Die multiple Regression

Stephanie Geise

Table of contents

1 Data Management **1**

2 Multiple lineare Regression **3**
2.1 Inhaltliche Interpretation des Outputs: Was bedeutet das Ergebnis? 4
2.2 Zusätzliche Voraussetzungsprüfung bei der multiplen linearen Regression: Liegt Multikollinearität vor? 5
2.3 Vorhersage des multivariaten Modells 6
2.4 Multiple Regression mit Dummy-Codierung der kategorialen Variable "sex" . . 6
2.4.1 Vorbereitung der Daten zum Zusammenhang von Alter, Bildung, Vertrauen ins TV, Geschlecht und TV-Konsum 6
2.5 Regressionsmodell zum Zusammenhang von Alter, Bildung, TV-Vertrauen, Geschlecht und TV-Konsum 7
2.5.1 Inhaltliche Interpretation des Outputs: Was bedeutet das Ergebnis? . . 8

In diesem Teilkapitel lernen wir nun - wie angekündigt - die *multiple lineare Regression* kennen, die es erlaubt, Zusammenhänge zwischen *mehreren x-Variablen* und einer y-Variablen zu analysieren.

1 Data Management

Zunächst laden wir wieder die Pakete des `tidyverse` und das Paket `broom` um die normale Ausgabe der Funktion `lm` (für die Berechnung linearer Modelle) in ein etwas anschaulicheres Format umwandeln zu können. Außerdem laden wir das Paket `performance`, dass wir für die Voraussetzungsprüfung brauchen, sowie die Pakete `lmtest` und `sandwich`, mit der wir fehlende Voraussetzungen korrigieren können (siehe unten).

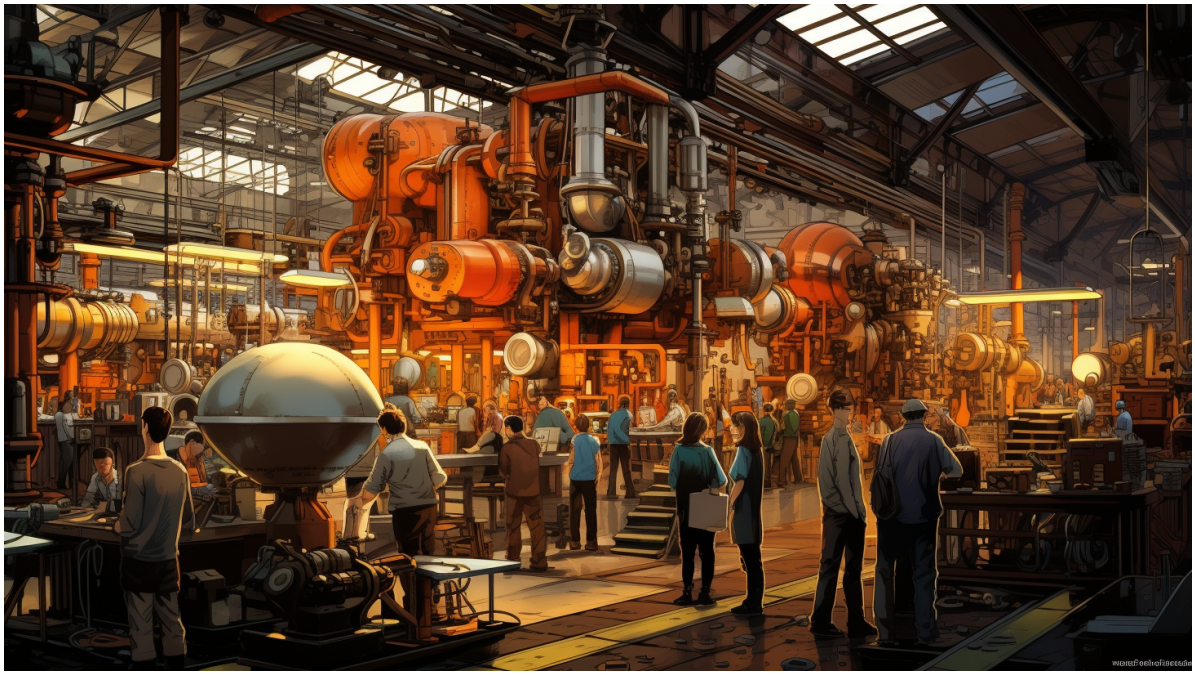

Figure 1: Eine Halle voller Maschinen, Bild generiert von Midjourney

```
if(!require("pacman")) {install.packages("pacman");library(pacman)}
p_load(tidyverse, lm.beta, lmtest, performance, easystats, haven, broom, see, haven, sandwich

theme_set(theme_classic())
options(scipen = 999)
```

Die Regression rechnen wir wieder auf Basis Allbus-Datensatzes, den wir entsprechend einlesen:

```
daten = haven::read_dta("Datensatz/Allbus_2021.dta")
```

Als abhängige Variable nutzen wir für unser Regressionsmodell wieder die TV-Nutzung (`lm02`); als unabhängige Variablen schauen wir uns wie beim letzten Mal das Alter, sowie heute zusätzlich die Variablen Bildung (`educ`) und Vertrauen ins Fernsehen (`pt09`). Damit der Output etwas nachvollziehbarer wird, benennen wir diese Variablen mit dem `rename`-Befehl zunächst wieder um. Außerdem filtern wir auch wieder die missings heraus (z.B. -9=Keine Angabe):

```
daten <- daten %>%
  rename(TV_Konsum = lm02)%>%
  rename(TV_Vertrauen= pt09)%>%
```

```r
  rename(Alter = age)%>%
  rename(Bildung = educ)%>%
  filter(between(TV_Konsum, 0, 1500))%>%
  filter(between(TV_Vertrauen, 1, 7))%>%
  filter(between(Alter, 18, 100))%>%
  filter(between(Bildung, 1, 5))
```

2 Multiple lineare Regression

Erinnerung: Einfache lineare Regression

Aus den letzten Sitzungen wissen wir bereits, dass das Alter einen signifikanten, positiven Einfluss auf die TV-Nutzung hat. Wir wissen aber auch, dass die TV-Nutzung nicht alleine über das Alter erklärt werden kann (weil unser R2 recht gering war).

```r
model <- lm(TV_Konsum ~ Alter, data = daten)
summary(lm.beta(model))
```

```
Call:
lm(formula = TV_Konsum ~ Alter, data = daten)

Residuals:
    Min      1Q  Median      3Q     Max
-243.01  -67.15  -22.38   34.77 1294.77

Coefficients:
            Estimate Standardized Std. Error t value             Pr(>|t|)
(Intercept)  59.5226           NA     7.2929   8.162 0.000000000000000474 ***
Alter         2.2555       0.2970     0.1299  17.364 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 124.4 on 3117 degrees of freedom
Multiple R-squared:  0.0882,    Adjusted R-squared:  0.0879
F-statistic: 301.5 on 1 and 3117 DF,  p-value: < 0.00000000000000022
```

Mit der *multiplen linearen Regression* wollen wir nun prüfen, wie Alter, Bildung sowie Vertrauen ins Fernsehen als UV die TV-Nutzung erklären könnten. Dazu bringen wir Vertrauen in das Fernsehen als möglicherweise zusätzlichen erklärenden Faktor in unser Regressionsmodell

ein. Als Funktion können wir weiterhin `lm` nutzen; die zusätzlichen Variablen können wir in der Klammer ganz simpel ergänzen, indem wir sie mit einem + hinten anhängen:

```
model2 <- lm(TV_Konsum ~ Alter + Bildung + TV_Vertrauen, data = daten)
summary(lm.beta(model2))
```

```
Call:
lm(formula = TV_Konsum ~ Alter + Bildung + TV_Vertrauen, data = daten)

Residuals:
    Min      1Q  Median      3Q     Max
-253.27  -65.90  -19.06   33.03 1243.54

Coefficients:
             Estimate Standardized Std. Error t value            Pr(>|t|)
(Intercept)  173.13362           NA   12.70427  13.628 <0.0000000000000002 ***
Alter          1.56297      0.20579    0.13805  11.322 <0.0000000000000002 ***
Bildung      -26.27887     -0.23361    1.98888 -13.213 <0.0000000000000002 ***
TV_Vertrauen   5.25668      0.05248    1.71576   3.064               0.0022 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 120.9 on 3115 degrees of freedom
Multiple R-squared:  0.1383,    Adjusted R-squared:  0.1375
F-statistic: 166.7 on 3 and 3115 DF,  p-value: < 0.00000000000000022
```

2.1 Inhaltliche Interpretation des Outputs: Was bedeutet das Ergebnis?

Der Output zeigt uns: Das Alter hat weiterhin einen positiven Einfluss auf die tägliche Fernsehnutzung in Minuten. Je älter ein Nutzer ist, desto mehr nutzt er das Fernsehen. Mit jeder Einheit, in der die unabhängige Variable Alter steigt (hier: mit jedem Jahr Alter), nimmt die unabhängige Variable TV-Konsum um 2,08 Messeinheiten (hier: Minuten) zu. Dieser Zusammenhang ist mit p=.000 statistisch höchst signifikant.

Auch die zweite Variable, die wir ins Regressionsmodell eingebracht haben, das Vertrauen in das Fernsehen, hat einen signifikanten, positiven Effekt auf den TV-Konsum der Befragten: Mit jedem Skalenpunkt steigt der TV-Konsum um 5,1 Minuten an. Der Einfluss des TV-Vertrauens ist hoch signifikant.

Schließlich hat auch der Bildungsabschluss einen höchst signifikanten Einfluss auf die Intensität der Fernsehnutzung - dieser ist allerdings negativ, so dass der TV-Konsum mit steigender Messeinheit des Bildungsniveaus sinkt. Um das zu inhaltlich interpretieren, ist es sinnvoll, sich

die Skalierung der Variable genauer anzuschauen: Je höher der Wert, desto höher der Abschluss (z.B. Bildung: 1=ohne Abschluss; 3=Mittlere Reife; 5=Abitur). Wie unser Regressionsmodell zeigt, geht also ein höherer Abschluss mit einem niedrigeren TV-Konsum einher.

Im Output sehen wir nun auch, dass sich unser R2 (im Vergleich zur einfachen Regression oben) deutlich verbessert hat - es liegt jetzt bei 13,8 Prozent Varianzaufklärung. Das ist schon ordentlich, aber es muss immer noch weitere Faktoren geben, die die TV-Nutzung substantiell erklären können. Welche das sein können, können Sie ja mal selbst ausprobieren. Fügen Sie dazu einfach weitere - theoretisch plausible! - Kandidaten in das Regressionsmodell ein, indem sie sie durch ein "+"-Zeichen in ihre Modellfunktion integrieren (Achtung, vorher müssen Sie die zusätzlichen Variablen natürlich in ihrem Datenobjekt definieren und ggf. aufbereiten).

2.2 Zusätzliche Voraussetzungsprüfung bei der multiplen linearen Regression: Liegt Multikollinearität vor?

Die multiple lineare Regression erfordert *alle Voraussetzungen*, die für die einfache Regression auch verlangt sind - wie Sie die Voraussetzungen der Regressionaanalyse prüfen, haben Sie ja in den letzten Teilkapiteln gelernt - hier könnt ihr das noch einmal nachlesen.

Zusätzlich müsst ihr bei einer multiplen Regresssion allerdings noch prüfen, ob *Multikollinearität* vorliegt. Multikollinearität bedeutet, dass mindestens einer unserer Prädiktoren durch einen oder mehrere der anderen Prädiktoren vorhergesagt werden kann. Die Prädiktoren wären in diesem Fall *nicht unabhängig* voneinander, sondern würden hoch miteinander korrelieren und hätten damit sozusagen *keine selbstständige Erklärungskraft* im Modell.

Ob Multikollinearität vorliegt, können wir durch den *VIF-Wert (variance inflation factor)* ermitteln. Dieser darf nicht über 10 liegen, idealerweise auch nicht über 5. Um dies zu prüfen, nutzen wir den check_collinearity-Befehl aus dem Performance package:

```
check_collinearity(model2)
```

```
# Check for Multicollinearity

Low Correlation

         Term  VIF   VIF 95% CI Increased SE Tolerance Tolerance 95% CI
        Alter 1.19 [1.15, 1.25]         1.09      0.84     [0.80, 0.87]
      Bildung 1.13 [1.09, 1.18]         1.06      0.88     [0.85, 0.92]
 TV_Vertrauen 1.06 [1.03, 1.12]         1.03      0.94     [0.90, 0.97]
```

Die VIF-Werte liegen zwischen 0 und 5 (sogar alle bei 1); wir können daher davon ausgehen, dass *keine Multikollinearität* vorliegt (grün = "Low Correlation").

2.3 Vorhersage des multivariaten Modells

Auch für die Kombination verschiedener Merkmale bzw. unabhängiger Variablen können wir uns über die `predict.lm`-Funktion Prognosen erstellen lassen. So können wir beispielsweise vergleichen, wie sich der Fernsehkonsum bestimmter "Idealtypen" von Befragten unterscheiden würde:

```
predict.lm(model2, data.frame(Alter = c(25, 25), Bildung = c(0,5), TV_Vertrauen = c(7, 1)))
```

```
        1         2
249.00453  86.07011
```

Eine Person, die 25 Jahre alt ist und keinen Schulabschluss, aber dafür sehr großes Vertrauen in das Fernsehen hat, hat einen prognostizierten täglichen TV-Konsum von 249 Minuten. Eine Person, die ebenfalls 25 Jahre alt ist und Abitur hat, dem Fernsehen aber "überhaupt kein Vertrauen" entgegenbringt, hat einen täglichen prognostizierten Fernsehkonsum von 86 Minuten.

2.4 Multiple Regression mit Dummy-Codierung der kategorialen Variable "sex"

Eine Besonderheit schauen wir uns nun noch zum Schluss an: Die lineare Regression ist ein Verfahren für *metrische Daten*. Sie untersucht den Zusammenhang zwischen einer metrischen abhängigen Variable und mindestens einer metrischen unabhängigen Variable. In den bisherigen Regressionsanalysen haben wir mit metrischen und quasi-metrischen Variablen gearbeitet, die die Voraussetzung erfüllen, dass eine Regression gerechnet werden kann. Das ist auch der übliche Fall. Es gibt aber die "Ausnahme", dass auch kategorische (v.a. binäre) Variablen bei der Regressionsanalyse grundsätzlich eingesetzt werden können, wenn diese durch eine *Dummy-Coding* passend gemacht werden. Und diesen Fall schauen wir un nun am Beispiel der Variable "Geschlecht" an. Dazu müssen wir die Variable aber als *binär codiert* bzw. *dichotom* betrachten - d.h. wir behandeln sie so, als hätten wir hier nur 2 Ausprägungen.

2.4.1 Vorbereitung der Daten zum Zusammenhang von Alter, Bildung, Vertrauen ins TV, Geschlecht und TV-Konsum

Wir wollen das biologische Geschlecht (`sex`) als unabhängige Variable mit in unser Regressionsmodell aufnehmen. Bei "sex" haben wir aber das Problem, dass diese Variable nicht metrisch skaliert ist. Es handelt sich vielmehr um eine kategoriale Variable (mit 3 Ausprägungen). Dennoch können wir die Variable mit einem "Trick" in die Regressionsanalyse einbringen - Sie müssen diese dann aber durch Dummy-Coding passend machen. Hier wollen uns nun mal

anschauen, wie das funktioniert. Dazu müssen wir die Variable mittels mutate-Befehl aber erst einmal umcodieren und dadurch in eine dichotome VAriable verwandeln.

Durch die Dummy-Codierung wird die kategoriale Variable in zwei Gruppen übersetzt, von denen die eine mit 1 und die andere mit 0 codiert wird. Die Gruppe, der der Wert 0 zugeordnet wird, ist dann die Referenzkategorie. In unserem Beispiel Beispiel machen wir "männlich" zur Referenzkategorie (und codieren es mit 0 um). Der Regressionskoeffizient b gibt dann genau die Menge an, um die sich die Internetnutzung ändert, wenn sich das Geschlecht gegenüber der Referenzkategorie verändert.

PS: Die Variable als numerischen Wert zu behandeln, ist in unserem Fall etwas kompliziert, weil diese eine Faktorvariable war, die wir zuerst in eine Charaktervariable umwandeln mussten.

```
daten <- daten %>%
 filter(between(sex, 1, 2)) %>%
  mutate(sex_d = case_when(sex == 1 ~ 0, sex == 2 ~ 1))
```

```
häufigkeitstabelle <- table(daten$sex_d)
print(häufigkeitstabelle)
```

```
   0    1
1546 1569
```

Die Dummy-Codierung war erfolgreich - wir haben nun 1546 Befragte mit der Ausprägung 0 (männlich) sowie 1569 Befragte mit der Ausprägung 2 (weiblich).

2.5 Regressionsmodell zum Zusammenhang von Alter, Bildung, TV-Vertrauen, Geschlecht und TV-Konsum

In die bereits bekannte Regressionsfunktion `lm()` fügen wir im hinteren Teil (d.h. hinter der Tilde) nun einfach die weitere unabhängige Variable sex_d ein, indem wir sie mit einem + Zeichen anhängen.

```
model3 <- lm(TV_Konsum ~ Alter + Bildung + TV_Vertrauen + sex_d, data = daten)
summary(lm.beta(model3))
```

```
Call:
lm(formula = TV_Konsum ~ Alter + Bildung + TV_Vertrauen + sex_d,
    data = daten)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-255.89  -64.84  -18.97   33.01 1245.99

Coefficients:
              Estimate Standardized Std. Error t value            Pr(>|t|)
(Intercept)  170.54603           NA   12.96665  13.153 <0.0000000000000002 ***
Alter          1.56629      0.20625    0.13830  11.325 <0.0000000000000002 ***
Bildung      -26.32891     -0.23397    1.99103 -13.224 <0.0000000000000002 ***
TV_Vertrauen   5.34102      0.05330    1.71809   3.109               0.0019 **
sex_d          4.60020      0.01766    4.34277   1.059               0.2896
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 121 on 3110 degrees of freedom
Multiple R-squared:  0.1386,    Adjusted R-squared:  0.1375
F-statistic: 125.1 on 4 and 3110 DF,  p-value: < 0.00000000000000022
```

2.5.1 Inhaltliche Interpretation des Outputs: Was bedeutet das Ergebnis?

Gender hat *keinen* signifikanten Einfluss auf den TV-Konsum. Wenn es einen hätte, wäre der Einfluss geringer als bei Alter und Ausbildungsjahren (ersichtlich an der Größe des standardisierten beta-Koeffizienten). Wichtig: Gender ist als eine Dummy-Variable in 0=männlich und 1=weiblich codiert, deshalb ist der Estimate hier etwas schwieriger zu lesen. Die Frauen sind als 1 codiert und stellen hier die Vergleichsgruppe zur Referenzgruppe der Männer (=0) dar.

Frauen haben eine um (4,6) Minuten höheren TV-Konsum als Männer (wobei dieser Befund statistisch ja nicht signifikant ist).