Lösungen zu den Aufgaben

1. **Aufgabe**

   Die Inferenzstatistik ist ein Sammlung an Verfahren zur Bemessung von Unsicherheit in statistischen Schlüssen.

   a. Für welche Statistiken - also Kennzahlen der Deskriptivstatistik wie etwa $\overline{X}, sd, r$ - kann man die Inferenzstatistik verwenden?

   b. Für welche Forschungsfragen oder -bereiche kann man die Inferenzstatistik verwenden?

   c. Gibt es besondere Fälle, in denen man nicht die Inferenzstatistik verwenden möchte? Wenn ja, welche?

   Lösung

   a. Für alle: Für jede Statistik kann man prinzipiell von der jeweiligen Stichprobe (auf Basis derer die Statistik berechnet wurde) auf eine zugehörige Grundgesamtheit schließen.

   b. Für alle: Die Methoden der Inferenzstatistik sind prinzipiell unabhängig von den Spezifika bestimmter Forschungsfragen oder -bereiche. In den meisten Forschungsfragen ist man daran interessiert *allgemeingültige* Aussagen zu treffen. Da Statistiken sich nur auf eine Stichprobe - also einen zumeist nur kleinen Teil einer Grundgesamtheit beziehen - wird man sich kaum mit einer Statistik zufrieden geben, sondern nach Inferenzstatistik verlangen.

   c. In einigen Ausnahmefällen wird man auf eine Inferenzstatistik verzichten. Etwa wenn man bereits eine Vollerhebung durchgeführt hat, z.B. alle Mitarbeiteris eines Unternehmens befragt hat, dann kennt man ja bereits den wahren Populationswert. Ein anderer Fall ist, wenn man nicht an Verallgemeinerungen interessiert ist: Kennt man etwa die Überlebenschance p des Titanic-Unglücks, so ist es fraglich auf welche Grundgesamtheit man die Statistik p bzw. zu welchem Paramter π man generalisieren möchte.

2. **Aufgabe**

   Für Statistiken (Stichprobe) verwendet man meist lateinische Buchstaben; für Parameter (Population) verwendet man entsprechend meist griechische Buchstaben.

   Vervollständigen Sie folgende Tabelle entsprechend!

Kennwert	**Statistik**	**Parameter**
Mittelwert	$\overline{X}$	NA
Mittelwertsdifferenz	$\overline{X}_1 - \overline{X}_2$	NA
Streuung	sd	NA
Anteil	p	NA
Korrelation	r	NA
Regressionsgewicht	b	NA

   Lösung

Kennwert	**Statistik**	**Parameter**
Mittelwert	$\overline{X}$	μ
Mittelwertsdifferenz	$\overline{X}_1 - \overline{X}_2$	$\mu_1 - \mu_2$
Streuung	sd	σ
Anteil	p	π
Korrelation	r	ρ
Regressionsgewicht	b	β

3. **Aufgabe**

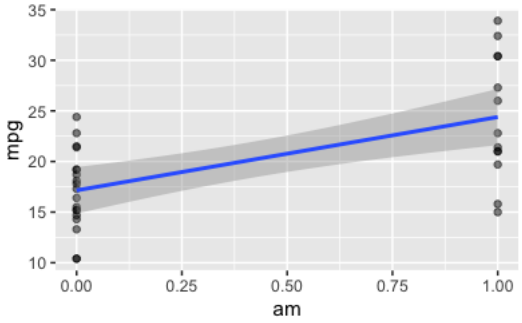
   Der *t-Test* kann als Spezialfall der Regressionsanalyse gedeutet werden.

   Hierbei ist es wichtig, sich das Skalenniveau der Variablen, die ein t-Test verarbeitet, vor Augen zu führen.

a. Benennen Sie die Skalenniveaus der UV eines t-Tests! Geben Sie nur ein Wort ein. Verwenden Sie nur Kleinbuchstaben (z.B. `regression`).
b. Benennen Sie die Skalenniveaus der AV eines t-Tests! Geben Sie nur ein Wort ein. Verwenden Sie nur Kleinbuchstaben (z.B. `regression`).
c. Nennen Sie eine beispielhafte Forschungsfrage für einen t-Test.
d. Skizzieren Sie ein Diagramm einer Regression, die analytisch identisch (oder sehr ähnlich) zu einem t-Test ist!

Lösung

a. UV: binär
b. AV: metrisch
c. Unterscheiden sich die mittleren Einparkzeiten von Frauen und Männern?
d. Aus dem Datensatz `mtcars`: `text data(mtcars) mtcars %>% ggplot() + aes(x = am, y = mpg) + geom_point(alpha = .5) +`



`geom_smooth(method = "lm")`

4. **Aufgabe**

Die *Varianzanalyse* (Analysis of Variance; Anova) ist ein statistisches Verfahren, um die Gleichheit zweier oder mehr Populationsmittelwerte zu testen: $\mu_1 = \mu_2 = \ldots = \mu_n$.

Wie viele andere Verfahren kann die Anova als ein Spezialfall der Regression bzw. des linearen Modells $y = \beta_0 + \beta_1 + \ldots \beta_n + \epsilon$ betrachtet werden.

Als ein spezielles Beispiel betrachten wir die Frage, ob Diamanten (Datensatz `diamonds`) verschiedener Schliffart (`cut`) sich nicht in ihrem mittleren Preis (`price`) unterscheiden.

Den Datensatz können Sie so laden:

```
library(tidyverse)
data(diamonds)
```

a. Nennen Sie UV und AV! Geben Sie jeweils das Skalenniveau an!

b. Nennen Sie die Regressionsformel für diese Forschungsfrage!

c. Betrachten Sie die Ausgabe von R:

```
Estimates:
              mean   sd     10%    50%    90%
(Intercept) 4062.0   25.9 4029.1 4062.6 4094.8
cut.L       -363.7   67.3 -449.8 -363.8 -278.3
cut.Q       -223.7   59.7 -300.2 -223.2 -147.3
cut.C       -700.8   51.7 -766.4 -701.7 -634.4
cut^4       -280.2   41.7 -333.5 -280.1 -226.5
sigma       3963.9   12.1 3948.4 3963.7 3979.5
```

Geben Sie die Punktschätzer β zu den Mittelwertsunterschieden an. Die Spalte `sd` quantifiziert die Unsicherheit bzw. Ungenauigkeit in der Schätzung. Die Prozentwerte kann man interpretieren, dass das Modell der Meinung ist, der wahre (zu schätzende Werte) ist mit 10% (50%, 90%) Wahrscheinlichkeit kleiner als der jeweils angegebene Wert.

Vor diesem Hintergrund: Würden Sie die Hypothese der Gleichheit aller Mittelwerte der Gruppen (an Schliffarten) ablehnen oder beibehalten?

Lösung

a. UV: `cut`, AV: `price`; `cut` ist nominal (mehrstufig) und `price` ist metrisch (verhältnisskaliert)

b. `price ~ cut`

c. Die Punktschätzer zu den Mittelwertsunterschieden der einzelnen Gruppenwerten - relativ zur Bezugsgruppe (Baseline) - sind in der Spalte `mean` aufgeführt.

Mit hoher Wahrscheinlichkeit ist die Hypothese der Gleichheit aller Gruppenmittelwerte abzulehnen, laut diesem Modell: Ein Großteil der Wahrscheinlichkeitsmasse ist für jeden Modellparameter außerhalb 0.

Das Modell wurde übrigens so berechnet:

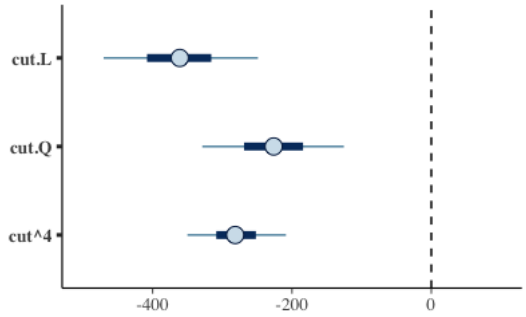
```
library(rstanarm)
stanlm1 <- stan_glm(price ~ cut,
```

```
                  data = diamonds,
                  refresh = FALSE)  # für weniger R-Output
summary(stanlm1)

##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      price ~ cut
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 53940
##  predictors:   5
##
## Estimates:
##               mean   sd     10%    50%    90%
## (Intercept) 4061.6   25.2 4029.8 4061.7 4092.6
## cut.L       -361.5   67.7 -447.0 -361.1 -274.7
## cut.Q       -226.8   61.7 -307.0 -226.3 -148.0
## cut.C       -699.4   53.6 -768.4 -699.6 -632.3
## cut^4       -280.3   42.3 -333.5 -281.6 -224.9
## sigma       3963.7   12.0 3948.4 3963.4 3979.2
##
## Fit Diagnostics:
##            mean   sd     10%    50%    90%
## mean_PPD 3932.5   24.2 3901.1 3933.0 3963.4
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for details see help('summary.stanreg'))
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.5  1.0  2495
## cut.L         1.3  1.0  2566
## cut.Q         1.3  1.0  2326
## cut.C         1.0  1.0  3041
## cut^4         0.7  1.0  3393
## sigma         0.2  1.0  3342
## mean_PPD      0.5  1.0  2641
## log-posterior 0.0  1.0  1715
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample size, and Rhat is the potential s

library(bayesplot)

plot(stanlm1,
     pars = c("cut.L", "cut.Q", "cut.L", "cut^4")) +
  scale_x_continuous(limits = c(-500, 100)) +
  geom_vline(xintercept = 0, linetype = "dashed")
```



Ein “klassisches” Regressionsmodell kommt übrigens zu ähnlichen Werten:

```
lm1 <- lm(price ~ cut, data = diamonds)
summary(lm1)

##
## Call:
## lm(formula = price ~ cut, data = diamonds)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4258  -2741  -1494   1360  15348
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4062.2       25.4  159.92  < 2e-16 ***
## cut.L         -362.7       68.0   -5.33  9.8e-08 ***
## cut.Q         -225.6       60.6   -3.72    2e-04 ***
## cut.C         -699.5       52.8  -13.25  < 2e-16 ***
## cut^4         -280.4       42.6   -6.59  4.5e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3960 on 53935 degrees of freedom
## Multiple R-squared:  0.0129, Adjusted R-squared:  0.0128
## F-statistic:  176 on 4 and 53935 DF,  p-value: <2e-16
```

Die einzelnen Stufen (*levels*) von `cut` lassen sich z.B. so ermitteln:

```
diamonds %>%
  summarise(stufen_von_cut = levels(cut))
```

```
## # A tibble: 5 × 1
##   stufen_von_cut
##   <chr>
## 1 Fair
## 2 Good
## 3 Very Good
## 4 Premium
## 5 Ideal
```

Alternativ:

```
diamonds %>%
  count(cut)
```

```
## # A tibble: 5 × 2
##   cut           n
##   <ord>     <int>
## 1 Fair       1610
## 2 Good       4906
## 3 Very Good 12082
## 4 Premium   13791
## 5 Ideal     21551
```

5. **Aufgabe**

Die Korrelation prüft, ob zwei Merkmale linear zusammenhängen.

Wie viele andere Verfahren kann die Korrelation als ein Spezialfall der Regression bzw. des linearen Modells $y = \beta_0 + \beta_1 + ...\beta_n + \epsilon$ betrachtet werden.

Als ein spezielles Beispiel betrachten wir die Frage, ob das Gewicht eines Diamanten (`carat`) mit dem Preis (`price`) zusammenhängt (Datensatz `diamonds`).

Den Datensatz können Sie so laden:

```
library(tidyverse)
data(diamonds)
```

a. Geben Sie das Skalenniveau beider Variablen an!

b. Betrachten Sie die Ausgabe von R:

```
lm1 <- lm(price ~ carat, data = diamonds)
summary(lm1)
```

```
##
## Call:
## lm(formula = price ~ carat, data = diamonds)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18585   -805    -19    537  12732
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2256.4       13.1    -173   <2e-16 ***
## carat         7756.4       14.1     551   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1550 on 53938 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.849
## F-statistic: 3.04e+05 on 1 and 53938 DF,  p-value: <2e-16
```

Wie (bzw. wo) ist aus dieser Ausgabe die Korrelation herauszulesen?

c. Macht es einen Unterschied, ob man Preis mit Karat bzw. Karat mit Preis korreliert?

d. In der klassischen Inferenzstatistik ist der p-Wert eine zentrale Größe; ist er klein ($p < .05$) so nennt man die zugehörige Statistik *signifikant* und verwirft die getestete Hypothese.

e. Im Folgenden sehen Sie einen Korrelationstest auf statistische Signifikanz, mit R durchgeführt. Zeigt der Test ein (statistisch) signifikantes Ergebnis? Wie groß ist der “Unsicherheitskorridor”, um den Korrelationswert (zugleich Punktschätzer für den Populationswert)?

```
library(rstatix)
diamonds %>%
  sample_n(30) %>%
  select(price, carat) %>%
  rstatix::cor_test() %>%
  gt()
```

```
## Error in gt(.): could not find function "gt"
```

Lösung

a. `carat` ist metrisch (verhältnisskaliert) und `price` ist metrisch (verhältnisskaliert)

b. R^2 kann bei einer einfachen (univariaten) Regression als das Quadrat von r berechnet werden. Daher $r = \sqrt{R^2}$.

```
sqrt(0.8493)
```

```
## [1] 0.92
```

Zum Vergleich

```
diamonds %>%
  summarise(r = cor(price, carat))
```

```
## # A tibble: 1 × 1
##       r
##   <dbl>
## 1 0.922
```

Man kann den Wert der Korrelation auch noch anderweitig berechnen (β umrechnen in ρ).

c. Nein. Die Korrelation ist eine symmetrische Relation.

d. Ja; die Zahl “3.81e-14” bezeichnet eine positive Zahl kleiner eins mit 13 Nullern vor der ersten Ziffer, die nicht Null ist (3.81 in diesem Fall). Der “Unsicherheitskorridor” reicht von etwa 0.87 bis 0.97.

6. **Aufgabe**

Eine statistische Analyse, wie eine Regression, ist mit mehreren Arten an Ungewissheit konfrontiert. Zum einen gibt es die *Ungewissheit in den Modellparametern*. Für die Regression bedeutet das: “Liegt die Regressionsgerade in”Wahrheit” (in der Population) genauso wie in der Stichprobe, sind Achsenabschnitt und Steigung in der Stichprobe also identisch zur Popuation?“. Zum anderen die *Ungewissheit innerhalb des Modells*. Auch wenn wir die”wahre” Regressionsgleichung kennen würden, wären (in aller Regel) die Vorhersagen trotzdem nicht perfekt. Auch wenn wir etwa wüssten, wieviel Klausurpunkte “in Wahrheit” pro Stunde Lernen heraussprингen (und wenn wir den wahren Achsenabschnitt kennen würden), so würde das Modell trotzdem keine perfekten Vorhersagen zum Klausurerfolg liefern. Vermutlich fehlen dem Modell wichtige Informationen etwa zur Motivation der Studentis.

Vor diesem Hintergrund, betrachten Sie folgendes statistisches Modell:

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      mpg ~ hp
##  observations: 32
##  predictors:   2
## ------
##             Median MAD_SD
## (Intercept) 30.1    1.6
## hp          -0.1    0.0
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 3.9    0.5
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

a. Welche Zahl kennzeichnet die Ungewissheit des Modells zum Achsenabschnitt?
b. Welche Zahl kennzeichnet die Ungewissheit des Modells zum Regressionsgewicht?
c. Welche Zahl(en) kennzeichnet/kennzeichnen die Ungewissheit des Modells gegeben der Modellparameter (die Ungewissheit innerhalb des Modells)?

Lösung

a. 1.7
b. 0.0
c. 3.9 und auch dazu 0.5

7. **Aufgabe**

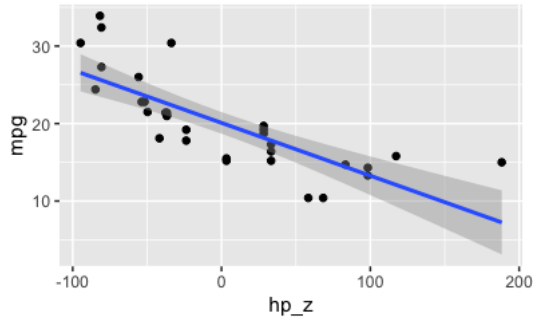
Betrachten Sie folgendes Modell, das den Zusammenhang von PS-Zahl und Spritverbrauch untersucht (Datensatz `mtcars`).

Aber zuerst zentrieren wir den metrischen Prädiktor `hp`, um den Achsenabschnitt besser interpretieren zu können.

```
Estimates:
              mean   sd   10%   50%   90%
(Intercept) 20.1    0.7 19.2  20.1  21.0
hp_z        -0.1    0.0 -0.1  -0.1  -0.1
sigma        4.0    0.5  3.4   3.9   4.7
```

Jetzt können wir aus dem Achsenabschnitt (Intercept) herauslesen, dass ein Auto mit `hp_z = 0` - also mit mittlerer PS-Zahl - vielleicht gut 20 Meilen weit mit einer Gallone Sprit kommt.

Zur Verdeutlichung ein Diagramm zum Modell:



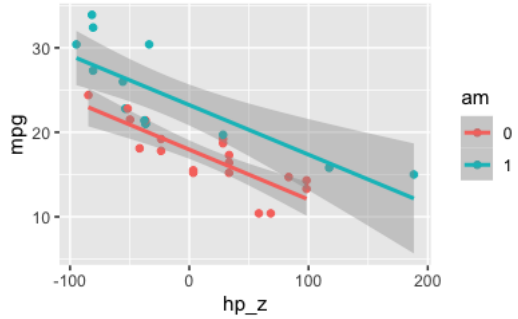
Adjustieren Sie im Modell die PS-Zahl um die Art des Schaltgetriebes (`am`), so dass das neue Modell den Effekt der PS-Zahl bereinigt bzw. unabhängig von der Art des Schaltgetriebes widerspiegelt!

(Hinweis `am=0` ist ein Auto mit Automatikgetriebe.)

Lösung

```
Estimates:
             mean   sd   10%   50%   90%
(Intercept) 26.6    1.5 24.7  26.6  28.5
hp          -0.1    0.0 -0.1  -0.1   0.0
am           5.3    1.1  3.8   5.3   6.6
sigma        3.0    0.4  2.5   3.0   3.5
```

Die Koeffizienten zeigen, dass der Achsenabschnitt für Autos mit Automatikgetriebe um etwa 5 Meilen geringer ist als für Autos mit manueller Schaltung: Ein durchschnittliches Auto mit manueller Schaltung kommt also etwa 5 Meilen weiter als ein Auto mit Automatikschaltung, glaubt unser Modell.



Man könnte hier noch einen Interaktionseffekt ergänzen.

8. **Aufgabe**

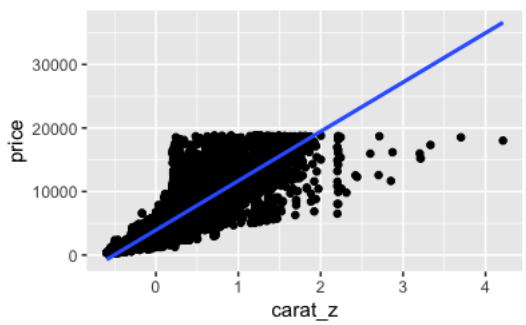
Betrachten Sie folgendes Modell, das den Zusammenhang des Preises (`price`) und dem Gewicht (`carat`) von Diamanten untersucht (Datensatz `diamonds`).

Aber zuerst zentrieren wir den metrischen Prädiktor `carat`, um den Achsenabschnitt besser interpretieren zu können.

Dann berechnen wir ein (bayesianisches) Regressionsmodell, wobei wir auf die Standardwerte der Prior zurückgreifen.

```
Estimates:
             mean   sd     10%    50%    90%
(Intercept) 3932.5    6.8 3923.7 3932.5 3941.1
carat_z     7756.3   14.2 7737.8 7756.2 7774.7
sigma       1548.6    4.8 1542.5 1548.6 1554.7
```

Zur Verdeutlichung ein Diagramm zum Modell:



a.

a. Was kostet in Diamant mittlerer Größe laut Modell `lm1`? Runden Sie auf eine Dezimale. Geben Sie nur eine Zahl ein. b) Geben Sie eine Regressionsformel an, die `lm1` ergänzt, so dass die Schliffart (`cut`) des Diamanten kontrolliert (adjustiert) wird. Anders gesagt: Das Modell soll die mittleren Preise für jede der fünf Schliffarten angeben. Geben Sie nur die Regressionsformel an. Lassen Sie zwischen Termen jeweils ein Leerzeichen Abstand. *Hinweis*: Es gibt (laut Datensatz) folgende Schliffarten (und zwar in der folgenden Reihenfolge): `## # A tibble: 5 × 1 ## cut ## <ord> ## 1 Ideal ## 2 Premium ## 3 Good ## 4 Very Good ## 5 Fair ## [1] "Fair" "Good" "Very Good" "Premium" "Ideal"`

b.

Lösung

a. Unser Modell `lm1` schätzt den Preis eines Diamanten mittlerer Größe auf etwa `3932.5` (was immer auch die Einheiten sind, Dollar vermutlich).

b. `price ~ carat_z + cut`

Das Modell könnten wir so berechnen:

Oder auch so, mit der klassischen Regression:

```
Estimates:
              mean   sd     10%    50%    90%
(Intercept) 3579.4    9.7 3566.9 3579.5 3591.9
carat_z     7871.5   14.2 7853.1 7871.4 7890.3
cut.L       1239.4   26.3 1205.7 1239.6 1272.6
cut.Q       -527.9   23.4 -557.9 -528.3 -497.7
cut.C        367.7   20.4  341.8  367.7  393.5
cut^4         74.9   16.5   53.6   75.0   95.5
sigma       1511.5    4.6 1505.6 1511.5 1517.4
```

```
##
## Call:
## lm(formula = price ~ carat_z + cut, data = diamonds)
##
## Coefficients:
## (Intercept)      carat_z        cut.L        cut.Q        cut.C        cut^4
##      3579.3       7871.1       1239.8       -528.6        367.9         74.6
```

Man könnte hier noch einen Interaktionseffekt ergänzen.

a.
b.

9. **Aufgabe**

Zwei Modelle, `m1` und `m2` produzieren jeweils die gleiche Vorhersage (den gleichen Punktschätzer).

`m1`:

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2438 -0.0659  0.0107  0.0595  0.2187
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.00187    0.00934    -0.2     0.84
## x            0.99795    0.00996   100.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.093 on 98 degrees of freedom
## Multiple R-squared:  0.99,   Adjusted R-squared:  0.99
## F-statistic: 1e+04 on 1 and 98 DF,  p-value: <2e-16
```

`m2`:

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1370 -0.5838 -0.0009  0.7129  2.5472
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.130      0.105    1.24     0.22
## x              1.058      0.104   10.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.1 on 98 degrees of freedom
## Multiple R-squared:  0.512,  Adjusted R-squared:  0.507
## F-statistic:  103 on 1 and 98 DF,  p-value: <2e-16
```

Die Modelle unterscheiden sich aber in ihrer Ungewissheit bezüglich β, wie in der Spalte `Std. Error` ausgedrückt.

Welches der beiden Modelle ist zu bevorzugen? Begründen Sie.

Lösung

Modell `m1` hat eine *kleinere* Ungewissheit im Hinblick auf die Modellkoeffizienten β_0, β_1 und ist daher gegenüber `m2` zu bevorzugen.

10. **Aufgabe**

Nennen Sie ein Beispiel für eine Vorhersagemodell (mit lineare Regression), wo Sie sich nicht mit dem Punktschätzer für die Modellkoeffizienten begnügen, sondern auch über die Ungewissheit in der Schätzung der Modellkoeffizienten informiert werden möchten!

Lösung

- Nehmen wir an, ein Modell sagt vorher, dass ich bei einer Investition Gewinn machen werden, in erwarteter (mittlerer) Höhe von 100 Euro. Es würde mich aber interessieren, wie sicher sich das Modell ist. Vor allem wäre ich brennend daran interessiert, ob das Modell auch "negative Gewinne" für plausibel hält. Sprich: Ich möchte die Ungewissheit in der Vorhersage gerne genauer wissen, am liebsten mit einer präzisen (und zuverlässigen) Zahl.
- Sagt der Wetterbericht vorher, dass morgen Mittag 15 Grad zu erwarten sind, wäre ich an der zusätzlichen Information interessiert, dass das Modell des Wetterfroschs auch (sagen wir) Werte von -5 bis +35 für plausibel hält.
- Angenommen, der Weltklimarat prognostiziert einen mittleren Temperaturanstieg von 2.5 Grad, so ist es wichtig zu wissen, dass das Modell der Forscheris auch Werte im Bereich von +0.5 bis +4.5 Grad für plausibel hält.

11. **Aufgabe**

Denken wir uns ein kausales System mit einer Ursache und einer Wirkung, etwa der Einfluss der Naturbelassenheit (N) eines Landkreises auf die Anzahl der Störche (S) dort (ein positiver Einfluss). Nehmen wir weiter an, die Naturbelassenenheit eines Landkreises hat einen (positiven) Einfluss auf die Anzahl Neugeborener (Babies, B).

Weitere kausale Einflüsse existieren in diesem kausalen System nicht (es handelt sich ja hier umn ein Gedankenexperiment, wir können frei bestimmen!).

Die Frage ist nun, ob wir erwarten müssen, dass Störche und Babies zusammenhängen in diesem System, dass es also dort, wo es viele Störche gibt auch viele Babies gibt. Das wäre deswegen beachtlich, weil wir in unserem System explizit keinen (kausalen) Zusammenhang zwischen diesen beiden Größen definiert haben.

Um die Sache etwas greifbarer zu machen, erstellen wir uns Daten, die zu diesem System passen. Sagen wir, wir haben 100 Landkreise, die in der Zahl der Störche und Babies und Naturbelassenheit variieren. Der Einfachheit halber seien alle Werte in z-Werten ausgedrückt. Gehen wir weiter (der Einfachheit halber) davon aus, alle Größen sind normalverteilt. Solche Werte kann man mit der R-Funktion `rnorm()` erzeugen.

Schließlich gehen wir noch davon aus, dass die Einflüsse linear sind und nicht perfekt. Der Zufall (zufälliger "Fehler", e) soll also auch einen Einfluss auf die Größen haben.

```
N <- rnorm(100, mean = 0, sd = 1)  # 100 normalverteilte z-Werte
e1 <- rnorm(100)  # das gleiche wie oben: normalverteilte z-Werte
e2 <- rnorm(100)  # das gleiche wie oben: normalverteilte z-Werte
S <- N + e1  # S wird determiniert durch N und e
B <- N + e2  # B wird determiniert druch N und e
```

Testen wir unsere simulierten Daten mit einer einfachen Regression, der Frage, ob die Anzahl der Störche (S) von der Natürlichkeit (N) abhängt:

```
lm1 <- lm(S ~ N)
summary(lm1)
```

```
##
## Call:
## lm(formula = S ~ N)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2528 -0.6076  0.0072  0.6931  2.0559
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0641     0.0976    0.66     0.51
## N             1.0894     0.1044   10.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.97 on 98 degrees of freedom
## Multiple R-squared:  0.526,  Adjusted R-squared:  0.521
## F-statistic:  109 on 1 and 98 DF,  p-value: <2e-16
```

Unser Modell `lm1` bringt unsere Annahmen deutlich zum Vorschein.

a. Bestimmen Sie den Zusammenhang (β oder ρ) zwischen Störchen und Babies!
b. Erklären Sie den Befund!

Lösung

a. Es findet sich ein nicht-kausaler, also ein *Scheinzusammenhang* zwischen Störchen und Babies:

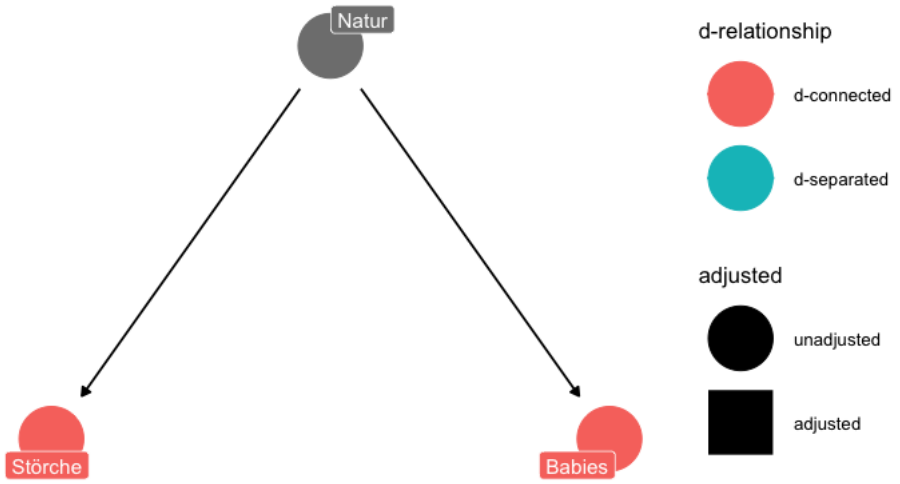
```
lm(B ~ S)
```

```
##
## Call:
## lm(formula = B ~ S)
##
## Coefficients:
## (Intercept)            S
##     -0.0698       0.5048
```

```
cor(S, B)
```

```
## [1] 0.52
```

b. Haben zwei Variablen eine gemeinsame Ursache, so sind sie durch eine Scheinkorrelation verbunden.



`dconnected` bedeutet, dass zwei Variablen *verbunden* (connected) sind, sie also voneinander (statistisch) abhängig (assoziiert) sind, z.B. korreliert. Das `d` steht für `directed` also über gerichtete Kanten, die Kausalpfeile, verbunden. Dabei ist zu beachten, dass die Assoziation in beide Richtungen des Kausalpfeils "fließen" kann; auch gegen "den Strom" (also von der Pfeilstütze anfangend rückwärts).

12. **Aufgabe**

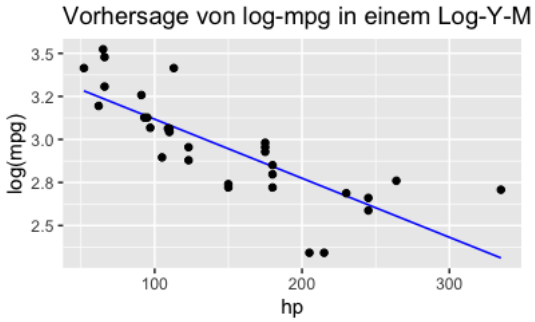
Wir suchen ein Modell, das einen *nichtlinearen* Zusammenhang von PS-Zahl und Spritverbrauch darstellt (Datensatz `mtcars`).

Geben Sie dafür ein mögliches Modell an! Nutzen Sie den R-Befehl `lm`.

Lösung

```
##
## Call:
## lm(formula = mpg_log ~ hp, data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.4158 -0.0658 -0.0174  0.0983  0.3962
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.460467   0.078584   44.04  < 2e-16 ***
## hp          -0.003429   0.000487   -7.05  7.9e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.19 on 30 degrees of freedom
## Multiple R-squared:  0.623,  Adjusted R-squared:  0.611
## F-statistic: 49.6 on 1 and 30 DF,  p-value: 7.85e-08
```

Visualisieren wir die Vorhersagen des Modells:



Möchte man auf der Y-Achse *mpg* und nicht log(mpg) anzeigen, muss man den Logarithmus wieder "auflösen", das erreicht man mit der Umkehrfunktion des Logarithmus, das Exponentieren (man "delogarithmiert"):

$$\begin{aligned} log\,(y) &= x && \text{|Y in Log-Form} \\ exp\,(log\,(y)) &= exp\,(x) && \text{|Jetzt exponenzieren wir beide Seiten} \\ y &= exp\,(x) \end{aligned}$$

Dabei gilt $exp\,(x) = e^x$, mit e als Eulersche Zahl (2.71…).