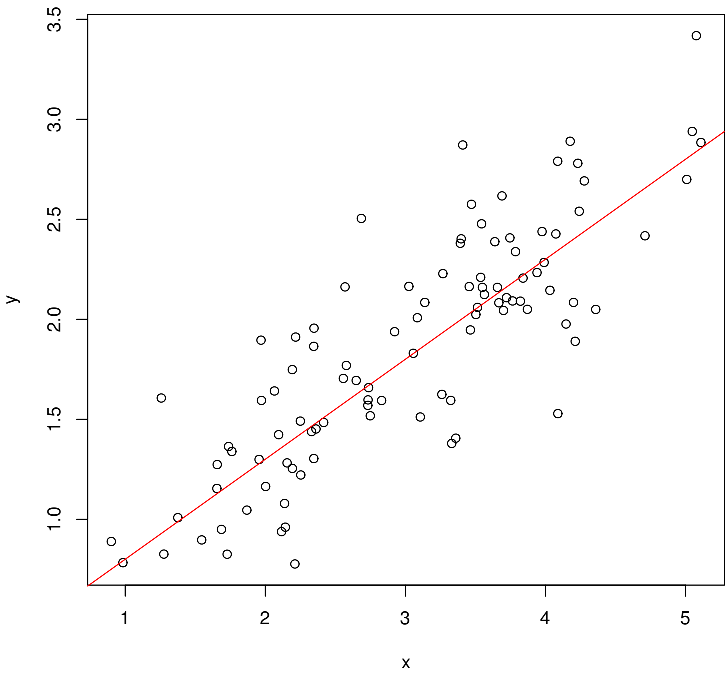
Mit „fixierter" unabhängiger Variablen

```
x <- rep(1:5,each=20)
y <- 0.3 + 0.5*x + rnorm(n=length(x),sd=.3)
plot(x,y)
abline(a=.3,b=.5,col="red")
```



Mit „zufälliger“ unabhängiger Variablen

```
x <- rnorm(n=100,mean=3)
y <- 0.3 + 0.5*x + rnorm(n=length(x),sd=.3)
plot(x,y)
abline(a=.3,b=.5,col="red")
```



Das Kriterium der kleinsten Quadrate (Ordinary Least Squares – OLS)

- Angenommen $y_i = \alpha + \beta x_i + \epsilon_i$ wobei $\mathrm{E}(\epsilon_i) = 0$
- Wähle $\hat{\alpha}$ und $\hat{\beta}$ so dass

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{wobei } \hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

  so klein wie möglich wird.
- Dies führt zum System der linearen Gleichungen:

$$\begin{aligned} n\alpha + \sum_{i=1}^{n} x_i \beta &= \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i \alpha + \sum_{i=1}^{n} x_i^2 \beta &= \sum_{i=1}^{n} x_i y_i \end{aligned}$$

Warum OLS?

- Intuition: Wir wollen α und β so wählen, dass sie die beobachteten Werte der abhängigen Variable so gut wie möglich „vorhersagen“.

- Was bedeutet „gut“ hier? Kleine *Residuen* $r_i = y_i - \hat{y}_i$
- Man könnte den mittleren Absolutwert $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$ nehmen …
- aber die Quadratsumme (Residual Sum of Squares) $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ ist mathematisch einfacher zu handhaben
- und außerdem hat OLS bestimmte Optimalitätseigenschaften.

Das Gauss-Markov-Theorem

- Wenn $y_i = \alpha + \beta x_i + \epsilon_i$ für bestimmte Werte von α und β
- und wenn die Regressionsfehler ϵ_i unkorreliert sind und eine endliche Varianz haben (sie müssen nicht normalverteilt sein)
- dann sind die OLS-Schätzungen $\hat{\alpha}$ und $\hat{\beta}$
  - unverzerrt, d.h. $\mathrm{E}(\hat{\alpha}) = \alpha <$ und $\mathrm{E}(\hat{\beta}) = \beta$ und
  - haben den minimalen erwarteten Schätzfehler $\mathrm{E}\left((\hat{\alpha} - \alpha)^2\right)$ und $\mathrm{E}\left((\hat{\beta} - \beta)^2\right)$ von allen unverzerrten linearen Schätzern.

Berechnung von OLS-Schätzungen mit *R*

- Die *R*-Funktion um OLS-Schätzungen von Regressionskoeffizienten zu erhalten ist **`lm()`**. („lm“ steht für **l**inear **m**odel)
- Aufruf mit abhängiger Variable **`y`** und unabhängiger Variablen **`x`**:

  ```
  lm(y~x)
  ```
- Wenn **`y`** oder **`x`** im Data Frame **`MyData`** enthalten sind:

  ```
  lm(y~x,data=MyData)
  ```

Ein Beispiel

```
# Wir erzeugen künstliche Daten
set.seed(2)
x <- rep(1:5,each=20)
y <- 0.3 + 0.5*x + rnorm(n=length(x),sd=.3)
DataFrame1 <- data.frame(x,y)
x <- rnorm(n=100,mean=3)
y <- 0.3 + 0.5*x + rnorm(n=length(x),sd=.3)
DataFrame2 <- data.frame(x,y)
# Der Arbeitsbereich wird aufgeräumt
rm(x,y)
```

```
# Wir holen uns die Daten vom Data Frame
lm1 <- lm(y~x,data=DataFrame1)
# Das "lm"-Object
print(lm1)
```

Output

```
Call:
lm(formula = y ~ x, data = DataFrame1)

Coefficients:
(Intercept)            x
      0.390        0.467
```

```
# Die Koeffizienten
coef(lm1)
```

Output

```
(Intercept)           x
      0.390       0.467
```

```
# Die Zusammenfassung des Modells
summary(lm1)
```

Output

```
Call:
lm(formula = y ~ x, data = DataFrame1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7595 -0.2361 -0.0433  0.2196  0.6634

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   0.3903     0.0813     4.8 0.0000056 ***
x             0.4668     0.0245    19.1   < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.347 on 98 degrees of freedom
Multiple R-squared: 0.787,        Adjusted R-squared: 0.785
F-statistic:  363 on 1 and 98 DF,  p-value: <2e-16
```

Interpretation

- **`Estimate`**: Schätzwert von Regressionskonstante (Intercept) und Regressionskoeffizient
- **`Std. Error`**: (Geschätzter) Standardfehler von Konstante und Koeffizienten, ein Maß der Unsicherheit der Schätzwerte (je kleiner, desto besser). Es handelt sich dabei um die Quadratwurzel der Schätzvarianz $\mathrm{Var}(\hat{\alpha})$ bzw. $\mathrm{Var}(\hat{\beta})$
- **`t value`**: Wert der Teststatistik für die Nullhypothese $\alpha = 0$ bzw. $\beta = 0$
- **`Pr(>`**t|)|: Wahrscheinlichkeit, einen Wert der Teststatistik mindestens so groß wie die berechnete zu erhalten, *wenn die Nullhypothese zutrifft*
- **`Residual standard error`**: Eine Schätzung der Standardabweichung von ϵ_i
- **`Multiple R-squared`**: der Determinationskoeffizient, ein Maß für die „Anpassungsgüte" (goodness of fit) des Regressionsmodells.
- **`F-statistic`**: Wert der Teststatistik für die Nullhypothese das *alle* Koeffizienten (außer der Konstanten) gleich Null sind.

Anpassungsgüte

- Determinationskoeffizient:

  $$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  Im bivariaten Fall ist R^2 gleich dem Quadrat der Korrelation zwischen der abh. und der unabh. Variable.
- Korrigierter Determinationskoeffizient, adjusted R^2:

  $$R_a^2 = 1 - \frac{n-1}{n-k}(1 - R^2)$$

  (k ist die Anzahl der unabhängigen Variablen)
- F-Statistik:

  $$F = \frac{n-k-1}{k}\frac{R^2}{1-R^2}$$

Die Struktur von Aufrufen zur Berechnung von Regressionsmodellen

Funktionen für die Schätzung von Regressionsmodellen und deren Verallgemeinerungen haben üblicherweise die folgenden Argumente:

- Ein **`formula`**-Argument, z.B. **`y~x+z`** das die abhängige Variable und die modellierten Effekte der unabhängigen Variablen festlegt. Das ist fast immer das erste Argument.

- Ein optionales **`data`**-Argument, das festlegt, aus welchem Data Frame die Variablen im Modell stammen. Es ist nicht erforderlich, wenn die Variablen im Arbeitsbereich „sichtbar“ sind.
- Einige zusätzlichen Argumente, die zusätzliche schätzbare Modellaspekte beschreiben.
- Ein **`control`**-Argument für die Steuerung des iterativen Schätzverfahrens (sofern für den Modelltyp nötig, nicht für lineare Regression).

Derartige Funktionen geben ein Objekt zurück, dass weiter verarbeitet werden kann.

Verwertung von Schätzergebnissen

Mehrere Funktionen sind verfügbar um auf Schätzergebnisse zuzugreifen und sie weiter zu verarbeiten.

- **`coef()`**, **`coefficients()`** gibt die Regressionskoeffizienten zurück
- **`summary()`** gibt eine umfassende Zusammenfassung des geschätzten Modells, einschließlich Schätzwerte, Standardfehler, Signifikanzniveaus und Goodness-of-fit-Statistiken.
- **`fitted()`** gibt die angepassten (vorhergesagten) Werte der abhängigen Variablen zurück.
- **`residuals()`** gibt, die Residuen, den „unerklärten“ Teil der abhängigen Variablen zurück.
- **`predict()`** erlaubt Vorhersagen innerhalb der Stichprobe und außerhalb der Stichprobe (d.h. für andere Daten als die für die Schätzung verwendeten)
- **`termplot()`** erlaubt eine graphische Darstellung des Einflusses unabhängiger Variablen
- **`plot()`** gibt Diagramme zur Regressionsdiagnostik

Ein reales Beispiel: Berufsprestige, Einkommen und Bildung

- Abhängige Variable: **`prestige`**, Berufsprestige - das Prestige das ein spezifischer Beruf mit sich bringt. Spezifischer: Pineo-Porter-Prestige-Score für den Beruf, von eine Bevölkerungsumfrage Mitte der 1960er Jahre.
- Unabhängige Variable: **`education`**, durchschnittliche Bildung der Inhaber des jeweiligen Berufs, in Ausbildungsjahren (Daten von 1971)
- Weitere unabhängige Variable: **`income`**, mittleres Einkommen der Berufstätigen in Dollar (Daten von 1971)
- Die Variablen sind im Data Frame **`Prestige`**, der im Paket **`car`** (Companion to *Applied Regression* – ein Buch von John Fox)

```
library(car)
# Bildungsdauer als unabh. Variable
lm.prestige_education <- lm(prestige~education,
                            data=Prestige)
summary(lm.prestige_education)
```

Output

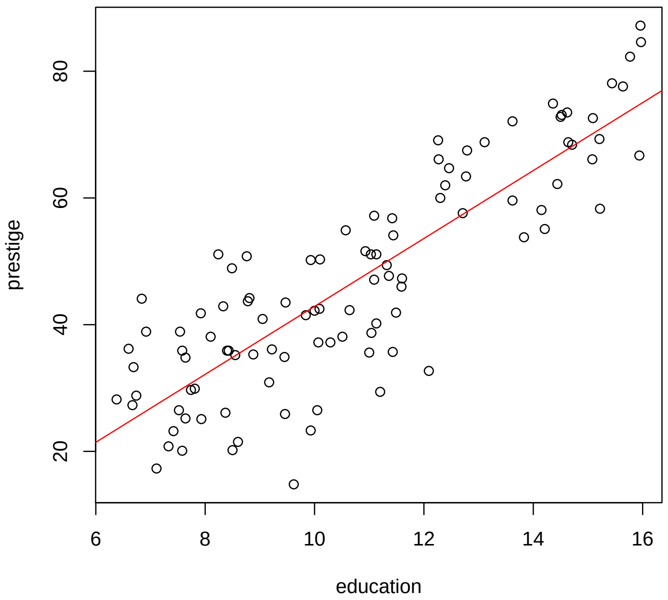
```
Call:
lm(formula = prestige ~ education, data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max 
-26.040  -6.523   0.661   6.743  18.164 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -10.732      3.677   -2.92   0.0043 ** 
education      5.361      0.332   16.15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.1 on 100 degrees of freedom
Multiple R-squared: 0.723,	Adjusted R-squared: 0.72 
F-statistic:  261 on 1 and 100 DF,  p-value: <2e-16
```

```
plot(prestige~education,data=Prestige)
abline(lm.prestige_education,col="red")
```




```
# Einkommen als unabh. Variable
lm.prestige_income <- lm(prestige~income,
                         data=Prestige)
summary(lm.prestige_income)
```

Output

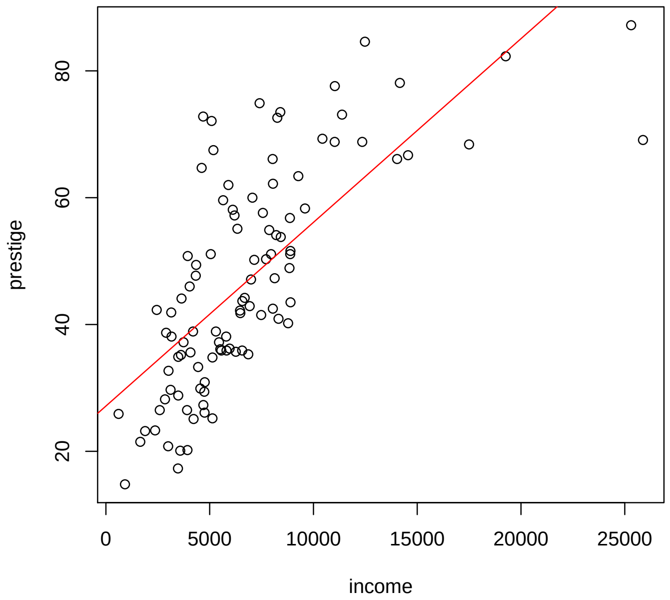
```
Call:
lm(formula = prestige ~ income, data = Prestige)

Residuals:
   Min     1Q Median     3Q    Max
-33.01  -8.38  -2.38   8.43  32.08

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.141176   2.267704    12.0   <2e-16 ***
income       0.002897   0.000283    10.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.1 on 100 degrees of freedom
Multiple R-squared: 0.511,	Adjusted R-squared: 0.506
F-statistic:  105 on 1 and 100 DF,  p-value: <2e-16
```

```
plot(prestige~income,data=Prestige)
abline(lm.prestige_income,col="red")
# Offensichtlich keine besonders gute Passung
```




```
# Logarithmus des Einkommens als unabh. Variable
lm.prestige_logincome <- lm(prestige~log(income),
                            data=Prestige)
summary(lm.prestige_logincome)
```

Output

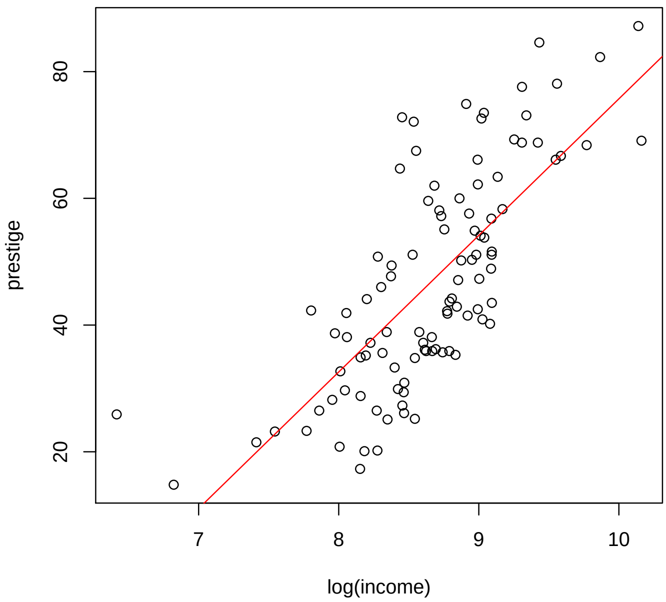
```
Call:
lm(formula = prestige ~ log(income), data = Prestige)

Residuals:
   Min      1Q  Median      3Q     Max
-19.11   -9.34   -1.22    8.10   30.45

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -139.86      16.95   -8.25  6.6e-13 ***
log(income)    21.56       1.95   11.04  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 11.6 on 100 degrees of freedom
Multiple R-squared: 0.549,        Adjusted R-squared: 0.545
F-statistic:  122 on 1 and 100 DF,  p-value: <2e-16
```

```
plot(prestige~log(income),data=Prestige)
abline(lm.prestige_logincome,col="red")
```




```
lm.prestige_biv <- lm(prestige~education+log(income),
                            data=Prestige)
summary(lm.prestige_biv)
```

Output

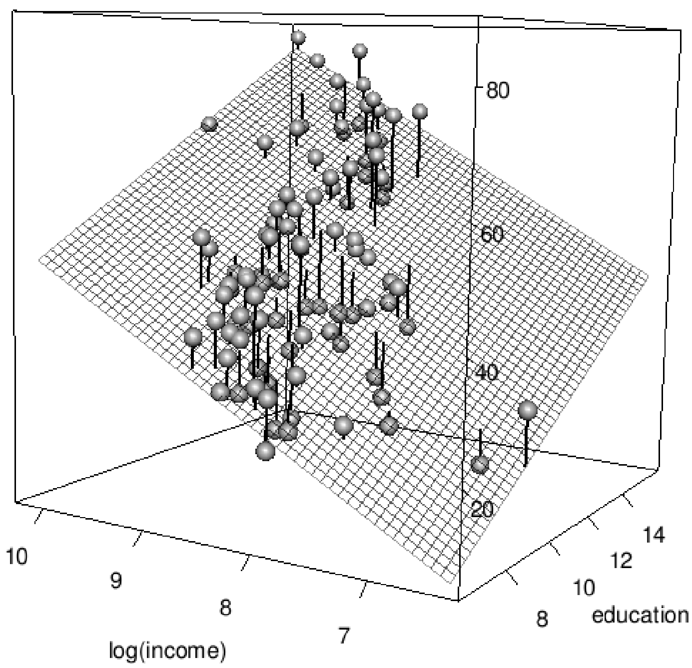
```
Call:
lm(formula = prestige ~ education + log(income), data = Prestige)

Residuals:
    Min       1Q   Median       3Q      Max
-17.035   -4.566   -0.186    4.058   18.127

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -95.194     10.998   -8.66  9.3e-14 ***
education      4.002      0.312   12.85  < 2e-16 ***
```

```
log(income)   11.437      1.437    7.96  2.9e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.14 on 99 degrees of freedom
Multiple R-squared: 0.831,       Adjusted R-squared: 0.828
F-statistic:  243 on 2 and 99 DF,  p-value: <2e-16
```



2 Unstandardisierte und standardisierte Koeffizienten

- In Regression mit mehreren unabh. Variablen: diese of auf unterschiedlichen Skalen gemessen
- Wie den Einfluss des Einkommens in Dollar vergleichen mit dem Einfluss der Ausbildungsdauer in Jahren auf die Partei-Identifikation?
- Standardisierung:
  - Variable $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ im Datensatz
  - Werte der „z-standardisierten“ Variablen $\boldsymbol{z}$:

$$z_i = \frac{x_i - \bar{x}}{\mathrm{sd}(\boldsymbol{x})} \quad \Rightarrow \bar{z} = 0 \text{ and } \mathrm{sd}(\boldsymbol{z}) = 1$$

(sd($\boldsymbol{x}$) meint die Standardabweichung von $\boldsymbol{x}$)

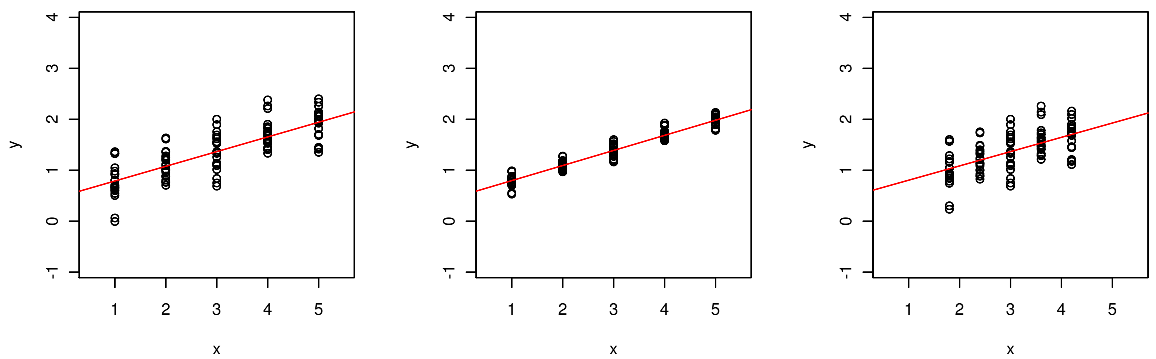
- Standardisierte Regressionskoeffienten: Koeffizienten, die man bekommen würde, wenn sowohl abh. als auch unabh. Variablen z-standardisiert wären.
- Beziehung zwischen standardisierten und unstandardisierten Koeffizienten:
  - $\beta_{0,\text{std}} = 0$ (wenn abh. und unabh. Variablen standardisiert sind, dann haben sie alle den Mittelwert Null, und die „standardisierte“ Regressionskonstante wird auch Null sein.
  - Für β_j ($j = 1, 2, \ldots, k$)
$$\beta_{j,\text{std}} = \frac{\text{sd}(\boldsymbol{x}_j)}{\text{sd}(\boldsymbol{y})}\hat{\beta}_j$$
  - Standardisierte Koeffizienten werden (leider!) häufig „Beta-Gewichte“ genannt (weil sie in SPSS und Stata so genannt werden)

Wann soll man standardisierte Koeffizienten betrachten und wann nicht?

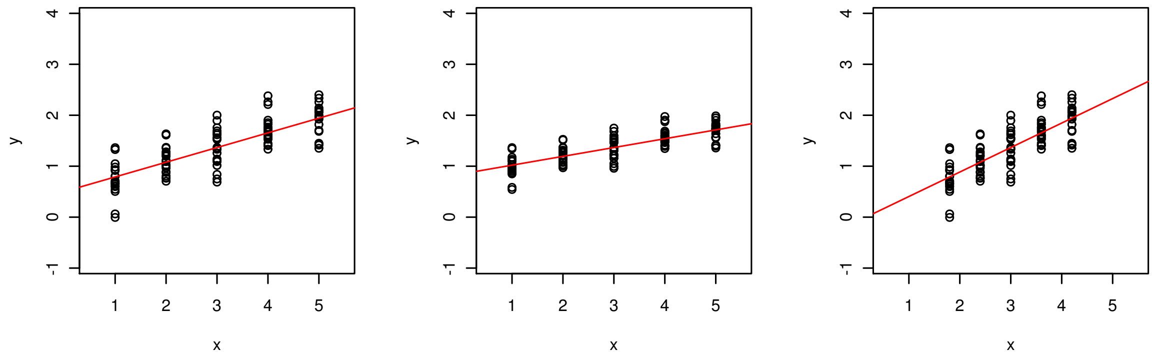
- Standardisierte Koeffizienten sind nützlich wenn
  - der Einfluss *unterschiedlicher* Variablen
  - innerhalb des *selben* Modells und der *selben* Stichprobe verglichen werden soll
- Standardisierte Koeffizienten sollten nicht benutzt werden um
  - den Einfluss der *selben* Variablen
  - zwischen *unterschiedlichen* Populationen oder *unterschiedlichen* Stichproben zu vergleichen

Illustration

Gleiche unstandardisierte, unterschiedliche standardisierte Koeffizienten



Unterschiedliche unstandardisierte, gleiche standardisierte Koeffizienten



Nochmals ein „reales“ Beispiel: Berufsprestige, Einkommen und Bildung

```
lm.prestige_biv <- lm(prestige~education+log(income),
                            data=Prestige)
summary(lm.prestige_biv)
```

Output

```
Call:
lm(formula = prestige ~ education + log(income), data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.035  -4.566  -0.186   4.058  18.127 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -95.194     10.998   -8.66  9.3e-14 ***
education      4.002      0.312   12.85  < 2e-16 ***
log(income)   11.437      1.437    7.96  2.9e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.14 on 99 degrees of freedom
Multiple R-squared: 0.831,	Adjusted R-squared: 0.828 
F-statistic:  243 on 2 and 99 DF,  p-value: <2e-16
```

```
# Wir wollen die Relevanz der unabh. Variablen vergleichen,
# daher holen wir uns standardisierte Koeffizienten
# 'scale' does the trick
```

```
lm.prestige_biv.std <- lm(scale(prestige)~
                                  scale(education)
                                  +scale(log(income))-1,
                            data=Prestige)
summary(lm.prestige_biv.std)
```

Output

```
Call:
lm(formula = scale(prestige) ~ scale(education) + scale(log(income)) -
    1, data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9901 -0.2654 -0.0108  0.2359  1.0536

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
scale(education)     0.6347     0.0492    12.9  < 2e-16 ***
scale(log(income))   0.3932     0.0492     8.0  2.3e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.413 on 100 degrees of freedom
Multiple R-squared: 0.831,       Adjusted R-squared: 0.828
F-statistic:  246 on 2 and 100 DF,  p-value: <2e-16
```

3 Hypothesentests und statistische Signifikanz

Statistische Hypothesen

- Eine *statistische Hypothese* ist, grob gesprochen, eine Vermutung über die Wahrscheinlichkeitsverteilung von der die beobachteten Daten generiert worden sind.
- Ein statistischer Test ist ein Verfahren, in dem der Wahrheitsgehalt einer statistischen Hypothese überprüft wird.
- Genauer gesagt, ein statistischer Test hilft zu entscheiden zwischen einer
  - *Nullhypothese* – gewöhnlich eine sehr spezifische Aussage über die Wahrscheinlichkeitsverteilung oder ihre Parameter, z.B. dass der Erwartungswert Null ist.

- *Alternativhypothese* – gewöhnlich eine weniger spezifische Aussage, die als zutreffend akzeptiert wird, wenn die Nullhypothese verworfen wird.

  Sie ist nicht einfach eine Verneinung der Nullhypothese, sondern hat gewisse grundlegende Annahmen mit ihr gemein.

4 F-Tests für den Modellvergleich

Genestete und nicht-genestete Modelle

- Ein Modell M_1 ist genested (eingebettet) in ein anderes Modell M_2, wenn M_1 und M_2 die gleiche abhängige Variable haben und jede unabhängige Variable in M_1 auch eine unabhängige Variable in M_2 ist.
- Ein Beispiel für ein Modell, das in ein anderes eingebettet ist:

$$M_1 : \widehat{\texttt{rent}} = \alpha + \beta_1 \texttt{hhinc}$$
$$M_2 : \widehat{\texttt{rent}} = \alpha + \beta_1 \texttt{hhinc} + \beta_2 \texttt{sqfeet}$$

- Ein Beispiel für nicht eingebettete Modelle:

$$M_1 : \widehat{\texttt{rent}} = \alpha + \beta_1 \texttt{hhinc}$$
$$M_2 : \widehat{\texttt{rent}} = \alpha + \beta_1 \texttt{sqfeet}$$

- Noch ein Beispiel für nicht eingebettete Modelle:

$$M_1 : \widehat{\texttt{rent}} = \alpha + \beta_1 \texttt{hhinc}$$
$$M_2 : \widehat{\texttt{rent}} = \alpha + \beta_1 \ln(\texttt{hhinc})$$

F-Test für den Vergleich zweier Modelle

- Einfacheres Modell M_0 mit weniger Parametern (Koeffizienten) k_0 als Modell M_1 (mit k_1 Parametern) in das es eingebettet ist.
- Nullhypothese: Alle Koeffizienten, die in M_1 aber nicht in M_0 enthalten sind sind „in Wirklichkeit“ gleich Null.
- Alternativhypothese: mindestens einer der in M_1 aber nicht in M_0 enthaltenen Koeffizienten ist von Null verschieden.

- Test-Statistik:

$$F = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{k_1 - k_0} \Bigg/ \frac{\mathrm{RSS}_1}{n - k_1 - 1} = \frac{n - k_1 - 1}{k_1 - k_0} \frac{R_1^2 - R_0^2}{1 - R_1^2}$$

  wobei RSS_1 die Fehlerquadratsumme von Modell M_1 und RSS_0 die Fehlerquadratsumme von Modell M_0 ist; R_1^2 und R_0^2 sind die jeweiligen Determinationskoeffizienten der Modelle M_1 und M_0.

Ein Beispiel für einen inkrementellen F-Test

```
library(car)
lm.prestige_ii <- lm(prestige~education+log(income),
                            data=Prestige,
                            subset=is.finite(type))
                            # Fehlende Werte werden aussortiert,
                            # um eine Fehlermeldung zu vermeiden

# Dem Modell 'lm.prestige_ii' wird der Faktor
# 'type' hinzugefügt
lm.prestige_iii <- update(lm.prestige_ii,
                          .~.+type)
# Das ist äquivalent mit
# lm.prestige_iii <- lm(prestige~education+log(income)+type,
#                             data=Prestige)

# Ein F-Test der zwei Modelle vergleicht und dabei
# die statistische Signifikanz des Einflusses von
# 'type' ermittelt
anova(lm.prestige_ii,lm.prestige_iii)
```

```
Output
Analysis of Variance Table

Model 1: prestige ~ education + log(income)
Model 2: prestige ~ education + log(income) + type
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     95 4565
2     93 4096  2       469 5.32 0.0065 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Hier wird das Modell

$$M_0 : \widehat{\texttt{prestige}} = \alpha + \beta_1 \texttt{education} + \beta_2 \ln(\texttt{income})$$

  verglichen mit dem Modell

$$\begin{aligned} M_1 : \widehat{\texttt{prestige}} &= \alpha + \beta_1 \texttt{education} + \beta_2 \ln(\texttt{income}) \\ &\quad + \beta_3 d_{\texttt{type=="prof"}} + \beta_4 d_{\texttt{type=="wc"}} \end{aligned}$$

- M_1 verbraucht zwei Freiheitsgrade mehr als M_0, da es zwei Koeffizienten mehr enthält.
- Der Wert der F-Statistik ist 5.32 und hat einen p-Wert von 0.0065 (oder 0.65 Prozent) und ist daher statistisch signifikant.