Thema 3: Darstellung eines linearen Modells

JSA

3/24/23

# https://quarto.org/docs/authoring/markdown-basics.html  
  
# knitr::purl("t3\_lm/t3\_ablauf.qmd", output = "t3\_lm/t3\_ablauf.R")  
  
suppressPackageStartupMessages(library(readr))  
suppressPackageStartupMessages(library(tidyverse))  
suppressPackageStartupMessages(library(ggplot2))  
suppressPackageStartupMessages(library(flextable))  
suppressPackageStartupMessages(library(officer))  
suppressPackageStartupMessages(library(car))  
suppressPackageStartupMessages(library(broom))

## Beispieldaten für Teil 3

Im R-Paket **car** bzw. **carData** ist der Datensatz Prestige enthalten.

Dieser enthält die folgenden Variablen.

This data frame contains the following columns (angezeigt mit *help(Prestige)*):

* education: Average education of occupational incumbents, years, in 1971.
* income: Average income of incumbents, dollars, in 1971.
* women: Percentage of incumbents who are women.
* prestige: Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.
* census: Canadian Census occupational code.
* type: Type of occupation. A factor with levels (note: out of order): bc, Blue Collar; prof, Professional, Managerial, and Technical; wc, White Collar.

head(Prestige) %>% flextable() %>% autofit()

| education | income | women | prestige | census | type |
| --- | --- | --- | --- | --- | --- |
| 13.11 | 12,351 | 11.16 | 68.8 | 1,113 | prof |
| 12.26 | 25,879 | 4.02 | 69.1 | 1,130 | prof |
| 12.77 | 9,271 | 15.70 | 63.4 | 1,171 | prof |
| 11.42 | 8,865 | 9.11 | 56.8 | 1,175 | prof |
| 14.62 | 8,403 | 11.68 | 73.5 | 2,111 | prof |
| 15.64 | 11,030 | 5.13 | 77.6 | 2,113 | prof |

Es wird ein lineares Modell mit

analysiert.

Die Grundausgabe erfolgt mit dem Befehl **summary()** erstellt.

Die Ausgabe ist nicht sehr übersichtlich.

lm1 <- lm(prestige ~ education + log2(income) + women, data = Prestige)  
summary(lm1)

Call:  
lm(formula = prestige ~ education + log2(income) + women, data = Prestige)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-17.364 -4.429 -0.101 4.316 19.179   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -110.9658 14.8429 -7.476 3.27e-11 \*\*\*  
education 3.7305 0.3544 10.527 < 2e-16 \*\*\*  
log2(income) 9.3147 1.3265 7.022 2.90e-10 \*\*\*  
women 0.0469 0.0299 1.568 0.12   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 7.093 on 98 degrees of freedom  
Multiple R-squared: 0.8351, Adjusted R-squared: 0.83   
F-statistic: 165.4 on 3 and 98 DF, p-value: < 2.2e-16

Ohne den Befehl **summary()** werden nur die Schätzer für das Modell ausgegeben.

lm1

Call:  
lm(formula = prestige ~ education + log2(income) + women, data = Prestige)  
  
Coefficients:  
 (Intercept) education log2(income) women   
 -110.9658 3.7305 9.3147 0.0469

Das R-Paket **broom** dient dazu, ein Modell formatiert auszugegeben.

Der Befehl **tidy()** gibt die Schätzer tabellarisch aus.

ft <- tidy(lm1) %>% flextable()  
ft

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -110.96582409 | 14.84292810 | -7.476006 | 0.000000000032697743335509 |
| education | 3.73050783 | 0.35438304 | 10.526767 | 0.000000000000000008730063 |
| log2(income) | 9.31466643 | 1.32651512 | 7.021907 | 0.000000000289556558862053 |
| women | 0.04689514 | 0.02989886 | 1.568459 | 0.119997441726498343950169 |

Der Befehl **glance()** gibt weitere Modellparameter auf einen Blick aus (at a glance).

# Modellparameter auf einen Blick  
gl <- glance(lm1)  
gl

# A tibble: 1 × 12  
 r.squared adj.r.squa…¹ sigma stati…² p.value df logLik AIC BIC devia…³  
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 0.835 0.830 7.09 165. 3.21e-38 3 -343. 695. 708. 4930.  
# … with 2 more variables: df.residual <int>, nobs <int>, and abbreviated  
# variable names ¹​adj.r.squared, ²​statistic, ³​deviance

# glance()-Ausgabe als data frame  
df\_gl <- data.frame(t(gl))  
df\_gl

t.gl.  
r.squared 8.350957e-01  
adj.r.squared 8.300476e-01  
sigma 7.092598e+00  
statistic 1.654280e+02  
p.value 3.213281e-38  
df 3.000000e+00  
logLik -3.425147e+02  
AIC 6.950295e+02  
BIC 7.081543e+02  
deviance 4.929885e+03  
df.residual 9.800000e+01  
nobs 1.020000e+02

# Reihennamen in Variable umwandeln  
df\_gl$parameter <- rownames(df\_gl)  
rownames(df\_gl) <- NULL  
df\_gl <- df\_gl %>% dplyr::rename("value" = "t.gl.")  
df\_gl$value <- sprintf("%.4f", df\_gl$value)  
df\_gl %>% select(parameter, value)

parameter value  
1 r.squared 0.8351  
2 adj.r.squared 0.8300  
3 sigma 7.0926  
4 statistic 165.4280  
5 p.value 0.0000  
6 df 3.0000  
7 logLik -342.5147  
8 AIC 695.0295  
9 BIC 708.1543  
10 deviance 4929.8854  
11 df.residual 98.0000  
12 nobs 102.0000

# Funktion zum Formatieren der p-Werte  
format\_p <- function(p) {  
 p\_c <- sprintf("%.3f", p)  
 if (p < 0.001) {  
 p\_c <- "< 0.001 \*\*\*"  
 } else if (p < 0.01) {  
 p\_c <- paste0(p\_c, " \*\*")  
 } else if (p < 0.05) {  
 p\_c <- paste0(p\_c, " \*")  
 }  
 p\_c  
}  
  
df <- tidy(lm1)   
df$p.value <- sapply(df$p.value, format\_p, simplify = TRUE)  
  
df$estimate <- sprintf("%.4f", df$estimate)  
df$std.error <- sprintf("%.4f", df$std.error)  
df$statistic <- sprintf("%.4f", df$statistic)  
  
ft <- df %>% flextable() %>% autofit()  
  
my\_border = fp\_border(color="black", width = 2)  
  
ft\_lm <- ft %>%   
 add\_header\_lines(c("Study title", "1.2.3 Linear Model", "FAS")) %>%  
 add\_footer\_lines(c("Abbreviations and comments", "Program name / Run date / Data cut date")) %>%  
 border\_remove() %>%   
 set\_header\_labels(  
 term = "Variable",  
 estimate = "Estimate",   
 std.error = "SE",  
 statistic = "Statistic",  
 p.value = "p") %>%  
 hline\_top(border = my\_border, part = "footer") %>%  
 hline\_bottom(border = my\_border, part = "header") %>%  
 hline\_bottom(border = my\_border, part = "body") %>%  
 hline(i = c(3), border = my\_border, part = "header") %>%  
 align(j = c(2:5), align = "right", part = "header") %>%  
 align(j = c(2:5), align = "right", part = "body")  
ft\_lm

| Study title | | | | |
| --- | --- | --- | --- | --- |
| 1.2.3 Linear Model | | | | |
| FAS | | | | |
| Variable | Estimate | SE | Statistic | p |
| (Intercept) | -110.9658 | 14.8429 | -7.4760 | < 0.001 \*\*\* |
| education | 3.7305 | 0.3544 | 10.5268 | < 0.001 \*\*\* |
| log2(income) | 9.3147 | 1.3265 | 7.0219 | < 0.001 \*\*\* |
| women | 0.0469 | 0.0299 | 1.5685 | 0.120 |
| Abbreviations and comments | | | | |
| Program name / Run date / Data cut date | | | | |

my\_doc <- read\_docx() %>% body\_add\_flextable(value = ft\_lm)  
print(my\_doc, target = "t3\_table\_lm.docx")