Software Übungen

Elliot Beck

Aufgabe 1

Aufgabe 13.3.6 im Buch von Ross (vom letzten Semester). Falls Sie das Buch nicht haben, hier ist eine "Kopie" der Aufgabe im Originalton:

The following data categorize a random selection of professors of a certain university according to their teaching performance (as measured by the students in their classes) in the most recent semester and the number of courses they were teaching at the time.

```
# Erstellen des Datensatzes
ross_13_3_6 <- matrix(c(12, 10, 4, 32, 40, 38, 7, 12, 25),
    nrow = 3, byrow = TRUE,
    dimnames = list(c("Above", "Average", "Below"), c("1", "2", "3+"))
)
ross_13_3_6
```

```
##          1  2 3+
## Above   12 10  4
## Average 32 40 38
## Below    7 12 25
```

Test, at the 5 percent level, the hypothesis that a professor's teaching performance is independent of the number of courses she or he is teaching.

```
# Durchführung des Chi-Quadrat-Tests
chisq_test <- chisq.test(ross_13_3_6)
chisq_test
```

```
##
##  Pearson's Chi-squared test
##
## data:  ross_13_3_6
## X-squared = 14.312, df = 4, p-value = 0.006363
```

Da der p-Wert kleiner ist als 1%, können wir die Null-Hypothese verwerfen, dass die Evaluation unabhängig von der Anzahl der Kurse ist.

```
round(chisq.test(ross_13_3_6)$expected, 1)
```

```
##            1    2   3+
## Above    7.4  9.0  9.7
## Average 31.2 37.9 40.9
## Below   12.5 15.2 16.4
```

Ein Vergleich der beobachteten Tabelle mit der erwarteten Tabelle zeigt, dass Professoren mit einem Kurs besser als erwartet (unter der Null) abschneiden während Professoren mit drei oder mehr Kursen schlechter als erwartet (unter der Null) abschneiden. Damit gibt es einen negativen Zusammenhang zwischen Evaluation und Anzahl der Kurse: wer mehr Kurse unterrichtet, wird im Durchschnit schlechter evaluiert.

Aufgabe 2

Der Datensatz "Medikament" enthält die Blutdrücke von n = 80 Patienten vor und nach der Behandlung mit einem Mittel, das hohen Blutdruck senken soll.

```
# Einlesen der Daten für Aufgabe 2
medikament <- read.csv("data/medikament.csv")
head(medikament)
```

```
##   before after
## 1  168.5 162.7
## 2  172.2 160.5
## 3  175.2 168.5
## 4  173.2 155.2
## 5  164.8 155.0
## 6  167.1 153.7
```

a) Finden Sie ein 95% Konfidenz-Intervall für die durchschnittliche Blutdruck-Senkung. Erscheint das Medikament somit wirksam?

```
# Berechnung des Konfidenz-Intervalls
senkung <- medikament$before - medikament$after
ci <- t.test(senkung)$conf.int
ci
```

```
## [1] 10.21158 13.29842
## attr(,"conf.level")
## [1] 0.95
```

Da dieses Konfidenz-Intervall die Null nicht enthält, ist es plausibel, dass das Medikament im Durchschnitt den Blutdruck senkt.

b) Berechnen Sie die Stichproben-Korrelation zwischen "before" und "after".

```
# Berechnung der Korrelation
cor(medikament$before, medikament$after)
```

```
## [1] 0.7778159
```

c) Wiederholen Sie anhand dieser Daten das Rechenbeispiel von Seite 28 der Folien 1

```
# Direkte Berechnung von s_D
sd(senkung)
```

```
## [1] 6.935506
```

```
# Indirekte Berechnung von s_D
sqrt(var(medikament$before) + var(medikament$after) - 2 *
    cor(medikament$before, medikament$after) * sd(medikament$before) * sd(medikament$after))
```

```
## [1] 6.935506
```

d) Erscheint das Medikament wirksamer als das Placebo?

Der Trick ist hier wie folgt. Zuerst berechnen wir die paarweise Differenzen, einmal für das Medikament ($n_1 = 80$) und dann noch einmal für das Placebo ($n_2 = 60$). Dann wenden wir das Konfidenz-Intervall für zwei unabhängige Stichproben auf die zwei Differenz-Stichproben an

```r
# Einlesen der Placebodaten
placebo <- read.csv("data/placebo.csv")

# Berechnung der Konfidenz-Intervalle für die Differenzen
senkung_placebo <- placebo$before - placebo$after
t.test(senkung, senkung_placebo)$conf.int
```

```
## [1] -1.703946  3.460613
## attr(,"conf.level")
## [1] 0.95
```

Da dieses Konfidenz-Intervall die Null enthält, ist es plausibel, dass das Medikament nicht besser ist als das Placebo.

Aufgabe 3

Der Datensatz enthält Aktienüberschussrenditen in Prozent für drei Aktien (Coca Cola, General Electrics und Sun Microsystems) sowie für den S&P 500 Aktien-Index von 06/1992 bis 04/2002.

```r
# Einlesen der Daten für Aufgabe 3
aktien <- read.csv("data/aktien.csv")
head(aktien)
```

```
##   coke.ex general.ex sun.ex sp500.ex
## 1   -9.40       1.49  -6.59    -2.05
## 2    4.42      -1.88   2.12     3.67
## 3    2.43      -3.53   0.21    -2.66
## 4   -6.05       5.50  16.05     0.67
## 5    0.07      -2.16  12.56    -0.03
## 6   -2.41       8.21 -10.19     2.77
```

a) Schätze die Parameter α und β des CAPM.

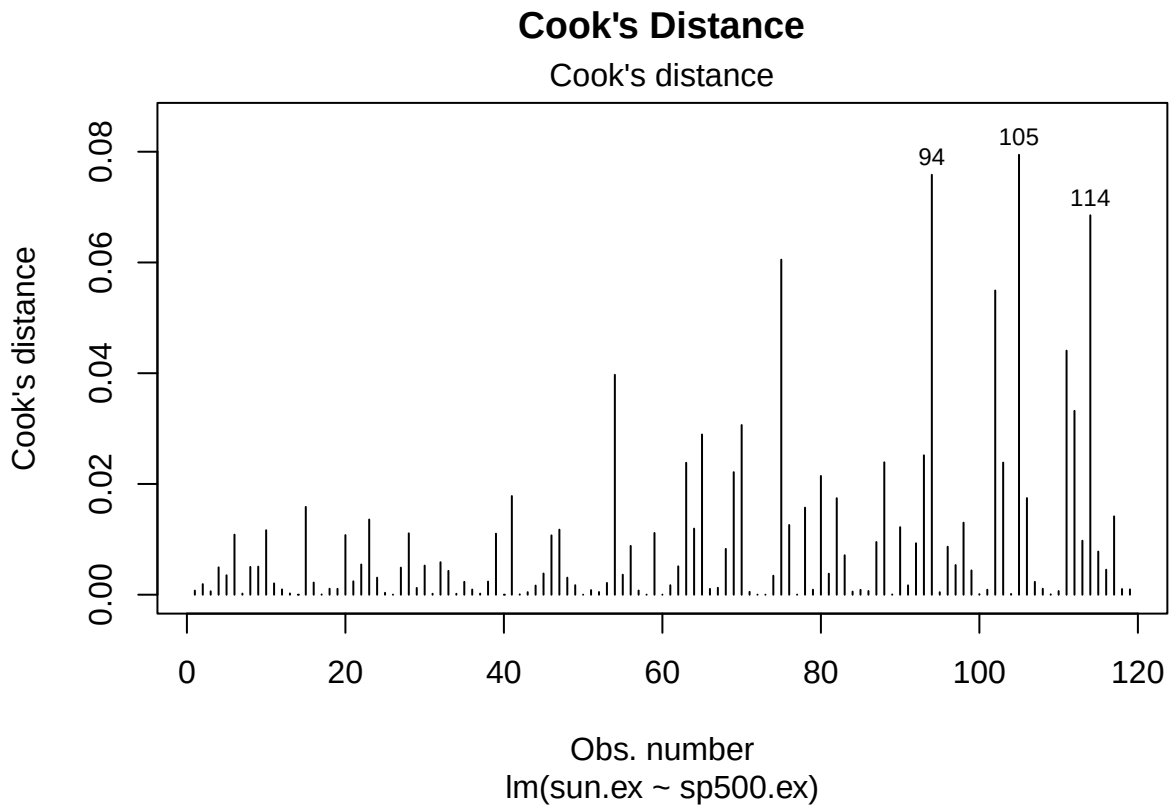
```r
# CAPM-Modell
fit <- lm(sun.ex ~ sp500.ex, data = aktien)
summary(fit)
```

```
##
## Call:
## lm(formula = sun.ex ~ sp500.ex, data = aktien)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.663  -8.056   0.973   8.251  24.558
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6060     1.1327   1.418    0.159
## sp500.ex      1.8990     0.2752   6.900 2.84e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.26 on 117 degrees of freedom
## Multiple R-squared:  0.2892, Adjusted R-squared:  0.2832
## F-statistic: 47.61 on 1 and 117 DF,  p-value: 2.84e-10
```

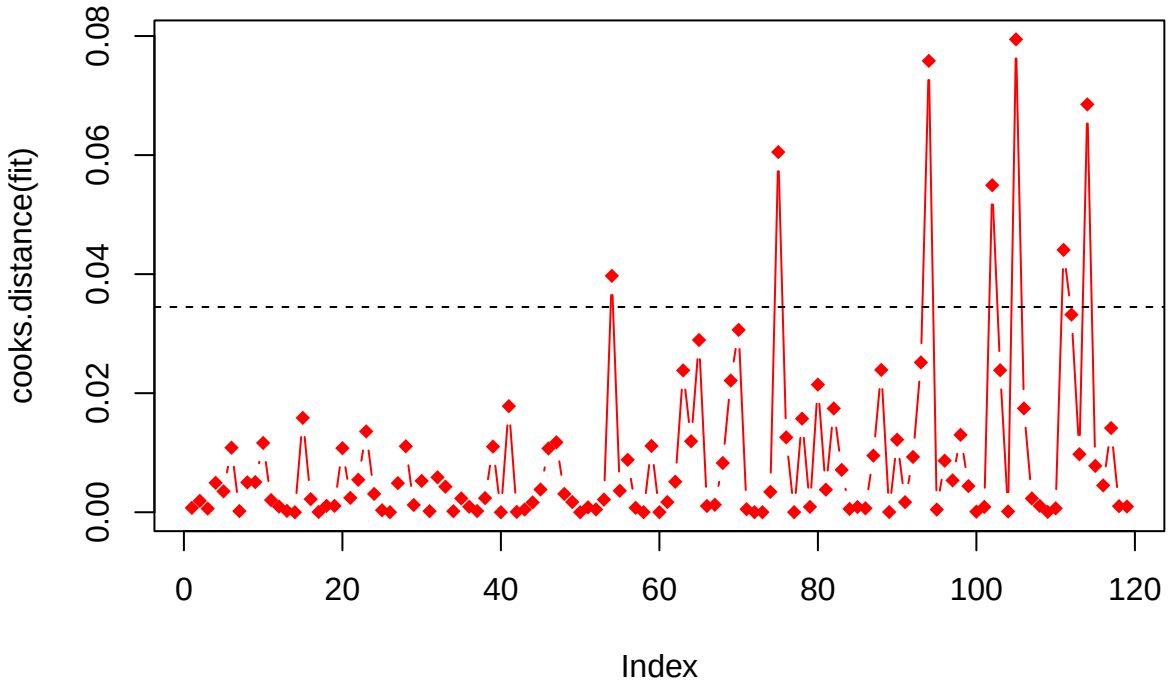
Wir erhalten die Schätzung $\hat{\alpha} = 1.6060$ und $\hat{\beta} = 1.8990$.

b) Welches sind die drei einflussreichsten Beobachtungen laut der Cook's Distance?

```
# Berechnung der Cook's Distance
plot(fit, main = "Cook's Distance", which = 4)
```




```
plot(cooks.distance(fit), type = "b", pch = 18, col = "red")
N <- nrow(aktien)
k <- 2
cutoff <- 4 / (N - k - 1)
abline(h = cutoff, lty = 2)
```



Es sind die Beobachtungen 94, 105 und 114. Allerdings hat keine dieser Beobachtungen eine Cook's Distance, die als "aussergewöhnlich" in Relation zur Grundgesamtheit beurteilt werden kann. Bemerkung: Es ist hier nicht hilfreich, sich univariate Boxplots anzuschauen. Datenpunkte, die univariate Ausreisser sind (z.B. in der Stichprobe der Sun-Daten) müssen keine Ausreisser in der bivariaten Beziehung S&P–Sun mehr sein und umgekehrt!

c) Kann die Aktie als 'defensiv' oder 'aggressiv' beurteilt werden?

```
# Konfidenz-Intervall für Beta
confint(fit)["sp500.ex", ]
```

```
##    2.5 %   97.5 %
## 1.353941 2.443974
```

Da dieses KI die 1 nicht enthält, können wir die Aktie als "aggressiv" in Relation zum Index beurteilen.

d) Gibt es Anzeichen dafür, dass das CAPM verletzt ist?

Wir testen $H_0 : \alpha = 0$ gegen $H_A : \alpha \neq 0$.

```
# Hypothesentest für Alpha
summary(fit)$coefficients["(Intercept)", ]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##  1.6059517  1.1327348  1.4177650  0.1589172
```

Der zugehörige p-Wert (direkt von R berechnet für uns) ist 0.159. Daher gibt es kein Anzeichen für eine Verletzung des CAPM. Wir kommen zum gleichen Ergebnis wenn wir ein 95% Konfidenz-Intervall für α in

der folgenden Weise berechnen: $1.606 \pm 1.96 \times 1.133 = [-0.62, 3.83]$. Da dieses Intervall die 0 enthält, ist es 'plausibel', dass $\alpha = 0$ und somit können wir H_0 nicht verwerfen.

Bemerkung: ein Test für die Differenz der beiden Mittelwerte (Aktivum und Markt) ist hier nicht angebracht! Wenn ein Aktivum ein $\beta < 1$ hat, dann kann $\alpha > 0$ sein, obwohl der Erwartungswert des Aktivums kleiner ist als der des Marktes.

e) Vorhersage der Überschussrendite bei einer neuen Monatsrisikoprämie von 3%

```
# Vorhersage
new_data <- data.frame(sp500.ex = 3)
predict(fit, newdata = new_data)
```

```
##        1
## 7.302824
```

Die Vorhersage ist 7.31. Allerdings ist diese Vorhersage nicht sehr zuverlässig, da $R^2 = 0.29$. D.h. die Risiko-Prämie des Indexes erklärt nur 29% der Variation der Risiko-Prämie von Sun Microsystems.

f) 95% Konfidenz-Intervall für die erwartete Überschussrendite

```
# Konfidenz-Intervall
predict(fit, newdata = new_data, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 7.302824 4.698463 9.907186
```

Aufgabe 4

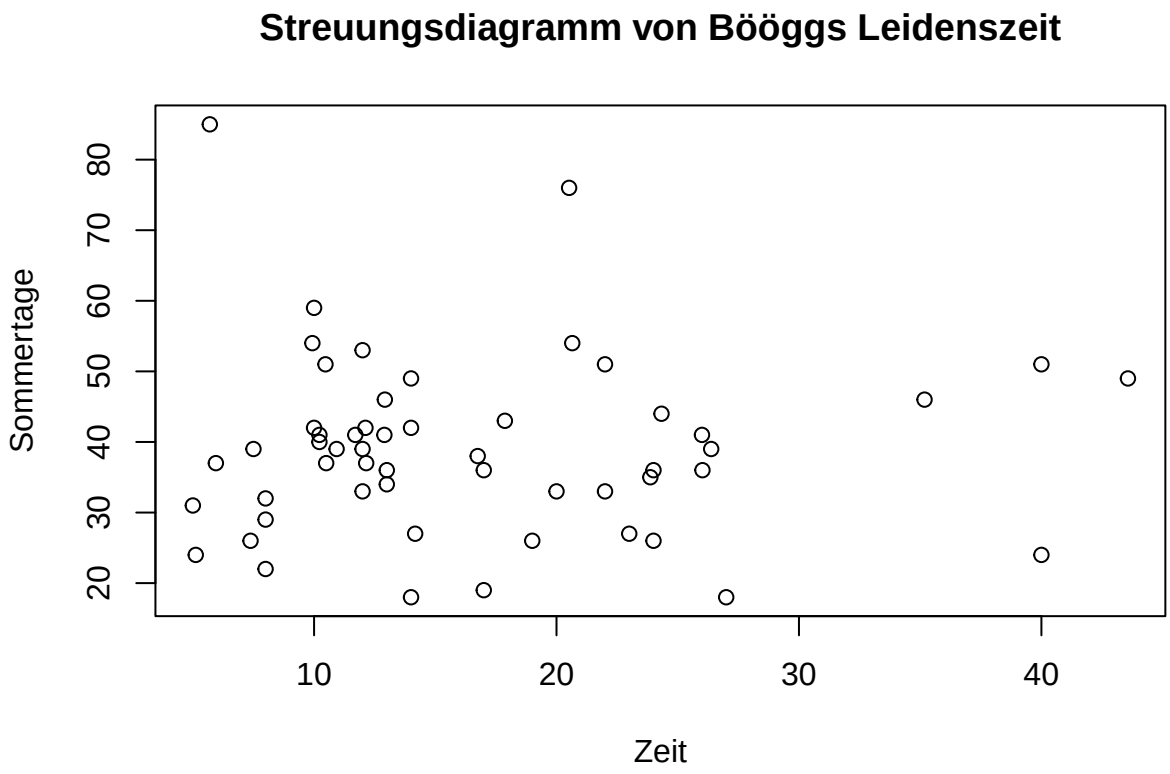
Der Datensatz enthält die Leidenszeit des Böögg (“time”) und die Anzahl der Sommertage (“days”) für die Jahre 1965 bis 2018.

```
# Einlesen der Daten für Aufgabe 4
boegg <- read.csv("data/boegg.csv")
head(boegg)
```

```
##   year days time
## 1 1965   33   20
## 2 1966   42   10
## 3 1967   59   10
## 4 1968   32    8
## 5 1969   49   14
## 6 1970   51   40
```

a) Erstellen Sie ein Streuungs-Diagramm und beurteilen Sie die Beziehung.

```
# Streudiagramm
plot(
    boegg$time, boegg$days,
    xlab = "Zeit", ylab = "Sommertage", main = "Streuungsdiagramm von Böögg Leidenszeit"
)
```



b) Wie lautet das geschätzte Modell?

```
# Lineares Modell
fit_boegg <- lm(days ~ time, data = boegg)
summary(fit_boegg)
```

```
##
## Call:
## lm(formula = days ~ time, data = boegg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.044  -6.843  -0.563   4.791  45.874
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.181278   3.686662   10.63 1.22e-14 ***
## time        -0.009776   0.194991   -0.05     0.96
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.84 on 52 degrees of freedom
## Multiple R-squared:  4.834e-05,  Adjusted R-squared:  -0.01918
## F-statistic: 0.002514 on 1 and 52 DF,  p-value: 0.9602
```

Die Beziehung ist sehr schwach, da weniger als 0.01% der beobachteten Variation in days durch time erklärt

wird.

c) Evidenz für die Behauptung des Volksmunds

```
# Hypothesentest
summary(fit_boegg)$coefficients["time", ]
```

```
##    Estimate  Std. Error     t value    Pr(>|t|)
## -0.00977622  0.19499126 -0.05013671  0.96020554
```

Der (einseitige) p-Wert ist $0.96/2 = 0.48$ (da die Test-Statistik negativ ist). Daher haben wir keine Evidenz für die Behauptung des Volksmunds.

d) Vorhersage der Sommertage im Jahr 2019

```
# Vorhersage für 2019
new_data <- data.frame(time = 17.73)
predict(fit_boegg, newdata = new_data, interval = "prediction", level = 0.90)
```

```
##        fit      lwr      upr
## 1 39.00795 17.30444 60.71145
```

Die Vorhersage ist $\hat{y}_{neu} = 39.0$. Das Vorhersage-Intervall ist $[17.3, 60.7]$. Es wäre aber keine gute Idee: sowohl vom Streuungs-Diagramm mit geschätzter Gerade als auch vom Normal-Quantil-Plot können wir sehen, dass die Residuen rechtsschief sind.

e) Änderung der Ergebnisse ohne einflussreiche Beobachtungen

```
# Entfernung der einflussreichen Beobachtungen
cooksd_boegg <- cooks.distance(fit_boegg)
influential_boegg <- which(cooksd_boegg > (4 / (nrow(boegg) - 2 - 1)))
boegg_no_infl <- boegg[-influential_boegg, ]

# Neues Modell ohne einflussreiche Beobachtungen
fit_boegg_no_infl <- lm(days ~ time, data = boegg_no_infl)
summary(fit_boegg_no_infl)
```

```
##
## Call:
## lm(formula = days ~ time, data = boegg_no_infl)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.1848  -5.1901  -0.1865   4.8135  21.8116
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.197221   3.405213  10.924 1.71e-14 ***
## time        -0.000886   0.203268  -0.004    0.997
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.738 on 47 degrees of freedom
## Multiple R-squared:  4.042e-07,  Adjusted R-squared:  -0.02128
## F-statistic: 1.9e-05 on 1 and 47 DF,  p-value: 0.9965
```

```r
predict(fit_boegg_no_infl, data.frame(year = 2019, days = 10, time = 17.73),
    interval = "prediction", level = 0.90
)
```

```
##        fit      lwr      upr
## 1 37.18151 20.65428 53.70875
```

Wenn diese Beobachtung entfernt wird, ändert sich so gut wie nichts.

Aufgabe 5

Die Datei "cocaine.csv" enthält 56 Observationen von Variablen welche im Zusammenhang mit dem Verkauf von Kokain in Nordosten von Kalifornieren zwischen 1984-1991 stehen.

Die Variablen sind:
- price: Preis pro Gramm in Dollar
- quant: Menge in Gram für eine Transaktion
- qual: Qualität des Kokains in Reinheitsgrad in Prozent
- trend: Eine Zeitvariable mit 1984 = 1, 1985 = 2, ..., 1991 = 8

Betrachten wir folgendes Regressionsmodell:
$price = \beta_1 + \beta_2 \cdot quant + \beta_3 \cdot qual + \beta_4 \cdot trend + \varepsilon$

```r
# Einlesen der Daten für Aufgabe 5
cocaine <- read.csv("data/cocaine.csv")
```

a) Welche Vorzeichen erwarten Sie für die Koeffizienten β_2, β_3 und β_4?

Negativ, Positiv, nicht klar.

b) Schätzen Sie das Modell und interpretieren Sie die Koeffizienten. Sind die Vorzeichen wie erwartet?

```r
fit_cocaine <- lm(price ~ quant + qual + trend, data = cocaine)
summary(fit_cocaine)
```

```
##
## Call:
## lm(formula = price ~ quant + qual + trend, data = cocaine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.479 -12.014  -3.743  13.969  43.753
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 90.84669    8.58025  10.588 1.39e-14 ***
## quant       -0.05997    0.01018  -5.892 2.85e-07 ***
## qual         0.11621    0.20326   0.572   0.5700
## trend       -2.35458    1.38612  -1.699   0.0954 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.06 on 52 degrees of freedom
## Multiple R-squared:  0.5097, Adjusted R-squared:  0.4814
```

```
## F-statistic: 18.02 on 3 and 52 DF,  p-value: 3.806e-08
```

c) Wie gross ist der Anteil der erklärten Variation von *price* durch die Variation von *quant*, *qual* und *trend*?

51%, siehe b).

d) Es wird behauptet, dass das Risiko aufzufliegen mit der Verkaufssumme steigt. Die Verkäufer wären also bereit, tiefere Preise zu akzeptieren, wenn die Menge steigt. Setzten sie eine passende H_0 und H_A auf und testen Sie die Hypothese.

Wir testen $H_0 : \beta_2 = 0$ gegen $H_A : \beta_2 < 0$. Die zugehörige Test-Statistik ist $t = -5.892$ (siehe b) und der zugehörige p-Wert ist ≈ 0. (Da es sich um einen einseitigen Test handelt, und die Test-Statistik negativ ist, müssen wir den angegebenen p-Wert 2.85e-07 halbieren.) Damit ist die Vermutung bestätigt.

e) Testen sie die Hypothese, dass die Qualität des Kokains keinen Einfluss auf den Preis hat gegen die Alternative, dass die Qualität einen Einfluss auf den Preis hat.

Wir testen $H_0 : \beta_3 = 0$ gegen $H_A : \beta_3 > 0$. Die zugehörige Test-Statistik ist $t = 0.572$ (siehe b) und der zugehörige (einseitige) p-Wert ist $0.57/2 = 0.285$. Damit können wir H_0 nicht verwerfen und es ist plausibel (aber nicht bewiesen), dass die Qualität keinen Einfluss auf den Preis hat.

f) Was ist die durchschnittliche Änderung des Preises pro Gramm pro Jahr? Können sie erklären, warum sich der Preis in diese Richtung entwickelt?

Preissenkung von \$2.35 pro Jahr. Eine mögliche Erklärung ist, dass mehr und mehr Kokain auf den Markt kommt und der Preis daher sinkt.

g) Kommentieren Sie die Gültigkeit der vorangegangenen Hypothesen-Tests. Benützen Sie hierzu, unter anderem, das verfeinerte Residuen-Diagramm anstelle des 'normalen' Residuen-Diagramms.

Bei den Daten handelt es sich teilweise um eine Zeitreihe. (Es ist keine Zeitreihe im "strikten Sinne", da es mehrere Beobachtungen pro Jahr gibt. Solche Daten heissen *Panel-Daten*.) Also ist Vorsicht geboten. Zudem erkennt man eine Fächer-Form im Residuen-Diagramm (recht klar) und eine halbe Fächer-Form im verfeinerten Residuen-Diagramm (etwas weniger klar). Die Inferenz ist also mit Vorsicht zu geniessen.

h) Ein Verkäufer bietet 1993 ein Paket an mit *quant* = 100 und *qual* = 60. Berechnen Sie ein 90% Vorhersage-Intervall für den Preis, den er erzielen wird. Inwieweit vertrauen Sie diesem Intervall?

```r
# option 1
predict(fit_cocaine, data.frame(price = 0, quant = 100, qual = 60, trend = 10), se = T)
```

```
## $fit
##        1
## 68.27623
##
## $se.fit
## [1] 8.094933
##
## $df
## [1] 52
##
## $residual.scale
## [1] 20.05778
```

```r
# option 2
predict(fit_cocaine, data.frame(price = 0, quant = 100, qual = 60, trend = 10), interval = "prediction"
```

```
##        fit      lwr      upr
## 1 68.27623 32.05327 104.4992
```

Das Vorhersage-Intervall ist daher $[32.0, 104.5]$. Wir haben weiterhin die Bedenken aus Teil g). Zusätzlich gibt das Normal-Quantil-Diagramm einen (leichten) Hinweis auf das "Heavy Tails" Muster. Also ist das Vorhersage-Intervall zusammengenommen nicht sehr vertrauenswürdig.

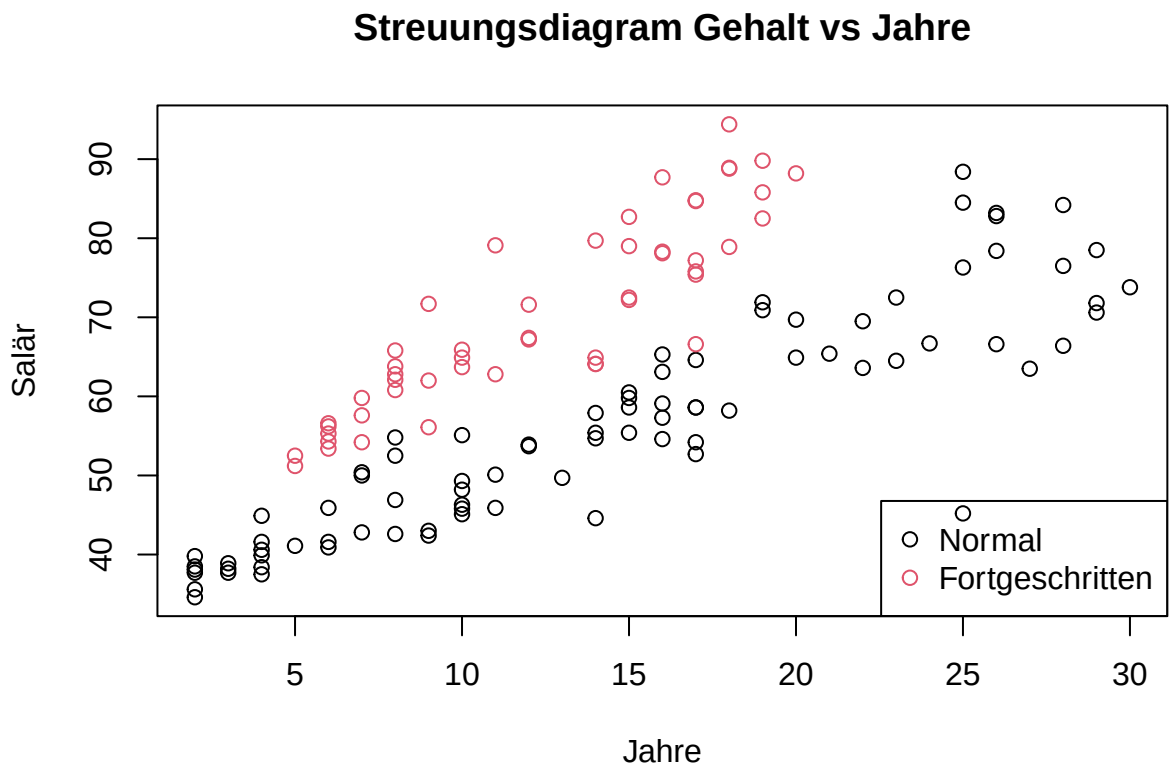
Aufgabe 6

Eine junge Person arbeitet seit drei Jahren in ihrem Beruf. Sie fragt sich, ob sie in ihrer Karriere weitermachen sollte oder ob sie einen fortgeschrittenen Studienabschluss machen sollte, um danach in das Berufsleben zurückzukehren. Der Datensatz "berufe.csv" enthält Daten von einer Zufalls-Stichprobe von Arbeitern in der zugehörigen Industrie:
- Gehalt (in 1'000 Euro)
- Arbeitsjahre
- Art des Studienabschlusses (0 für normal und 1 für fortgeschritten)

```r
# Einlesen der Daten für Aufgabe 6
berufe <- read.csv("data/berufe.csv")
```

a) Erstellen Sie ein Streuungsdiagramm mit unterschiedlichen Farben für die beiden Abschlüsse. Welche 'Botschaft' vermittelt Ihnen dieses Diagramm?

```r
plot(berufe$jahre, berufe$gehalt,
    col = berufe$abschluss + 1, xlab = "Jahre", ylab = "Salär", main =
        "Streuungsdiagram Gehalt vs Jahre"
)
legend("bottomright", legend = c("Normal", "Fortgeschritten"), col = 1:2, pch = 1)
```



Es suggeriert, was die formale Analyse später bestätigen wird. Die Gerade für die Gruppe mit dem fortgeschrittenen Abschluss hat nicht nur einen höheren Abschnitt sondern auch eine grössere Steigung. Ausserdem enthält die Gruppe mit dem normalen Abschluss einen klaren Ausreisser.

b) Schätzen das Modell Gemeinsame Gerade für beide Gruppen (normaler und fortgeschrittener Abschluss) und interpretieren Sie das geschätzte Modell.

```
fit_beruf <- lm(gehalt ~ jahre, data = berufe)
summary(fit_beruf)
```

```
##
## Call:
## lm(formula = gehalt ~ jahre, data = berufe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.592  -7.711  -1.650   7.868  26.752
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.5649     1.9051   21.82   <2e-16 ***
## jahre         1.4491     0.1237   11.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.45 on 133 degrees of freedom
```

```
## Multiple R-squared:  0.5079, Adjusted R-squared:  0.5042
## F-statistic: 137.3 on 1 and 133 DF,  p-value: < 2.2e-16
```

G = Gehalt, J = Jahre und A = Abschluss.
Geschätzes Modell: $\widehat{G} = 41.56 + 1.45J$. Somit ist die geschätzte durchschnittliche Gehaltserhöhung 1'450 pro Jahr. Das geschätzte durchschnittliche Anfangsgehalt beträgt 41'5600.

c) Sollten Sie zu dem Modell Gemeinsame Steigung übergehen? Falls ja, interpretieren Sie das geschätze Modell.

```
fit_beruf_2 <- lm(gehalt ~ jahre + abschluss, data = berufe)
summary(fit_beruf_2)
```

```
##
## Call:
## lm(formula = gehalt ~ jahre + abschluss, data = berufe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.6683  -3.4676  -0.4734   3.5420  15.5317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.85059    1.22763   26.76   <2e-16 ***
## jahre        1.60071    0.07229   22.14   <2e-16 ***
## abschluss   17.61713    1.08448   16.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.058 on 132 degrees of freedom
## Multiple R-squared:  0.8359, Adjusted R-squared:  0.8334
## F-statistic: 336.2 on 2 and 132 DF,  p-value: < 2.2e-16
```

Der p-Wert der zusätzlichen Variable A ist ungefähr gleich Null, also sollten wir zum Modell Gemeinsame Steigung übergehen. Das geschätzte Modell ist $\widehat{G} = 32.85 + 17.62A + 1.60J$. Somit ist die geschätzte gemeinsame durchschnittliche Gehaltserhöhung 1'600 pro Jahr. Das geschätzte durchschnittliche Anfangsgehalt beträgt 32'850 in der Gruppe mit dem normalen Abschluss und 32'850 + 17'620 = 50'470 in der Gruppe mit dem fortgeschrittenen Abschluss.

d) Sollten Sie zu dem Modell Total Verschieden übergehen? Falls ja, interpretieren Sie das geschätzte Modell.

```
fit_beruf_3 <- lm(gehalt ~ jahre + abschluss + jahre:abschluss, data = berufe)
summary(fit_beruf_3)
```

```
##
## Call:
## lm(formula = gehalt ~ jahre + abschluss + jahre:abschluss, data = berufe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.3552  -3.2322  -0.2437   3.2656  16.8448
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      34.6081     1.2268  28.211  < 2e-16 ***
## jahre             1.4779     0.0739  19.999  < 2e-16 ***
## abschluss         7.4245     2.6042   2.851  0.00507 **
## jahre:abschluss   0.8046     0.1891   4.254 3.97e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.7 on 131 degrees of freedom
## Multiple R-squared:  0.8558, Adjusted R-squared:  0.8525
## F-statistic: 259.2 on 3 and 131 DF,  p-value: < 2.2e-16
```

Der p-Wert der zusätzlichen Variable $(A \times J)$ ist ungefähr gleich Null, also sollten wir zu "Total Verschieden" übergehen. Das geschätzte Modell ist:

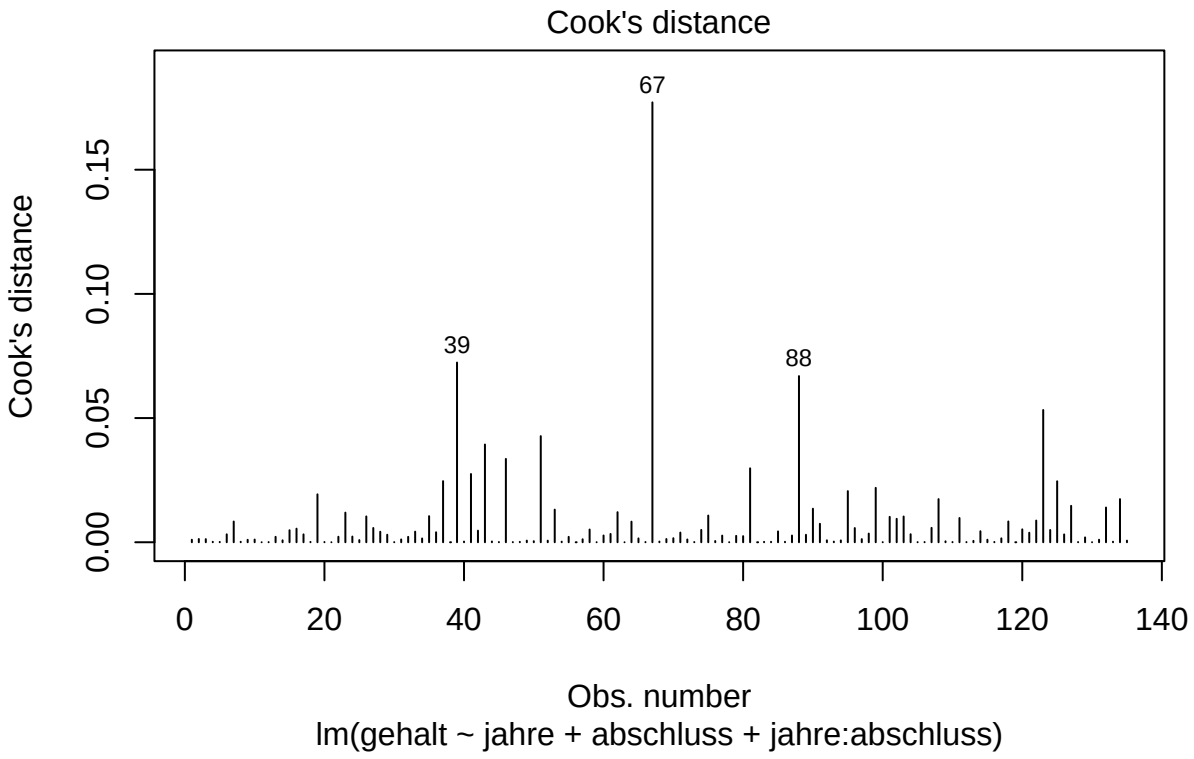
$$\hat{G} = 34.61 + 7.42A + 1.48J + 0.80(A \times J)$$

.

Somit ist die geschätzte durchschnittliche Gehaltserhöhung 1'480 pro Jahr in der Gruppe mit dem normalen Abschluss und 1'480 + 800 = 2'280 in der Gruppe mit dem fortgeschrittenen Abschluss. Das geschätzte durchschnittliche Anfangsgehalt beträgt 34'610 in der Gruppe mit dem normalen Abschluss und 34'610 + 7'420 = 40'030 in der Gruppe mit dem fortgeschrittenen Abschluss.

e) In dem Modell, das Ihnen am besten erscheint, suchen Sie nach (klaren) Ausreissern und entfernen Sie diese. Für den Rest der Aufgabe benützen Sie dann dieses Modell. Was ist der Prozentsatz der beobachteten Variation von Gehalt, der von diesem Modell erklärt wird.

```r
# Suchen nach Ausreissern (Cook's Distance, Obs. 67 sieht nach Ausreisser aus)
plot(fit_beruf_3, which = 4)
```




```
fit_beruf_4 <- lm(gehalt ~ jahre + abschluss + jahre:abschluss, data = berufe[-67, ])
summary(fit_beruf_4)
```

```
##
## Call:
## lm(formula = gehalt ~ jahre + abschluss + jahre:abschluss, data = berufe[-67,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.235  -3.436  -0.331   3.047  15.998
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     34.23259    1.12537  30.419  < 2e-16 ***
## jahre            1.52676    0.06831  22.350  < 2e-16 ***
## abschluss        7.80007    2.38506   3.270  0.00138 **
## jahre:abschluss  0.75570    0.17340   4.358 2.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.218 on 130 degrees of freedom
## Multiple R-squared:  0.8791, Adjusted R-squared:  0.8763
## F-statistic: 314.9 on 3 and 130 DF,  p-value: < 2.2e-16
```

Datenpunkt Nr. 67 ist ein klarer Ausreisser. Das geschätze Modell nach dem Entfernen diese Punktes ist dann:

$$\hat{G} = 34.23 + 7.8 * A + 1.53 * J + 0.76 * (A \times J)$$

.

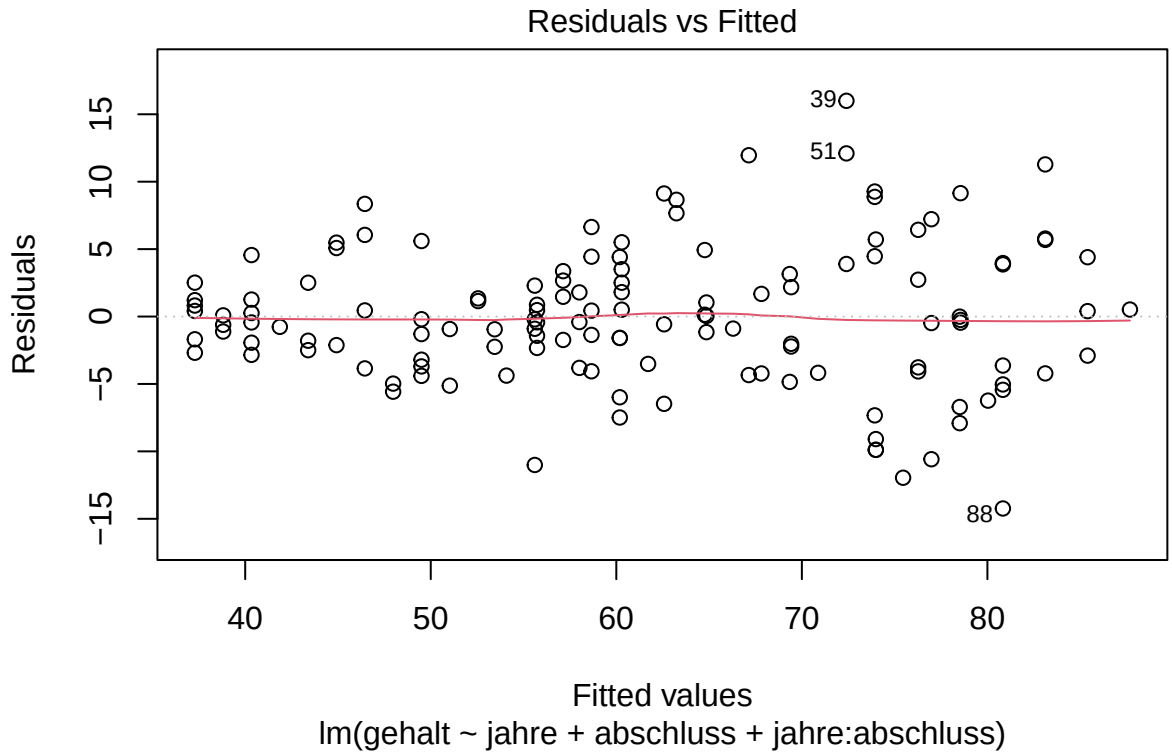
Dieses Modell erklärt 88% der beobachteten Variation des Gehalts.

f) Finden Sie ein 90% Konfidenz-Intervall für die Steigung von Arbeitsjahren in der Gruppe mit dem normalen Abschluss. Benützen Sie hierzu die standard OLS-Inferenz. Glauben Sie, dass Sie diesem Intervall vertrauen können?

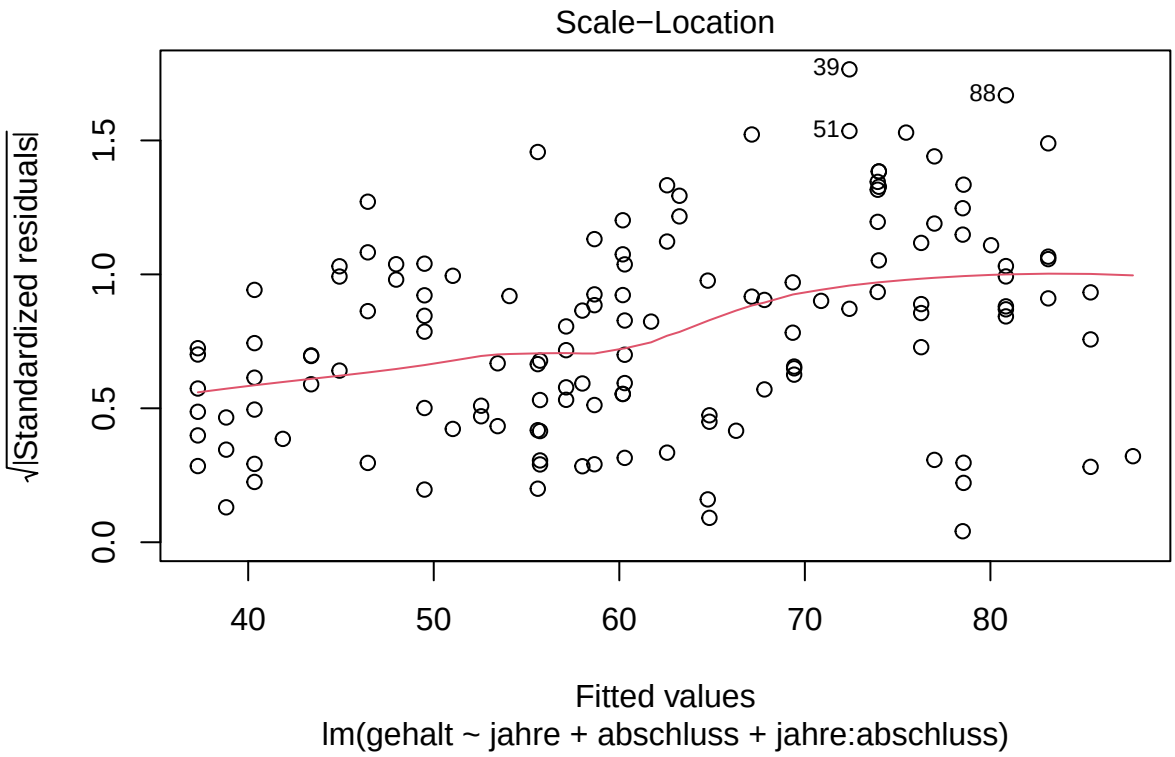
```
confint(fit_beruf_4, level = 0.9)["jahre", ]
```

```
##      5 %     95 %
## 1.413588 1.639930
```

```
plot(fit_beruf_4, which = 1)
```




```
plot(fit_beruf_4, which = 3)
```



Das 90% Konfidenz-Intervall ist $[1.41, 1.64]$. Jedoch zeigt das Residuen-Diagramm eine Fächer-Form und, äquivalent, das verfeinerte Residuen-Diagramm eine halbe Fächer-Form. Also sollten wir diesem Intervall nicht trauen.

g) Falls Ihre Antwort in (f) nein war, finden Sie ein KI, dem Sie vertrauen können. Benützen Sie hierzu geeignete HC-Inferenz.

```
# HC3 Inference
lmtest::coeftest(fit_beruf_4, vcov = sandwich::vcovHC(fit_beruf_4, type = "HC3"))["jahre", ]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 1.526759e+00 7.961971e-02 1.917564e+01 1.027643e-39
```

Das 90% Konfidenz-Intervall basierend auf dem HC3 Standardfehler ist gegeben als $1.53 \pm 1.645 \times 0.080 = [1.40, 1.66]$, und diesem Intervall können wir vertrauen. In der Regel (wie auch hier), jedoch nicht immer, werden die Konfidenz-Intervalle für Steigungen im Falle der Berücksichtigung von HC etwas länger.

h) Nehmen wir an die Person ist momentan 27 Jahre alt und dass es zwei Jahre dauern würde, den fortgeschrittenen Abschluss zu erwerben. Sagen Sie ihr Gehalt im Alter von 30 Jahren und 45 Jahren vorher für jede der beiden folgenden Strategien (und ohne die Inflation in Betracht zu ziehen):

h1) Sie arbeitet in ihrer Industrie weiter.

h2) Sie erwirbt den fortgeschrittenen Abschluss und kehrt dann in ihre Industrie zurück, um dort weiterzuarbeiten.

(**Bemerkung**: nur Vohersagen, keine Vorhersage-Intervalle.)

Die Person ist 27 Jahre alt und hat momentan 3 Arbeitsjahre. Dies ergibt die folgenden Arbeitsjahre für die betrachteten Alter und Strategien. (Bemerkung: den fortgeschrittenen Abschluss zu erwerben "kostet" die Person zwei Arbeitsjahre).

	Strategie (g1)	Strategie (g2)
Alter = 30	6	4
Alter = 45	21	19

Wenn man diese Werte in das geschätzte Modell von e) einsetzt, erhält man die folgenden Vorhersagen.

	Strategie (g1)	Strategie (g2)
Alter = 30	43.41	66.36
Alter = 45	51.19	85.54

Bemerkung: Strenggenommen ist die standard OLS-Inferenz in b) und c) nicht gültig, da die Annahme der konstanten Fehler-Standardabweichung verletzt ist. Aber die p-Werte sind so extrem klein, dass sich die Resultate nicht ändern würden wenn man eine allgemeinere Inferenzmethode wählen würde, die diese Annahme nicht benötigt. Sie können dies nachprüfen, indem Sie die Inferenz stattdessen auf den HC3 Standardfehlern basieren. Alternativ könnte man auch das Kriterium der adjustierten R^2-Statistik wühlen und küme zu dem gleichen Schluss: man muss schlussendlich zum Modell "Total Verschieden" übergehen.

Aufgabe 7

Betrachten Sie Produktionsfunktionen der Art $Q = f(L, K)$, wobei Q ein Mass für Output ist, L Labor-Input ist und K Kapital-Input ist. Eine populäre funktionale Form ist die Cobb-Douglas-Gleichung:

$$\log Q_i = \beta_1 + \beta_2 \log L_i + \beta_3 \log K_i + u_i$$

Die Daten für diese AUfgabe sind in der Datei "production.csv" gespeichert.

```
# Einlesen der Daten für Aufgabe 7
production <- read.csv("data/production.csv")
```

a) Testen Sie die constant returns to scale Hypothese, d.h., $H_0 : \beta_2 + \beta_3 = 1$

```
fit_production <- lm(log(Q) ~ log(L) + log(K), data = production)
fit_constant_returns <- lm(I(log(Q) - log(K)) ~ I(log(L) - log(K)), data = production)
summary(fit_production)
```

```
##
## Call:
## lm(formula = log(Q) ~ log(L) + log(K), data = production)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55363 -0.09179 -0.00146  0.13888  0.49386
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  -0.1287     0.5461  -0.236    0.815
## log(L)        0.5590     0.8164   0.685    0.499
## log(K)        0.4877     0.7039   0.693    0.494
##
## Residual standard error: 0.2167 on 30 degrees of freedom
## Multiple R-squared:  0.6883, Adjusted R-squared:  0.6675
## F-statistic: 33.12 on 2 and 30 DF,  p-value: 2.547e-08
```

```
summary(fit_constant_returns)
```

```
##
## Call:
## lm(formula = I(log(Q) - log(K)) ~ I(log(L) - log(K)), data = production)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5514 -0.1054 -0.0100  0.1289  0.4818
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.02010    0.05293   0.380    0.707
## I(log(L) - log(K))   0.39840    0.55927   0.712    0.482
##
## Residual standard error: 0.2135 on 31 degrees of freedom
## Multiple R-squared:  0.01611,    Adjusted R-squared:  -0.01563
## F-statistic: 0.5075 on 1 and 31 DF,  p-value: 0.4816
```

Somit haben wir $SSR_0 = 0.2135^2 \times 31 = 1.413$ und $SSR_A = 0.2167^2 * 30 = 1.409$. Die Test-Statistik ist dann:

$$F = \frac{(1.413 - 1.409)/1}{1.409/30} = 0.0852$$

Der p-Wert ist damit $P(F_{1,30} \geq 0.0852) = 0.77$. Somit ist die "constant returns to scale" Hypothese plausibel. Alternativ können wir den F-Test auch "direkt" ausführen:

```
car::linearHypothesis(fit_production, hypothesis.matrix = c(0, 1, 1), rhs = 1)
```

```
## Linear hypothesis test
##
## Hypothesis:
## log(L)  + log(K) = 1
##
## Model 1: restricted model
## Model 2: log(Q) ~ log(L) + log(K)
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     31 1.4129
## 2     30 1.4094  1 0.0035203 0.0749 0.7862
```

Die leichten Unterschiede hierbei sind auf die Rundungsfehler zurückzuführen, wenn der Test "von Hand" ausgeführt wird.

b) Man beobachtet $K = 20$ und $L = 25$. Finden Sie ein 95% Vorhersage-Intervall für Q unter der Annahme, dass die constant returns to scale Hypothese gültig ist. Vertrauen Sie diesem Intervall? (Achtung: etwas trickreich...) Bemerkung: Gehen Sie von Homoskedastie hier aus.

```
# Berechnung des Vorhersage-Intervalls "von Hand"
predicition <- predict(fit_constant_returns, data.frame(K = 20, L = 25, Q = 0), se = T)
quantile <- qt(0.975, 31)
lower_bound <- exp(predicition$fit - quantile * sqrt(predicition$se.fit^2 + predicition$residual.scale^
upper_bound <- exp(predicition$fit + quantile * sqrt(predicition$se.fit^2 + predicition$residual.scale^
print(paste0("[", lower_bound, ", ", upper_bound, "]"))
```

```
## [1] "[13.8515036357709, 35.9118785901467]"
```

Das geschützte Modell unter H_0 gibt uns ein Vorhersage-Intervall für $\log(Q) - \log(20)$. Um dieses in ein Vorhersage-Intervall für Q umzuwandeln, müssen wir auf jeden Endpunkt die Invers-Transformation $f(a) = \exp(a + \log(20))$ anwenden. Das Ergebnis ist das Intervall $[13.9, 35.9]$. Etwas trickreich, zugegebenermassen.

Schneller erhalten wir das Vorhersage-Intervall in der transformierten Welt wie folgt:

```
exp(predict(fit_constant_returns, data.frame(K = 20, L = 25, Q = 0),
    interval = "prediction", level = 0.95
) + log(20))
```

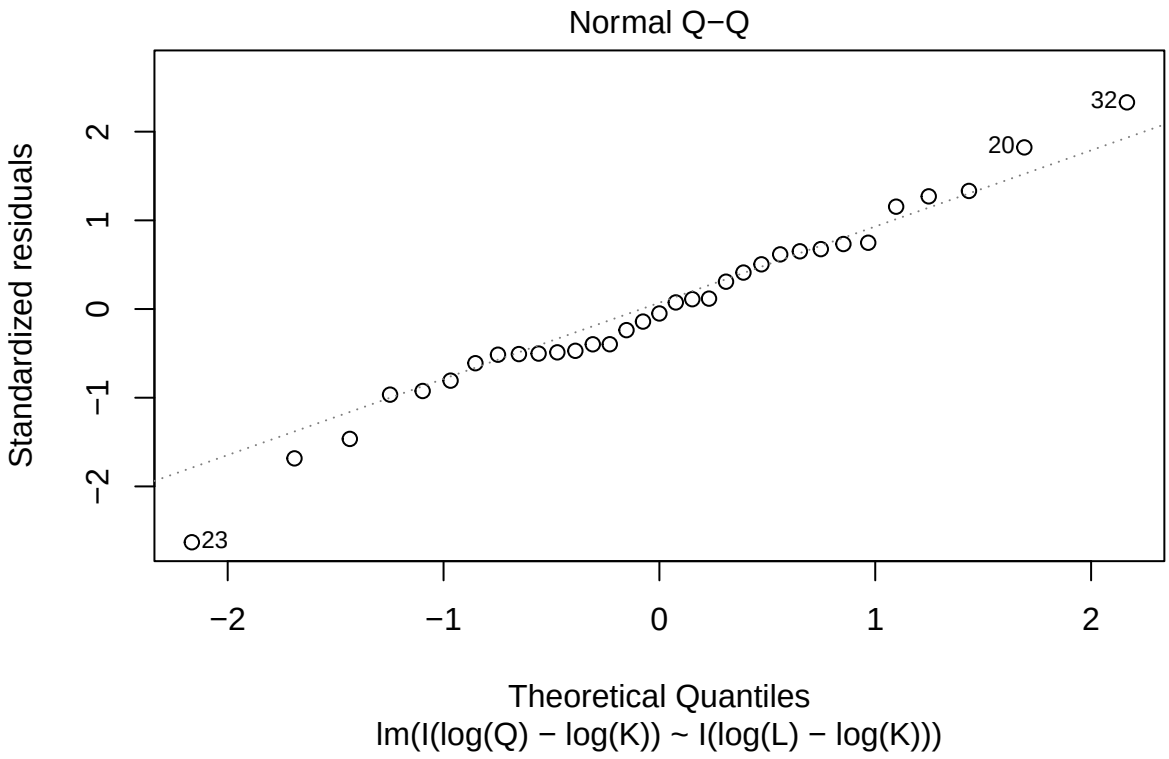
```
##        fit     lwr      upr
## 1 22.30322 13.8515 35.91188
```

Das Endergebnis ist (bis auf kleine Rundungsfehler) natürlich identisch.

Allerdings ist der Stichprobenumfang recht klein mit $n = 30$ und das Normal-Quantil-Diagramm zeigt das "Heavy Tails" Muster. Man sollte dem Vorhersage-Intervall daher nicht übermässig vertrauen.

Gefahrendiagramm:

```
plot(fit_constant_returns, which = 2)
```


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(I(log(Q) − log(K)) ~ I(log(L) − log(K)))
32
20
23
−2
−1
0
1
2