Übungszettel Multiple Regression, Exercise Sheet Multiple Regression

M.Psy.205, Dozent: Peter Zezula

Johannes Brachem (johannes.brachem@stud.uni-goettingen.de)

25 Mai, 2022 08:58

German

Links

Übungszettel als PDF-Datei zum Drucken

Übungszettel mit Lösungen

Lösungszettel als PDF-Datei zum Drucken

Der gesamte Übugszettel als .Rmd-Datei (Zum Downloaden: Rechtsklick > Speichern unter...)

Hinweise zur Bearbeitung

1. Bitte beantworten Sie die Fragen in einer .Rmd Datei. Sie können Sie über `Datei > Neue Datei > R Markdown...` eine neue R Markdown Datei erstellen. Den Text unter dem *Setup Chunk* (ab Zeile 11) können Sie löschen. Unter diesem Link können Sie auch unsere Vorlage-Datei herunterladen.
2. Informationen, die Sie für die Bearbeitung benötigen, finden Sie auf der Website der Veranstaltung
3. Zögern Sie nicht, im Internet nach Lösungen zu suchen. Das effektive Suchen nach Lösungen für R-Probleme im Internet ist tatsächlich eine sehr nützliche Fähigkeit, auch Profis arbeiten auf diese Weise. Die beste Anlaufstelle dafür ist der R-Bereich der Programmiererplattform Stackoverflow
4. Auf der Website von R Studio finden Sie sehr hilfreiche Übersichtszettel zu vielen verschiedenen R-bezogenen Themen. Ein guter Anfang ist der Base R Cheat Sheet

Ressourcen

Da es sich um eine praktische Übung handelt, können wir Ihnen nicht alle neuen Befehle einzeln vorstellen. Stattdessen finden Sie hier Verweise auf sinnvolle Ressourcen, in denen Sie für die Bearbeitung unserer Aufgaben nachschlagen können.

Ressource	Beschreibung
Field, Kapitel 7	Buchkapitel, das Schritt für Schritt erklärt, worum es geht, und wie man Regressionen in R durchführt. **Große Empfehlung!**

Tipp der Woche

Diese Woche gibt es zwei Tipps:

Mehrere Zeilen auf einmal auskommentieren

Mit der Tastenkombination `strg` + `shift` + `c` (Windows), bzw. `cmd` + `shift` + `c` (Mac) wird der gerade markierte Code auskommentiert. So können Sie beliebig viele Code-Zeilen innerhalb eines Chunks auf einmal auskommentieren.

Kein Stress mehr mit Anführungszeichen

Ihnen ist vielleicht aufgefallen, dass R automatisch zwei Anführungzeichen einfügt, wenn man `"` drückt. Das ist oft praktisch, nervt aber, wenn man ein Wort, das schon im Code steht, in Anführungszeichen setzen will. Dann führt es häufig dazu, dass der Code so aussieht: `"Text""`. Dafür gibt es einen Trick:

1. Markieren Sie das Wort, das Sie in Anführungszeichen setzen wollen.
2. Drücken Sie jetzt `"`. Das Wort wird automatisch in Anführungszeichen gesetzt.

Das gleiche funktioniert auch mit allen Arten von Klammern!

1) Daten einlesen

1. Setzen Sie ein sinnvolles Arbeitsverzeichnis für den Übungszettel (in der Regel der Ordner, in dem Ihre .Rmd liegt). Fügen Sie eine passende Code-Zeile an den Anfang ihres .Rmd-Dokuments ein.
2. Laden Sie die Pakete des `tidyverse` und fügen Sie eine entsprechende Code-Zeile an den Anfang ihrs .Rmd-Dokuments ein.
3. Lesen Sie den Datensatz `child_aggression.csv` mit dem Befehl `read_csv()` direkt aus der URL `https://md.psych.bio.uni-goettingen.de/mv/data/div/child_aggression.csv` in R ein.
4. Nutzen Sie den Befehl `write_csv()`, um den Datensatz in Ihrem Arbeitsverzeichnis in Ihrem Ordner für Daten abzuspeichern.
5. Falls Sie Probleme mit dem einlesen per URL hatten, können Sie den Datensatz unter diesem Link wie gewohnt herunterladen, in Ihrem Arbeitsverzeichnis speichern und einlesen.

Übersicht über den Datensatz

Hier zunächst einmal eine Übersicht über die Variablen im Datensatz. Der Datensatz enthält Daten von 666 Kindern.

Variable	Bedeutung
aggression	Je höher, desto mehr Aggression zeigt das Kind.
television	Je höher, desto mehr Zeit verbirngt das Kind vor dem Fernseher.
computer_games	Je höher, desto mehr Zeit verbringt das Kind mit Computerspielen.
sibling_aggression	Je höher, desto mehr Aggression zeigt der/die ältere Bruder/Schwester.
diet	Je höher, desto gesünder ist die Ernährung des Kindes.
parenting_style	Je höher, desto dysfunktionaler ist der Erziehungsstil der Eltern.

Lösung

Unteraufgabe 1 Bitte folgen Sie den Anweisungen im Aufgabentext.

```
library(tidyverse)
```

Unteraufgabe 2

```
child_data <- read_csv("https://md.psych.bio.uni-goettingen.de/mv/data/div/child_aggression.csv")
```

Unteraufgabe 3

```
## Rows: 666 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (6): aggression, television, computer_games, sibling_aggression, diet, parenting_style
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# example code, not executed
write_csv(child_data, "data/child_aggression.csv")
```

Unteraufgabe 4

```
# example code, not executed
child_data <- read_csv("data/child_aggression.csv")
```

Unteraufgabe 5

2) Multiple Regression: Hypothesentest

Aufgrund vorheriger Forschung haben Sie die Hypothesen, dass der Erziehungsstil und die Aggresionswerte der Geschwister gute Prädiktoren für das Aggressionslevel von Kindern sind.

1. Spezifizieren Sie ein Regressionsmodell, mit dem Sie die Aggression durch den Erziehungsstil und die Aggressionswerte der Geschwister vorhersagen lassen.
2. Schauen Sie sich den Output Ihres Modells an und beantworten Sie die folgenden Fragen:
    a) Wie viel Varianz in den Aggressionswerten der Kinder wird durch das Modell insgesamt aufgeklärt?
    b) Sagt das Modell die Aggressionswerte der Kinder besser vorher als ein Nullmodell?
    c) Tragen beide Prädiktoren signifikant zur Vorhersage bei?
    d) Schreiben Sie das geschätzte Regressionsmodell auf.
    e) Errechnen Sie das durch das Modell vorhergesagte Aggressionslevel eines Kindes, wenn der Erziehungsstil den Wert `1` und die Geschwisteraggression den Wert `0.5` hat.
3. Nun wäre es noch interessant zu wissen, welcher Prädiktor einen stärkeren Einfluss auf die Aggressionswerte hat. (Tipp: Werfen Sie einen Blick in den Field, wenn Sie nicht weiterkommen.)
    a) Die Stärke des Einflusses eines Prädiktors können wir nicht anhand des p-Wertes des Prädiktors ablesen. Warum ist das so? (Tipp: Was bedeutet der p-Wert?)
    b) Welcher Prädiktor stärker ist, können wir auch nicht ohne weiteres anhand der Schätzung für den zugehörigen Regressionskoeffizienten ablesen. Warum ist das so?
    c) Installieren Sie das Paket `lm.beta` und laden Sie es anschließend. Fügen Sie eine Code-Zeile zum Laden des Pakets an den Anfang ihrer .Rmd-Datei ein.
    d) Wenden Sie die Funktion `lm.beta()` auf Ihr Regressionsmodell aus 1. an.
    e) Was bedeutet ein standardisierter Regressionskoeffizient von `1`?
    f) Welcher Prädiktor hat einen stärkeren Einfluss auf das kindliche Aggressionslevel?

Lösung

```
m_1 <- lm(aggression ~ parenting_style + sibling_aggression,
          data = child_data)
```

Unteraufgabe 1

```
summary(m_1)
```

Unteraufgabe 2

```
##
## Call:
## lm(formula = aggression ~ parenting_style + sibling_aggression,
##     data = child_data)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.09755 -0.17180  0.00092  0.15405  1.23037
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.005784   0.012065  -0.479    0.632
## parenting_style     0.061984   0.012257   5.057 5.51e-07 ***
## sibling_aggression  0.093409   0.037505   2.491    0.013 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3113 on 663 degrees of freedom
## Multiple R-squared:  0.05325,    Adjusted R-squared:  0.05039
## F-statistic: 18.64 on 2 and 663 DF,  p-value: 1.325e-08
```

a) Wichtig ist hier das Multiple R^2. Der Wert ist hier .053, das entspricht einer Varianzaufklärung von 5.3%
b) Ja, das Modell sagt die Aggressionswerte signifikant besser vorher als das Nullmodell, mit $F(2, 663) = 18.64$ und $p < .001$. Das ist der F-Test am unteren Ende des Outputs.
c) Ja: Die t-Tests für beide Prädiktoren sind signifikant.
    i) Erziehungsstil ist signifikant mit $t(663) = 5.06$ und $p < .001$
    ii) Geschwisteraggression ist signifikant mit $t(663) = 2.49$ und $p = .013$
d) Das Geschätzte Regressionsmodell ist

$$\hat{y}_i = -0.006 + 0.062 \cdot Erziehungsstil + 0.093 \cdot Geschwisteraggression$$

e) Dafür setzen wir einfach die Werte 1 und 0.5 an den entsprechenden Stellen in der Gleichung ein:

$$-0.006 + 0.062 \cdot 1 + 0.093 \cdot 0.5 = 0.1025$$

Für diesen Fall sagen wir also ein Aggressionslevel von 0.1025 vorher.

Unteraufgabe 3

a) Der p-Wert beantwortet lediglich die Frage: Ist der Regressionskoeffizient signifikant verschieden von null? D.h.: Hat der Prädiktor überhaupt einen Einfluss? Daraus können wir noch nicht ableiten, wie stark der Einfluss eines Prädiktors ist.
b) In einem einfachen Regressionsmodell ist die Höhe des geschätzten Regressionskoeffizienten nicht zur davon abhängig, wir stark der Einfluss des Prädiktors ist, sondern auch, in welcher Einheit der Prädiktor gemessen wurde. Wenn man z.B. das Gewicht durch die Körpergröße vorhersagen möchte, dann wird der Regressionskoeffizient unterschiedlich groß sein, je nachdem ob die Körpergröße in cm oder in Metern gemessen wurde, obwohl der Einfluss der Körpergröße in beiden Fällen gleich stark ist. Es gibt aber die Möglichkeit, die Koeffizienten zu standardisieren und so vergleichbar zu machen. Mehr dazu in Abschnitt 7.8.3.2. (Model parameters) in in *Discovering Statistics Using R* (Field, Miles & Field, 2012).
c) Code zum installieren: `install.packages("lm.beta")` (Installieren Sie Pakete immer nur über die Konsole, nicht über das Skript. Es kann sonst Probleme beim Rendern geben.) Code zum laden:

```
library(lm.beta)
```

d) Code:

```
lm.beta(m_1)
```

```
##
## Call:
## lm(formula = aggression ~ parenting_style + sibling_aggression,
##     data = child_data)
##
## Standardized Coefficients::
##        (Intercept)    parenting_style sibling_aggression
##                 NA         0.19406149         0.09557412
```

e) Ein standardisierter Regressionskoeffizient von 1 bedeutet, dass eine Änderung im Prädiktor um eine Standardabweichung zu einer Änderung im Outcome von einer Standardabweichung führt.

f) Der Output des Codes von d) zeigt: Der Erziehungsstil hat einen etwa doppelt so großen Einfluss auf das Aggressionslevel wie die Geschwisteraggression.

3) Multiple Regression: Exploration

Der Datensatz enthält noch weitere Daten, die potentiell als Prädiktoren interessant sein könnten. Wir haben bei diesen Prädiktoren noch keine Hypothesen darüber, welchen Effekt sie haben, und möchten erst einmal schauen, ob wir ein Muster finden können.

1. Spezifizieren Sie ein Regressionsmodell, das das Aggressionslevel der Kinder aus allen potentiellen Prädiktoren im Datensatz (inkl. Erziehungsstil und Geschwisteraggression) vorhersagt.
2. Nutzen Sie den Befehl `anova()`, um einen R^2-Differenzentest zum Vergleich der beiden Modelle durchzuführen.
3. Interpretieren Sie den Output von `anova()`. Was können Sie daraus schließen?
4. Lassen Sie sich eine Zusammenfassung des besser passenden Modells aus Aufg. 3 mit `summary()` anzeigen. Welche Prädiktoren tragen signifikant zur Varianzaufklärung bei?
5. Wenden Sie die Funktion `vif()` aus dem Paket `car` auf das Regressionsmodell an, um die Prädiktoren des Modells auf Multikollinearität zu überprüfen, indem Sie sich den *Variance Inflation Factor* für jeden Prädiktor ausgeben lassen. Besteht Anlass zur Sorge? (Hinweis: Vergessen Sie nicht, das Paket ggf. zu installieren und zu laden.)
6. Nutzen Sie die Funktion `plot()`, um sich die vier bekannten diagnostischen Plots zu Ihrem Regressionsmodell anzeigen zu lassen. Besteht Anlass zur Sorge?
7. Welche zwei Prädiktoren haben den stärksten Einfluss auf das kindliche Aggressionslevel?

Lösung

Teilaufgabe 1 Ich füge hier nach jedem Prädiktor einen Zeilenumbruch ein, um die Lesbarkeit zu erhöhen. Die Reihenfolge, in der man die Prädiktoren hier in das Modell schreibt, ist egal.

```
m_2 <- lm(aggression ~ television +
            computer_games +
            sibling_aggression +
            diet +
            parenting_style,
          data = child_data)
```

Teilaufgabe 2 Bei diesen Modellen sollte in der Klammer des Befehls `anova()` zuerst das weniger komplexe Modell mit weniger Prädiktoren stehen, und danach das komplexere.

```
anova(m_1, m_2)
```

```
## Analysis of Variance Table
##
## Model 1: aggression ~ parenting_style + sibling_aggression
## Model 2: aggression ~ television + computer_games + sibling_aggression +
##     diet + parenting_style
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    663 64.23
## 2    660 62.24  3      1.99 7.0339 0.0001166 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Teilaufgabe 3 Das Modell `m_2` klärt signifikant mehr Varianz in der abhängigen Variable Aggression auf, als das Modell `m_1` ($F(3, 660) = 7.0339$, $p = 0.001$). Das heißt, dass Modell `m_2` besser auf die Daten passt als `m_1`, und dass die Prädiktoren, die wir in `m_2` aufgenommen haben, zusätzliche Varianz in der abhängigen Variable aufklären. Wäre der Test nicht signifikant geworden, dann hieße das, dass wir keinen Hinweis darauf haben, dass die zusätzlichen Prädiktoren sich sinnvoll zur Vorhersage von Aggression eignen. In dem Fall sollten wir nur `m_1` interpretieren. Hier ist es aber anders, deshalb können wir in der Folge mit `m_2` weiterarbeiten und es interpretieren.

```
summary(m_2)
```

Teilaufgabe 4

```
##
## Call:
## lm(formula = aggression ~ television + computer_games + sibling_aggression +
##     diet + parenting_style, data = child_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12629 -0.15253 -0.00421  0.15222  1.17669
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.004988   0.011983  -0.416 0.677350
## television          0.032916   0.046057   0.715 0.475059
## computer_games      0.142161   0.036920   3.851 0.000129 ***
## sibling_aggression  0.081684   0.038780   2.106 0.035550 *
## diet               -0.109054   0.038076  -2.864 0.004315 **
## parenting_style     0.056648   0.014557   3.891 0.000110 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3071 on 660 degrees of freedom
## Multiple R-squared:  0.08258,    Adjusted R-squared:  0.07563
## F-statistic: 11.88 on 5 and 660 DF,  p-value: 5.025e-11
```

Alle Prädiktoren außer der Variable `television` weisen signifikante t-Tests auf, d.h. sie tragen signifikant zur Varianzaufklärung bei. Das heißt, dass zusätzlich zum Erziehungsstil und der Aggression der Geschwister wohl auch die Ernährung und das Spielen von Computerspielen zur Vorhersage von Aggression bei Kindern geeignet sein könnten. **Hier ist allerdings Vorsicht geboten:** Für die ersten beiden Prädiktoren hatten wir eine Hypothese, die wir überprüft haben (d.h. es war eine *konfirmatorische* Analyse). Hier haben wir nun eine *explorative* Analyse durchgeführt. Wir hatten keine Hypothese darüber, welche der zusätzlichen Prädiktoren einen Einfluss haben würden und in welche Richtung dieser Einfluss gehen würde. Die Ergebnisse sind deshalb noch nicht belastbar. Wir können sie lediglich nutzen, um nun daraus eine Hypothese abzuleiten, z.B. dass eine gute Ernährung das Aggressionspotential von Kindern reduzieren kann. Diese Hypothese müssten wir in der Folge mit neue Daten in einer *konfirmatorischen* Analyse überprüfen, bevor wir wirklich eine starke Aussage treffen können.

Teilaufgabe 5 Das Paket kann mit `install.packages("car")` installiert und mit `librayr(car)` geladen werden. Schreiben Sie den ersten Befehl *nur in die Konsole*, und fügen Sie den zweiten Befehl am Anfang ihrer Syntax hinzu. Sie sollten dort einen Code-Chunk haben, in dem Sie alle benötigten Pakete laden.

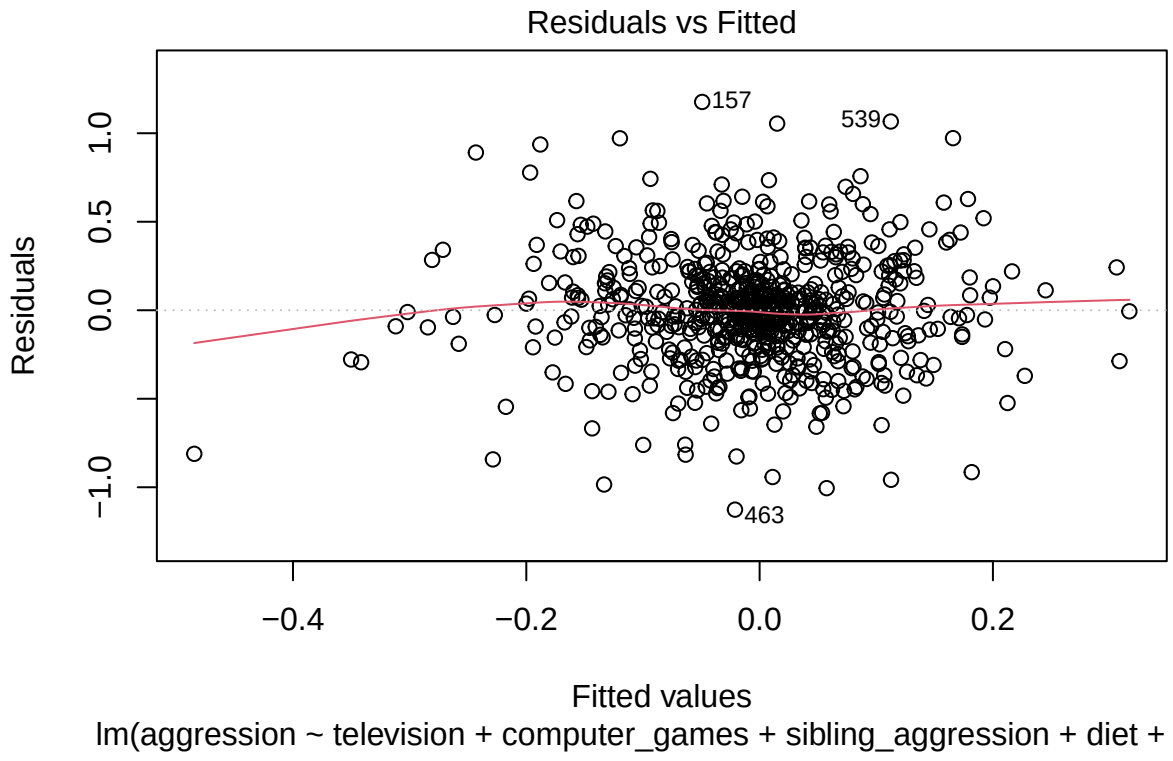
```
library(car)
```

```
vif(m_2)
```

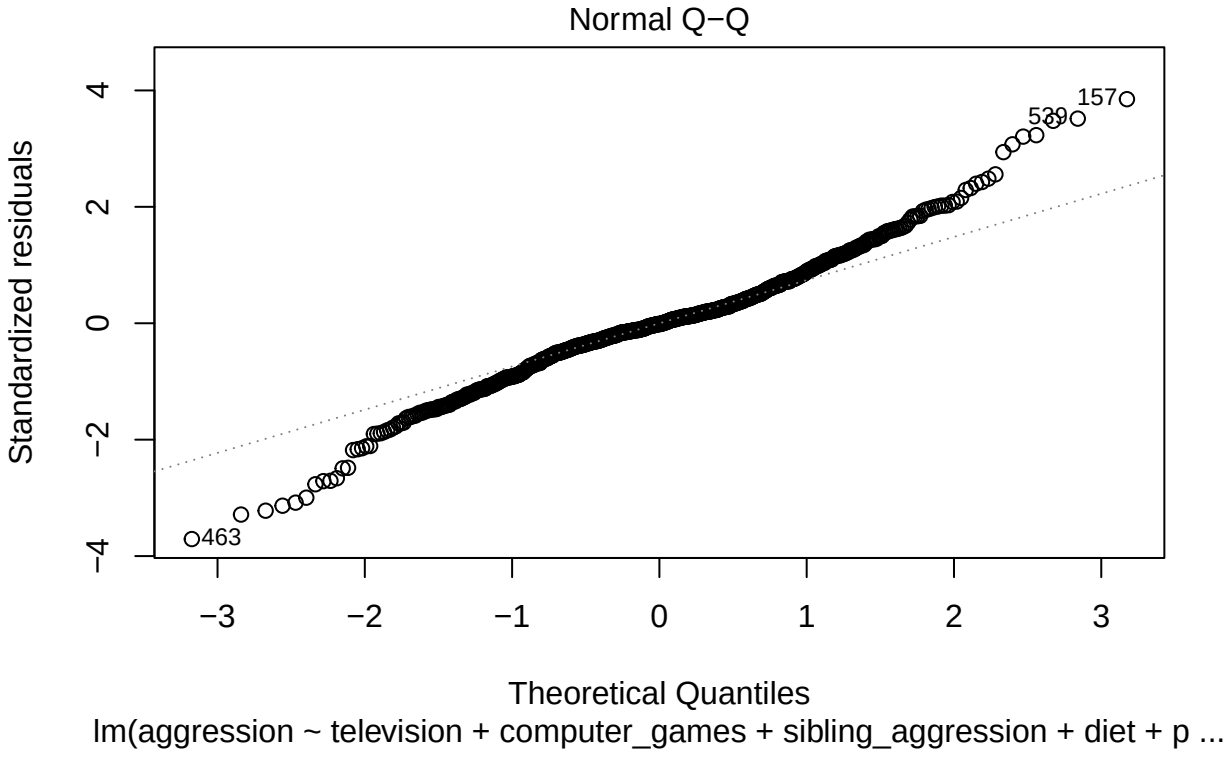
```
##         television     computer_games sibling_aggression               diet    parenting_style
##           1.435525           1.122719           1.132618           1.160466           1.494296
```

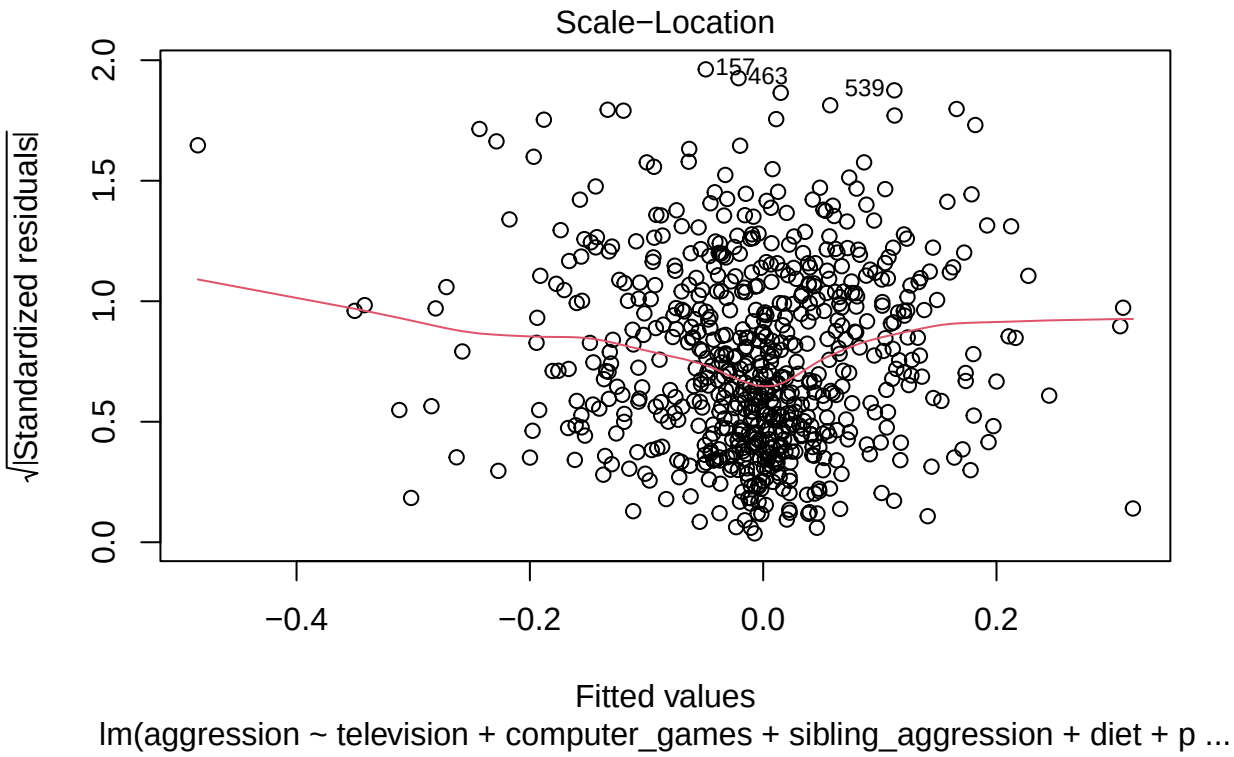
Grund zur Sorge besteht, als Faustregel, ab einem VIF von 10 bei einem Prädiktor. Bei unseren Prädiktoren ist der Wert weit davon entfernt, damit besteht kein Anlass zur Sorge.

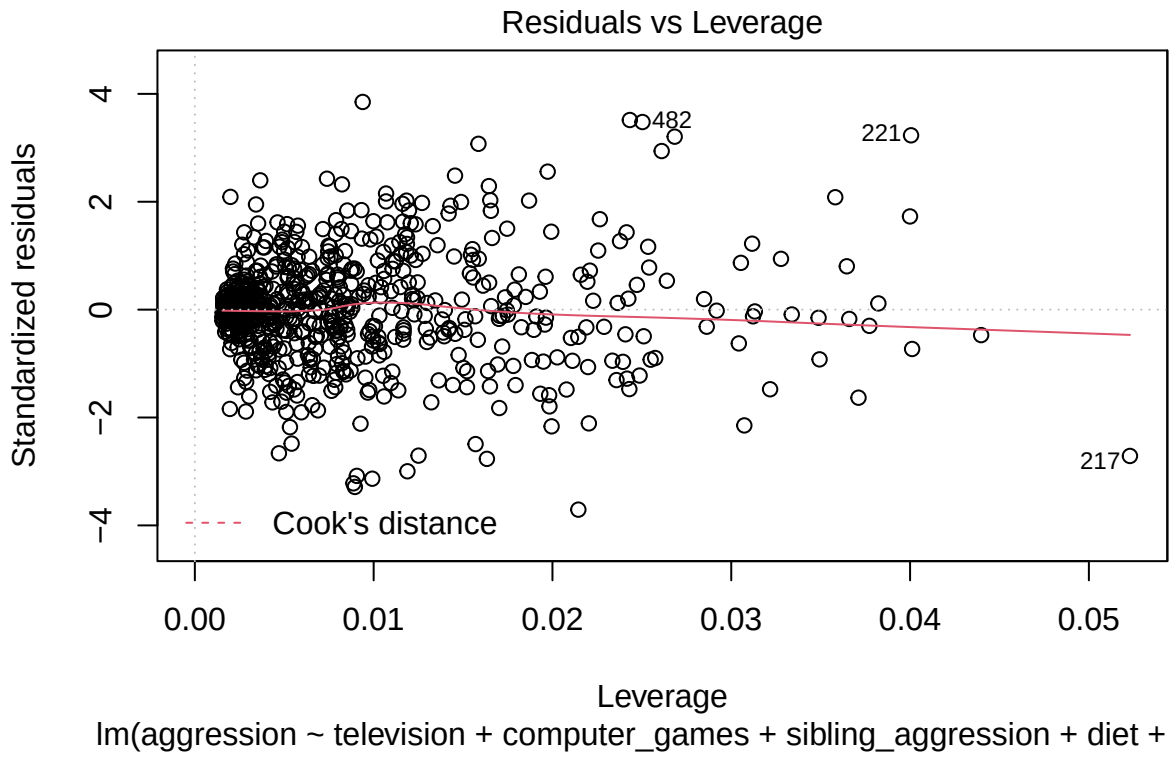
```
plot(m_2)
```



Teilaufgabe 6




Scale−Location
157
463
539
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
−0.4
−0.2
0.0
0.2
Fitted values
lm(aggression ~ television + computer_games + sibling_aggression + diet + p ...



Plot 1 zeigt eine zufällige Punktewolke, was darauf hindeutet, dass die Annahme eines linearen Zusammenhangs zwischen den Prädiktoren und der abhängigen Variable zutrifft.

Plot 2 zeigt eine annähernd gerade Linie, was darauf hindeutet, dass die Annahme normalverteilter Fehler zutrifft.

Plot 3 zeigt eine einigermaßen horizontale Linie und zufällige Punktewolke, was darauf hindeutet, dass die Annahme der Homoskedasdizität zutrifft.

Plot 4 zeigt keine Werte jenseits der gestrichelten roten Linien (wir sehen nicht einmal gestrichelte rote Linien), was darauf hindeutet, dass wir keine einzelnen Fälle mit übermäßigem Einfluss haben. Fall 217 ist allerdings in der unteren rechten Ecke, was auf einen großen *leverage*-Wert hindeutet. Diesen Fall könnte man genauer untersuchen. Das lassen wir aber hier jetzt mal sein.

Teilaufgabe 7 Um diese Frage beantworten zu können, müssen wieder die standardisierten Regressionskoeffizienten berechnet werden. Dazu verwenden wir, wie oben, die Funktion `lm.beta()` aus dem Paket `lm.beta`. Stellen Sie sicher, dass Sie das Paket installiert und geladen haben (`install.packages("lm.beta")` und `library(lm.beta)`).

```
lm.beta(m_2)
```

```
##
## Call:
## lm(formula = aggression ~ television + computer_games + sibling_aggression +
##     diet + parenting_style, data = child_data)
##
## Standardized Coefficients::
```

```
##        (Intercept)         television     computer_games sibling_aggression               diet     pa
##                 NA         0.03192490         0.15211518         0.08357717        -0.11503080
```

Die zwei Prädiktoren mit dem stärksten Einfluss auf das Aggressionslevel der Kinder sind der Erziehungsstil mit einem standardisierten $\hat{\beta}$ von 0.177 und die Intensität des Spielens von Computerspielen mit einem standardisierten $\hat{\beta}$ von 0.152.

4) Methoden der Regression

1. Lesen Sie Abschnitt 7.6.4. (Methods of regression) in *Discovering Statistics Using R* (Field, Miles & Field, 2012).
2. Welche Regressionsmethode haben wir in Aufgabe 3 angewendet?

Lösung

Unsere Herangehensweise war eine Mischung aus hierarchischem Vorgehen und Forced-Entry:

- Wir haben zwei Modelle miteinander verglichen: Eines, das unsere theoriegestützten Prädiktoren enthielt, und eines, das zusätzlich explorativ zu untersuchende Prädiktoren enthielt. Dieses Vorgehen ist hierarchisch.
- Innerhalb der beiden Modelle haben wir alle Prädiktoren, die uns interessiert haben, auf einmal in die Regression aufgenommen. Dieses Vorgehen ist forced-entry.

5) Beurteilung einer Regression

Szenario

Der Zusammenhang zwischen Wassertemperatur und Genuss beim Duschen soll untersucht werden. Die Rohdaten finden Sie hier: https://md.psych.bio.uni-goettingen.de/mv/data/div/shower_data.csv Sie sind ChefIn der Forschungsabteilung, und Ihre MitarbeiterInnen legen Ihnen die folgende Auswertung vor:

```
dd <- read_csv("https://md.psych.bio.uni-goettingen.de/mv/data/div/shower_data.csv")
```

```
## Rows: 200 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): temperatur, genuss
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
temp_model <- lm(genuss ~ temperatur, data=dd)
summary(temp_model)
```

```
##
## Call:
## lm(formula = genuss ~ temperatur, data = dd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

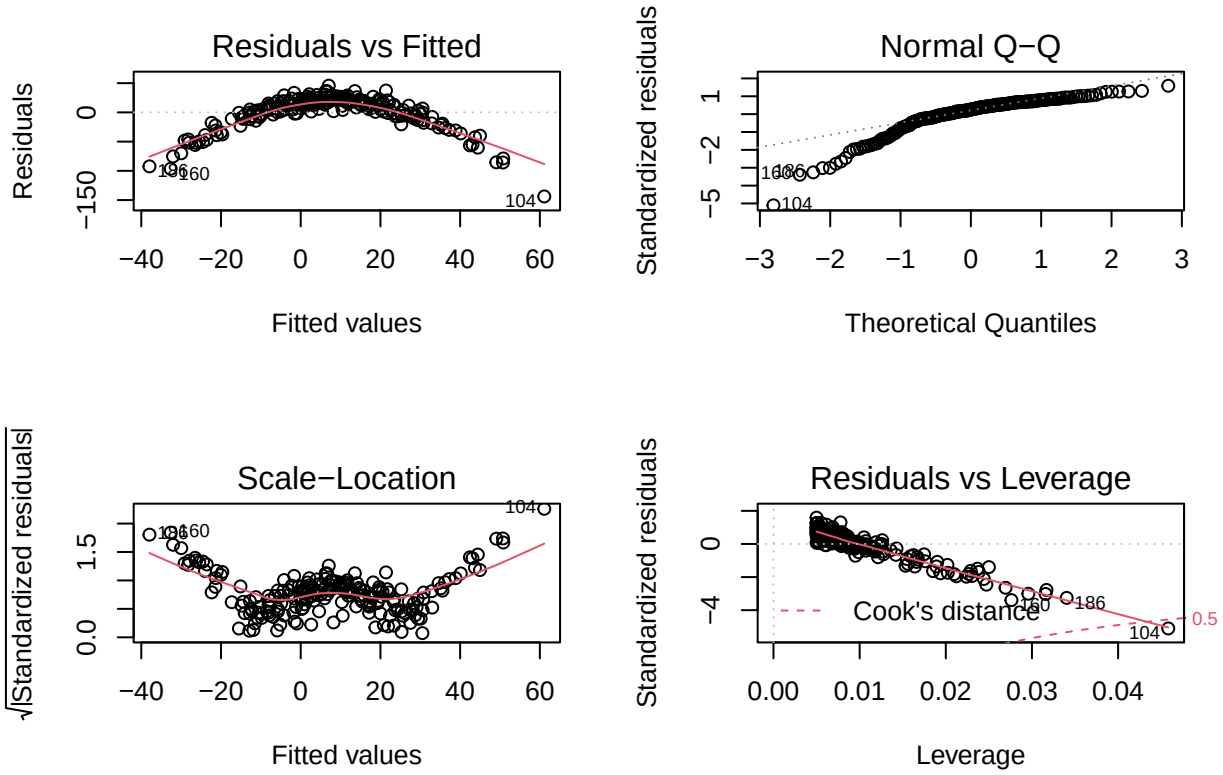
```
## -143.811   -7.291    7.326   19.296   45.700
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.7374     2.4260  -1.953   0.0523 .
## temperatur    5.7487     0.6235   9.220   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.85 on 198 degrees of freedom
## Multiple R-squared:  0.3004, Adjusted R-squared:  0.2968
## F-statistic:    85 on 1 and 198 DF,  p-value: < 2.2e-16
```

Kommentar der Mitarbeiter:

> Die Analyse ergibt eindeutig, dass der Genuss beim Duschen linear mit der Wassertemperatur ansteigt. Wir haben eine Varianzaufklärung von 0.3003517%, das Modell passt hochsignifikant auf die Daten mit $R^2 = 0.3003517$ $F(1,198) = 84.9993445$ und der Regressionskoeffizient für Temperatur ist hochsignifikant mit $t(198) = 9.2195089$; $p < 0.001$. Wir schlagen eine Pressemitteilung vor, in der wir den Menschen empfehlen, immer so heiß wie irgend möglich zu duschen, um Ihren Genuss zu maximieren.

Sie sind interessiert, denn die Ergebnisse überraschen Sie. Zur Sicherheit fordern Sie diagnostische Plots an:

```
par(mfrow = c(2,2)) # Diese Zeile sorgt für die kompakte Darstellung
plot(temp_model)
```



Aufgaben

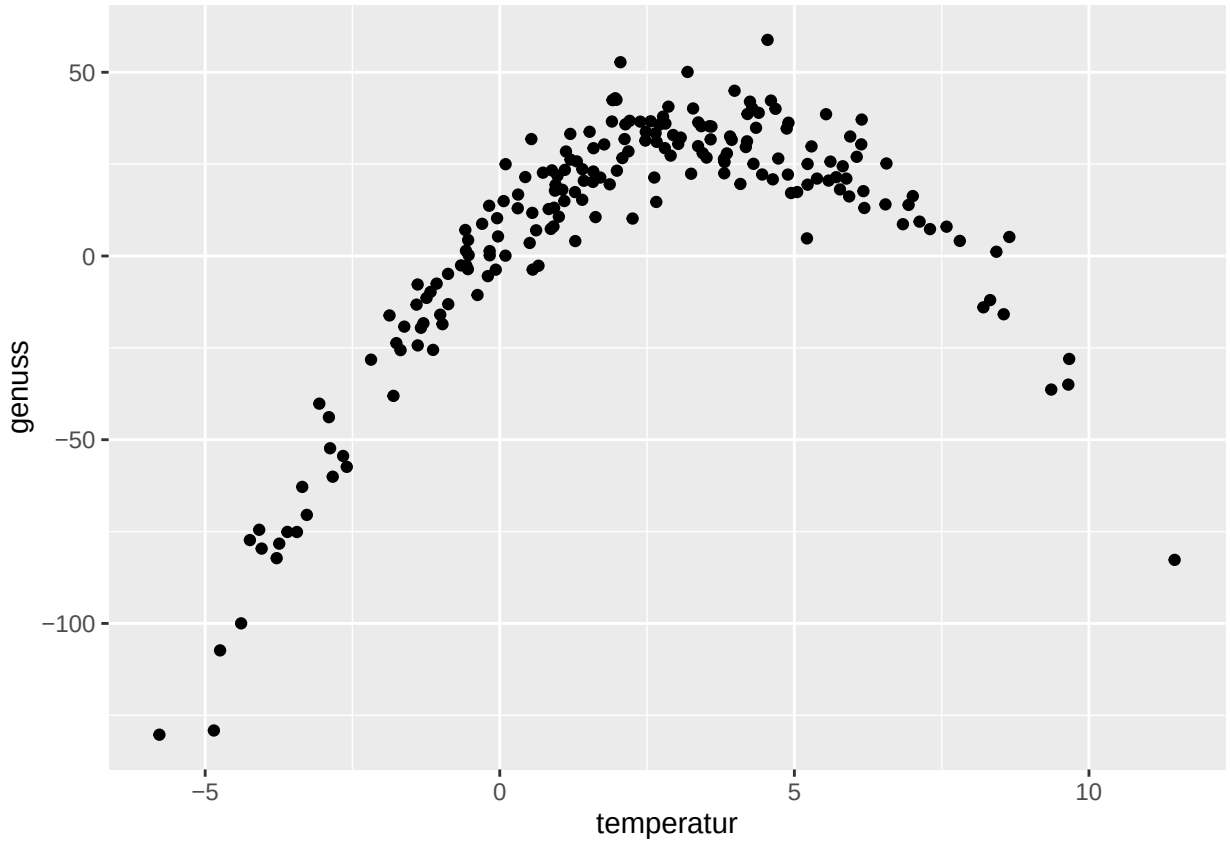
1. Beurteilen Sie, ob die von Ihren MitarbeiterInnen vorgeschlagene Analyse zulässig ist.
2. Entwickeln Sie auf Grundlage der Daten eine alternative Hypothese zum Zusammenhang zwischen Wassertemperatur und Duschgenuss. Tipp: Nützlich dafür können z.B. Plots sein.
3. Setzen Sie Ihre alternative Analysestrategie in R um und vergleichen Sie sie mit dem ursprünglichen Vorschlag.

Lösung

Unteraufgabe 1 Im ersten diagnostischen Plot zeigt sich ein deutliches Muster in Form eines Bogens, ein Hinweis darauf, dass eventuell kein linearer Zusammenhang zwischen Prädiktor und Outcome besteht. Auch die übrigen Plots geben Anlass zur Sorge. Im zweiten scheint die Abweichung von der geraden Linie relativ stark zu sein: Die Fehler sind womöglich nicht normalverteilt. Das Muster in Plot 3 deutet darauf hin, dass die Annahme der Homoskedasdizität möglicherweise verletzt ist.

Unteraufgabe 2 Die diagnostischen Plots deuten bereits darauf hin, dass eventuell ein quadratischer Zusammenhang zwischen Prädiktor und Outcome bestehen könnte. Wir schauen uns das mit einem Scatterplot genauer an:

```
temp_plot <- ggplot(dusch_data, aes(x = temperatur, y = genuss))
temp_plot + geom_point()
```



Tatsächlich deutet der Plot in die gleiche Richtung, wir haben einen umgekehrt U-förmigen Zusammenhang zwischen der Wassertemperatur und dem Genuss beim Duschen: Sowohl zu niedrige als auch zu hohe Temperatur schmälern den Genuss, angenehm scheint vor allem eine mittlere Temperatur zu sein.

Unteraufgabe 3 Wir können den quadratischen Zusammenhang in unser Regressionsmodell mit aufnehmen. Dafür schreiben wir das Modell wie folgt:

```
temp_model_2 <- lm(genuss ~ temperatur + I(temperatur^2), data = dusch_data)
```

In einem Regressionsmodell werden Rechenoperationen, die innerhalb von `I()` angegeben werden, direkt ausgeführt. So können wir das Quadrat von `temperatur` direkt im Code spezifizieren. Wir hätten das ganze aber auch manuell machen können:

```
temperatur_sq <- temperatur^2
# to understand the background better, we fit also a model with squared temperature only
temp_model_1 <- lm(genuss ~ temperatur_sq, data = dusch_data)
# the model with linear and quadratic explanatory variable
temp_model_2 <- lm(genuss ~ temperatur + temperatur_sq, data = dusch_data)
```

Wir vergleichen nun die Modelle mit `anova()`

```
anova(temp_model, temp_model_2)
```

```
## Analysis of Variance Table
##
## Model 1: genuss ~ temperatur
## Model 2: genuss ~ temperatur + temperatur_sq
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    198 164781
## 2    197  14380  1    150401 2060.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(temp_model, temp_model_1)
```

```
## Analysis of Variance Table
##
## Model 1: genuss ~ temperatur
## Model 2: genuss ~ temperatur_sq
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1    198 164781
## 2    198 229733  0    -64952
```

Das zweite Modell klärt signifikant mehr Varianz auf, als das erste. Dies ist das Modell, das wir interpretieren sollten. Wir lassen uns den Output und die standardisierten Koeffizienten anzeigen.

```
summary(temp_model_2) # Output
```

```
##
## Call:
```

```
## lm(formula = genuss ~ temperatur + temperatur_sq, data = dusch_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.2866  -5.7581  -0.4822   5.7685  26.7778
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.19091    0.75776    8.17 3.66e-14 ***
## temperatur    14.50442    0.26704   54.32  < 2e-16 ***
## temperatur_sq -1.94005    0.04274  -45.39  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.544 on 197 degrees of freedom
## Multiple R-squared:  0.9389, Adjusted R-squared:  0.9383
## F-statistic:  1515 on 2 and 197 DF,  p-value: < 2.2e-16
```

```
lm.beta(temp_model_2) # Standardisierte Koeffizienten
```

```
##
## Call:
## lm(formula = genuss ~ temperatur + temperatur_sq, data = dusch_data)
##
## Standardized Coefficients::
##   (Intercept)    temperatur temperatur_sq
##            NA      1.382751     -1.155564
```

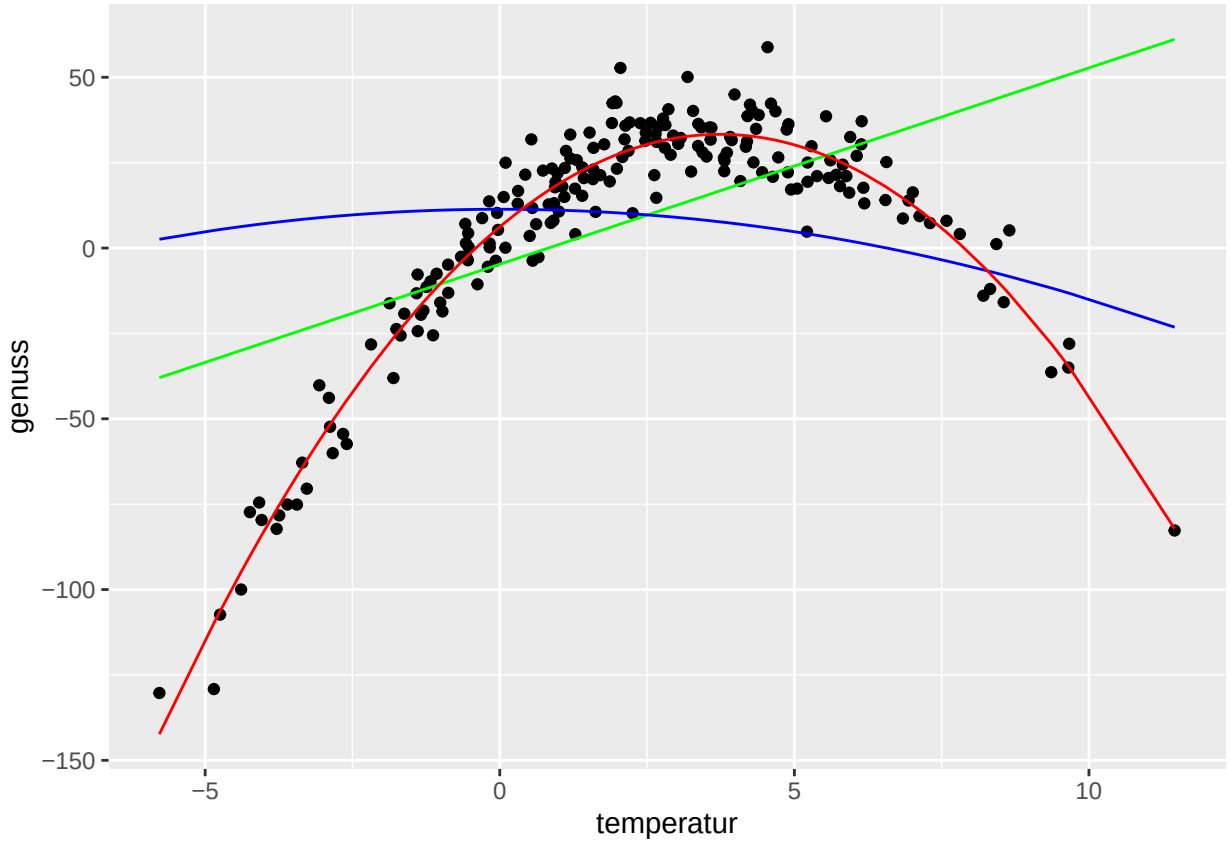
```
lm.beta(temp_model_1) # Standardisierte Koeffizienten, quadratischer Term alleine
```

```
##
## Call:
## lm(formula = genuss ~ temperatur_sq, data = dusch_data)
##
## Standardized Coefficients::
##   (Intercept) temperatur_sq
##            NA    -0.1567515
```

Der Blick auf den Output verrät uns, dass wir hier tatsächlich gut 94% der Varianz im Dusch-Genuss aufklären können. Das Modell passt signifikant auf die Daten mit $R^2 = .94$, $F(2, 197) = 1384$, $p < .001$ und sowohl der lineare ($\beta = 1.38$, $t(197) = 52.07$, $p < .001$) als auch der quadratische Koeffizient ($\beta = -1.14$, $t(197) = -42.80$, $p < .001$) sind signifikant verschieden von 0.

Auch ein Blick auf die Grafik verrät, dass die Kombination des linearen mit dem quadratischen Terms die bestangepasste Vorhersagelinie (rot) erzeugen.

```
temp_plot <- ggplot(dusch_data, aes(x = temperatur, y = genuss))
temp_plot +
  geom_point() + geom_line(aes(y=predict(temp_model), group=1), color="green") +
  geom_line(aes(y=predict(temp_model_1), group=1), color = "blue") +
  geom_line(aes(y=predict(temp_model_2), group=1), color = "red")
```



Literatur

Anmerkung: Diese Übungszettel basieren zum Teil auf Aufgaben aus dem Lehrbuch *Dicovering Statistics Using R* (Field, Miles & Field, 2012). Sie wurden für den Zweck dieser Übung modifiziert, und der verwendete R-Code wurde aktualisiert.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. London: SAGE Publications Ltd.

English

Links

Exercise sheet in PDF

Exercise sheet with solutions included

Exercise sheet with solutions included as PDF

The source code of this sheet as .Rmd (Right click and "store as" to download ...)

Some hints

1. Please give your answers in a .Rmd file. You may generate one from scratch using the file menu: 'File > new file > R Markdown ...' Delete the text below *Setup Chunk* (starting from line 11). Alternatively you may use this sample Rmd by downloading it.

2. You may find the informations useful that you can find on the start page of this course.
3. Don't hesitate to google for solutions. Effective web searches to find solutions for R-problems is a very useful ability, professionals to that too ... A really good starting point might be the R area of the programmers platform Stackoverflow
4. You can find very useful cheat sheets for various R-related topics. A good starting point is the Base R Cheat Sheet.

Ressources

This is a hands on course. We cannot present you all the useful commands in detail. Instead we give you links to useful ressources, where you might find hints to help you with the exercises.

Ressource	Description
Field, Chapter 7	Book chapter with a step for step introduction to regression, what that is all about and, how to do it in R. **Recommendation!**
Peters Multiple Regression Pages	Peters unit on multiple regression. A resource to find running examples.

Tip of the week

This week, we have two tips:

Commenting out a block of code

The shortcut `strg` + `shift` + `c` (Windows) or. `cmd` + `shift` + `c` (Mac) comments out the currently marked code lines. That way you can comment out several lines at a time inside a chunk.

No more mess while quoting

You may have noticed already, that R inserts automatically two quote charaters if we press `"`. This is sometimes cool, but sometimes not, f. e. if we want to quote a word already in the source code. We find us frequently in a situation, where we have too much quote characters, like: `"Text""`. But there is a trick:

1. Mark the word you want to quote.
2. After marking press `"`. The word will be correctly quoted automatically.

The same works for all types of brackets!

1) Read data

1. Define an appropriate working directory for this exercise sheet. This should usually be the folder, where your Rmd-file is located. But be careful: The render process always assumes that your working directory is the directory, your Rmd-file is in. This is expecially important if you work with relative links.
2. Assure, that the packages of `tidyverse` are loaded. Insert a code line for that in the beginning of your Rmd-file.

3. Read datafile `child_aggression.csv` directly from URL `https://md.psych.bio.uni-goettingen.de/mv/data/div/c` using the command `read_csv()`.
4. Use the command `write_csv()` to store this datafile in your current working directory in your subdirectory for data.
5. If you have problems reading from URL you can find and download the data using this link You may then read the data locally.

Something about the data

An overview of the variables in the data that were collected from 666 children.

Variable	Meaning
aggression	The higher the value, the more aggression is shown by the child.
television	The higher the value, the more time the child spents in front of the TV.
computer_games	Higher values indicate more time spnt playing computer games.
sibling_aggression	Higher values indicate higher aggression shown by older siblings.
diet	High values indicate more salutable nutrition of the child.
parenting_style	Higher values indicate more dysfunctional education style of the parents.

Solutions

Subtask 1 Please follow the recommendation of the exercise text.

```
library(tidyverse)
```

Subtask 2

```
child_data <- read_csv("https://md.psych.bio.uni-goettingen.de/mv/data/div/child_aggression.csv")
```

Subtask 3

```
## Rows: 666 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (6): aggression, television, computer_games, sibling_aggression, diet, parenting_style
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
write_csv(child_data, "data/child_aggression.csv")
```

Subtask 4

```
child_data <- read_csv("data/child_aggression.csv")
```

Subtask 5

2) Multiple Regression: Test Hypotheses

Based on previous research you have the hypothesis, that educational style and aggressivity values of the siblings are good predictors for the aggression level of children.

1. Specify a regression model, on base of which you predict aggression by combining educational style and aggression level of siblings.
2. Take a look at the output and answer the following questions:
    a) Which part of the variance of the aggression values of children is explained by the model?
    b) Is the prediction quality superior to the prediction of the null model?
    c) Do both predictor variables contribute significantly to the prediction quality?
    d) Write down the estimated regression model
    e) Calculate the aggressin level of a child on base of our model, that has educational style `1` and siblings aggression of `0.5`.
3. Now let's see, which predictor is the one with the higher influence on the aggression value prediction. (Tip: Take a look in Andy Fields book, if you don't know how to do that.)
    a) We cannot see the strength of the influence of a predictor from its p-value. Why not? (Tip: What says a p-value?)
    b) The estimate for a given regression coefficient is not directly comparable with the others. Explain why not?
    c) Install the package `lm.beta()` if it isn't installed already. Insert a code line in your .Rmd that assures, that `lm.beta()` is loaded.
    d) Apply `lm.beta` to the regression model from exercise 1.
    e) What does a standardized regression coefficient of 1 mean?
    f) Which of our predictors has the stronger impact on child aggression?

Solutions

```
m_1 <- lm(aggression ~ parenting_style + sibling_aggression,
          data = child_data)
```

Subtask 1

```
summary(m_1)
```

Subtask 2

```
##
## Call:
## lm(formula = aggression ~ parenting_style + sibling_aggression,
##     data = child_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.09755 -0.17180  0.00092  0.15405  1.23037
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.005784   0.012065  -0.479    0.632
## parenting_style     0.061984   0.012257   5.057 5.51e-07 ***
## sibling_aggression  0.093409   0.037505   2.491    0.013 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3113 on 663 degrees of freedom
## Multiple R-squared:  0.05325,    Adjusted R-squared:  0.05039
## F-statistic: 18.64 on 2 and 663 DF,  p-value: 1.325e-08
```

a) The important point here is the multiple R^2. This is .053, which would be 5.3% of explained variance.
b) Yes! The model predicts aggression values significantly better than the null model, with $F(2, 663) = 18.64$ and $p < .001$. We find this at the end of the output under 'F-Test'.
c) Yes, both t-tests for the predictors are significant.
    i) Education style is significant with $t(663) = 5.06$ and $p < .001$
    ii) Siblings aggression is significant with $t(663) = 2.49$ and $p = .013$
d) The estimated regression model is

$$\hat{y}_i = -0.006 + 0.062 \cdot parenting\_style + 0.093 \cdot sibling\_aggression$$

e) We simply use the values 1 und 0.5 at the corresponding positions in the equation

$$-0.006 + 0.062 \cdot 1 + 0.093 \cdot 0.5 = 0.1025$$

So we predict an aggression level of 0.1025.

Subtask 3

a) The p-value tells us only, whether the regression coefficient is different from 0. In other words: We get an answer to the question, whether there is an influence of the predictor at all. From that we cannot infer any idea about the strength of the influence of an predictor.
b) In a simple regression model the regression coefficient not only depends on the strength of its influence, but also on the measuring unit of the predictor. If we want to predict weight by height f. e. the coefficient varies considerably depending on whether height is measured in cm or in m, although the influence is always the same. But we can standardize the coefficients to make them comparable. More about that in section 7.8.3.2. (Model parameters) in *Discovering Statistics Using R* (Field, Miles & Field, 2012).
c) The code to install ´lm.beta`would be:`install.packages("lm.beta")‘. Please install packages only via console, not in a (Rmd) script. This can lead to problems when we render the Rmd file. The code to load would be:

```
# library(lm.beta)
# or better
require(lm.beta)
```

d) Code:

```
lm.beta(m_1)
```

```
##
## Call:
## lm(formula = aggression ~ parenting_style + sibling_aggression,
##     data = child_data)
##
## Standardized Coefficients::
##        (Intercept)    parenting_style sibling_aggression
##                 NA         0.19406149         0.09557412
```

e) A standardized regression coefficient of 1 means, that a change of 1 sd in the predictor would cause a change of 1 sd in the outcome.

f) From the output in d) we can see: The parenting styles influence on the outcome is double as big as the influence of siblings aggression level.

3) Multiple Regression: Exploration

We have more variables in our dataset that could serve as predictors. But we don't have any hypotheses about theyr effect on the outcome. So we only want to see, if we can detect some hint of an influence.

1. Specify a regression model, where aggression level of our children is predicted by all potencial predictors in our dataset (incl. parenting style and siblings aggression).
2. Use the command `anova()` to do a R^2 difference test to compare the two models.
3. Interprete the output of `anova()`. What can we conclude?
4. Check the `summary()` of the better model of exercise 3. Which predictors contribute significantly to the explanation of variance in the outcome?
5. Use `vif()` of package `car` to check multicollinearity of the predictors of our model. Inspect the *Variance Inflation Factor* for each predictor. Should we be preoccupied about vif? (Hint: Don't forget to eventually install or load the package).
6. Use function `plot()` to get the well known diagnostic plots for our regression model. Any hint for problems?
7. Which are the two predictors with the highest influence on child aggression?

Solutions

Subtask 1 We put every predictor in its own line to aument readibility. The sequence in which we specify the predictors doesn't matter.

```
m_2 <- lm(aggression ~ television +
            computer_games +
            sibling_aggression +
            diet +
            parenting_style,
          data = child_data)
```

Subtask 2 The parameters of `anova()` are the models to compare, ordered by the number of parameters they use. The most simple model is the first. Take care, the models have to be hierarchical.

```
anova(m_1, m_2)
```

```
## Analysis of Variance Table
##
## Model 1: aggression ~ parenting_style + sibling_aggression
## Model 2: aggression ~ television + computer_games + sibling_aggression +
##     diet + parenting_style
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    663 64.23
## 2    660 62.24  3      1.99 7.0339 0.0001166 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Subtask 3 Our model `m_2` explains significantly more variance in outcome child aggression than model `m_1` ($F(3, 660) = 7.0339$, $p = 0.001$). This means, that model `m_2` fits the data better than `m_1`. This means also, that the predictors, we included in `m_2` explain additional variance in our outcome variable. Wouldn't our anova not show significant differences, the inclusion of our additional variables wouldn't make sense to explain more variance of the dependent variable. In this case we should only interprete `m_1`. But `anova` tells us that it makes sense to go on with model `m_2` and interprete it.

```
summary(m_2)
```

Subtask 4

```
##
## Call:
## lm(formula = aggression ~ television + computer_games + sibling_aggression +
##     diet + parenting_style, data = child_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12629 -0.15253 -0.00421  0.15222  1.17669
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.004988   0.011983  -0.416 0.677350
## television          0.032916   0.046057   0.715 0.475059
## computer_games      0.142161   0.036920   3.851 0.000129 ***
## sibling_aggression  0.081684   0.038780   2.106 0.035550 *
## diet               -0.109054   0.038076  -2.864 0.004315 **
## parenting_style     0.056648   0.014557   3.891 0.000110 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3071 on 660 degrees of freedom
## Multiple R-squared:  0.08258,    Adjusted R-squared:  0.07563
## F-statistic: 11.88 on 5 and 660 DF,  p-value: 5.025e-11
```

All predictors but `television` show significant t-tests. So all of them increase predictive power of child aggression. This means, that above parenting style and siblings aggression diet and computer games could contribute to the prediction of aggression. **But take care:** We had a hypothesis for the first two predictors, that we checked. So we were doing a *confirmatory* analysis. But now we did a *explorative* analysis. We didn't have a hypothesis on which additional predictors could have an effect and in which way could be the influence. Therefore our results are not reliable yet. We can only use them to deduct new hypotheses, f. e. that good diet reduces child aggression. We could test this new hypothesis by collecting new data for an *confirmatory* analysis and thus come to stronger conclusions.

Subtask 5 We can install the package using `install.packages("car")` and activate it with `require(car)` or `library(car)` You should use the command `install.packages()` in the console only. Of course you may put the second command in an script to automize the activation of the needed packages. It is good practice to have a chunk at the beginning of a Rmd file where we load all the needed packages of the rest of the file.

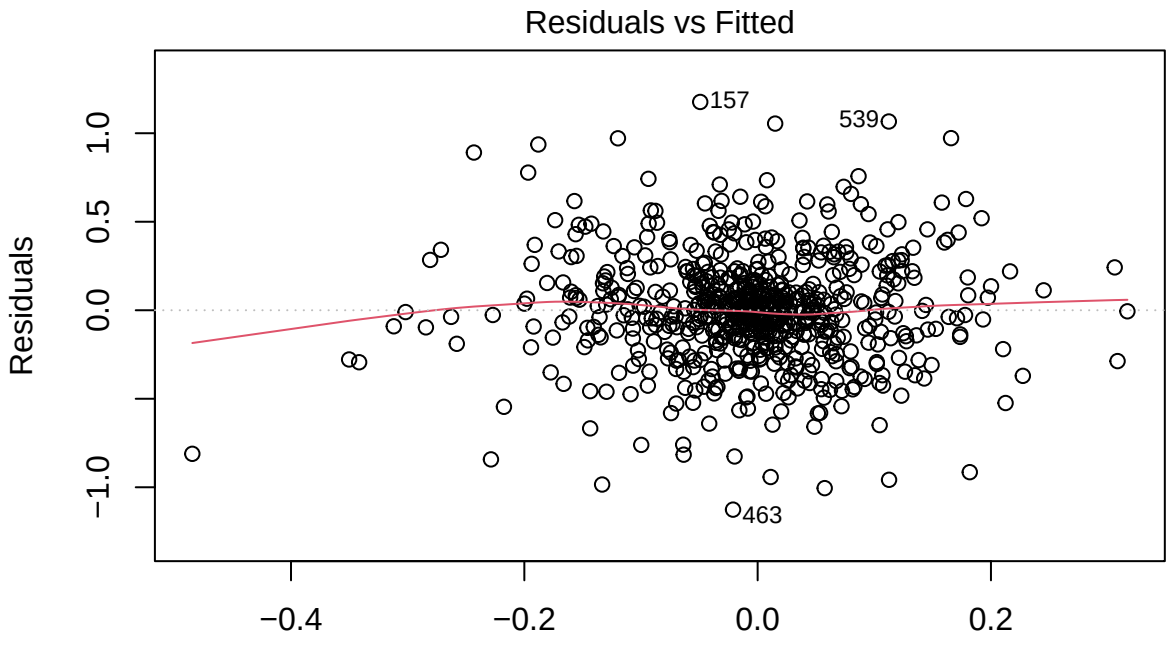
```
# library(car)
# or
require(car)
```

```
vif(m_2)
```

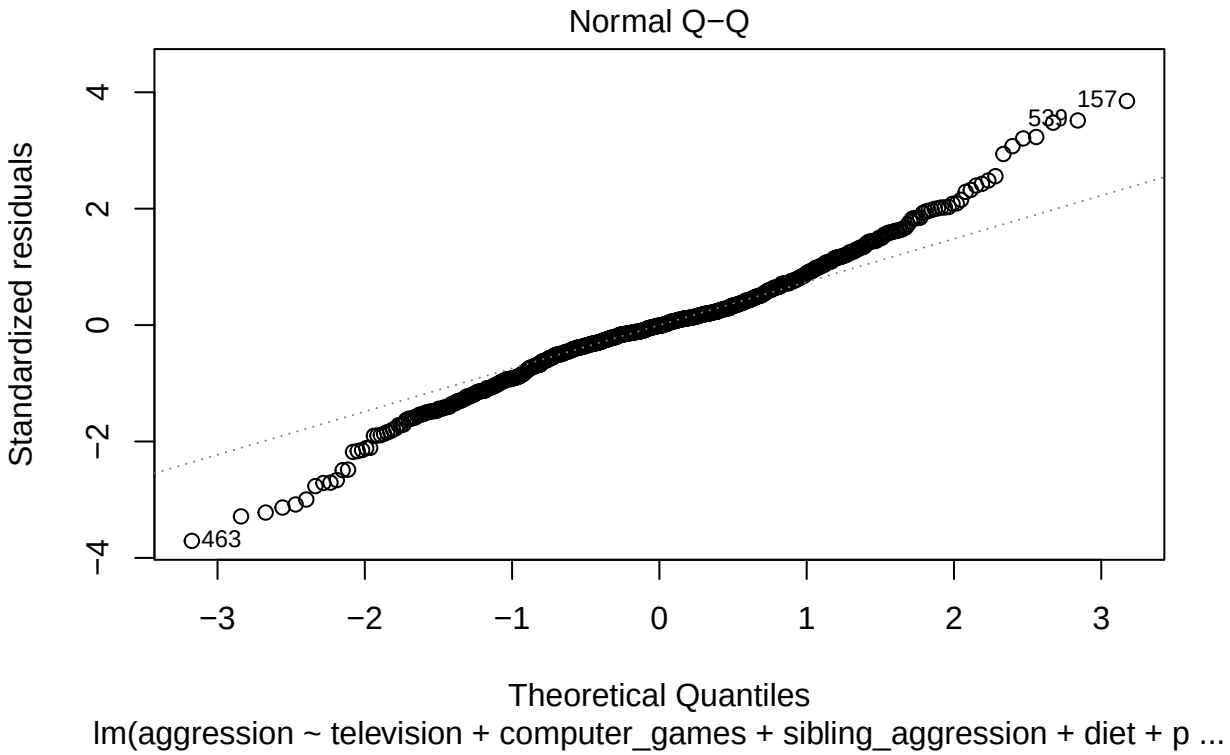
```
##         television     computer_games sibling_aggression               diet    parenting_style
##           1.435525           1.122719           1.132618           1.160466           1.494296
```

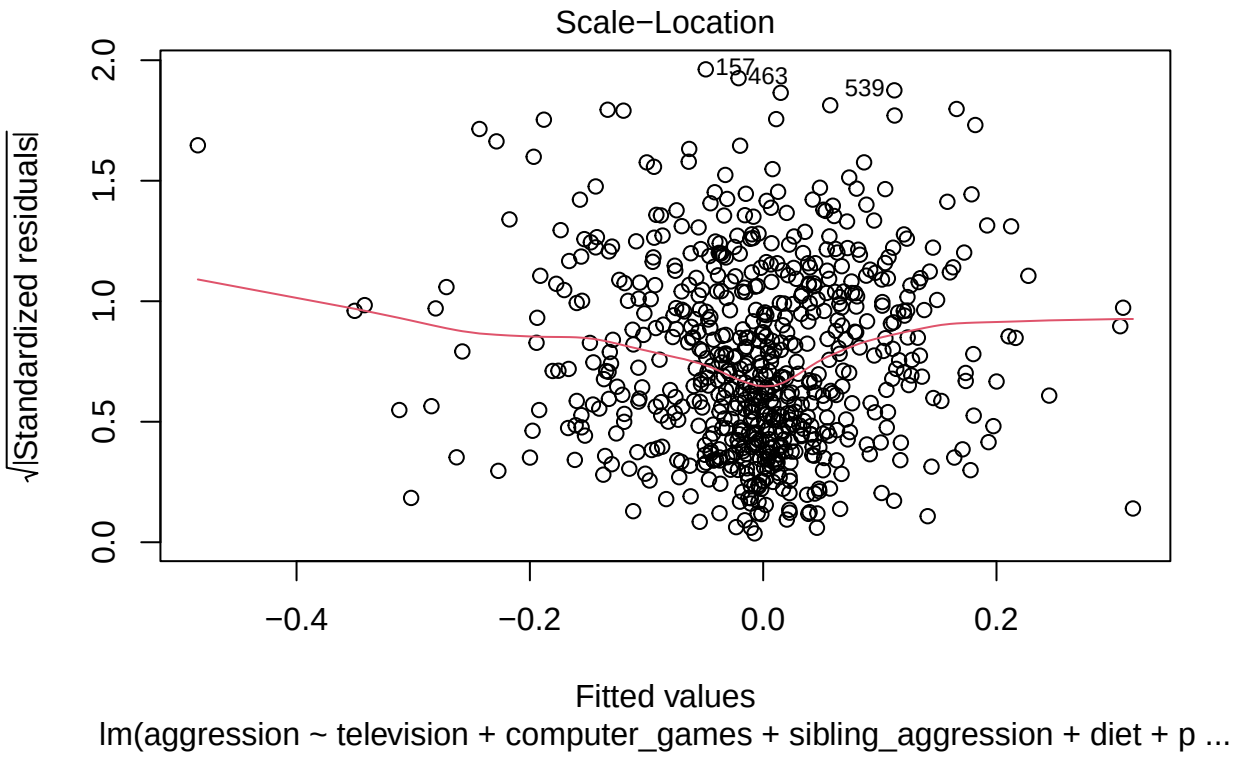
As a rule of thumb VIF values higher than 10 are suspicious. In our predictors the vif are way lower. So no need to be alarmed.

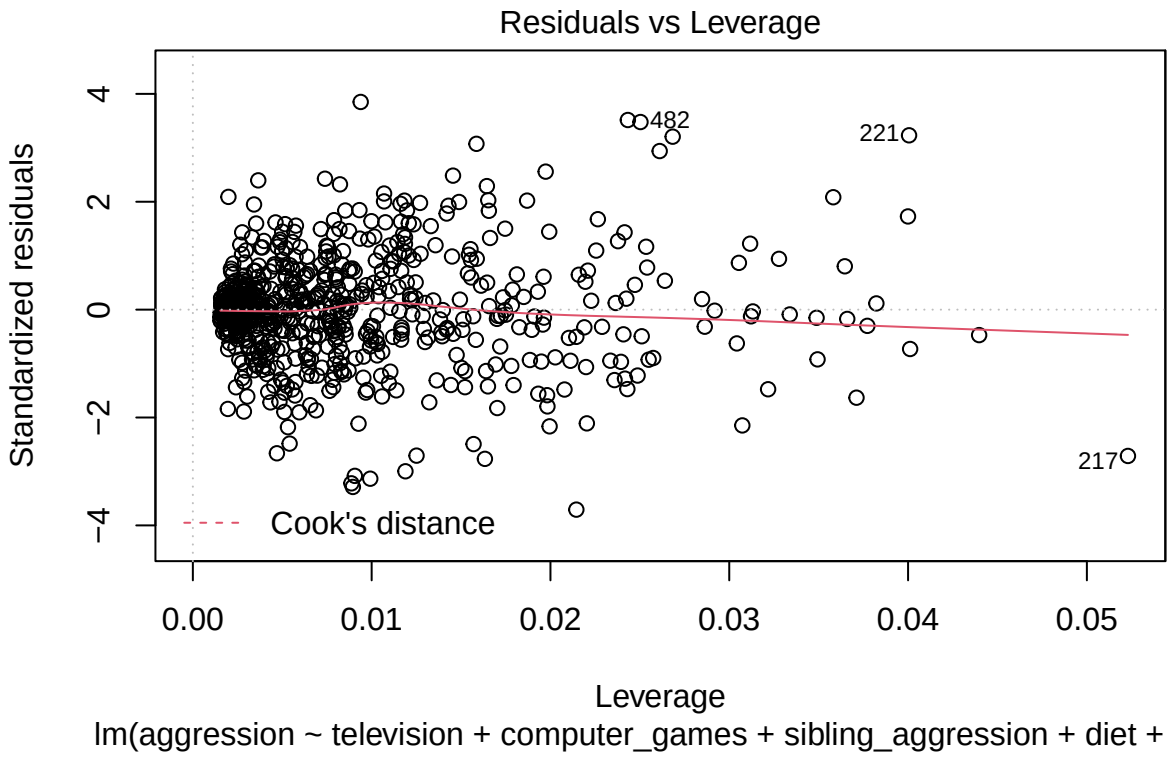
```
plot(m_2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(aggression ~ television + computer_games + sibling_aggression + diet + p ...
157
539
463
1.0
0.5
0.0
−1.0
−0.4
−0.2
0.0
0.2

Subtask 6


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(aggression ~ television + computer_games + sibling_aggression + diet + p ...
157
539
463
4
2
0
−2
−4
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
157
463
539
−0.4
−0.2
0.0
0.2
Fitted values
lm(aggression ~ television + computer_games + sibling_aggression + diet + p ...



Plot 1 shows a random distribution of our scattered points. We interprete this as evidence for a linear relation between predictors and dependent variable.

Plot 2 shows an almost straight line. This means, that we should have normally distributed errors.

Plot 3 shows an mostly horizontal line and randomly scattered points around it. We interprete this as homoskedacity.

Plot 4 shows no values outside the dashed red line. We do not even see the dashed red line. So we conclude, that we do not have too influencial cases in our sample. But case 217 is in the lowe right corner, indicating a relatively large *leverage* value. We should have a closer look at this case. But not for now ;-)

Subtask 7 We need to calculate the standardized regression coefficients to answer this question. So we run the function `lm.beta()` of package `lm.beta`. We have to assure, that this package is installed (`install.packages("lm.beta")`) and loaded (`require(lm.beta)`).

```
lm.beta(m_2)
```

```
##
## Call:
## lm(formula = aggression ~ television + computer_games + sibling_aggression +
##     diet + parenting_style, data = child_data)
##
## Standardized Coefficients::
##        (Intercept)         television     computer_games sibling_aggression               diet     pa
##                 NA         0.03192490         0.15211518         0.08357717        -0.11503080
```

`parenting style` and `computer_games` are the two variables with the strongest influence on child agression. Their standardized betas are $\hat{\beta}$ 0.177 and $\hat{\beta}$ von 0.152 respectively.

4) Methods of Regression

1. Read chapter 7.6.4 (Methods of regression) in *Discovering Statistics Using R* (Field, Miles & Field, 2012).
2. What is the regression method we used in exercise 3?

Solution

The approach we used above was a mixture of hierachical and forced entry.

- We compared two models: One with the predictors we had hypotheses with and one that had additional predictors that we entered in an exploratory way. This approach is hierarchical.
- With both of our models we entered all variables of interest at once. This is forced entry.

5) Evaluation of Regressions

Szenario

Imagine we want to find out the relation between water temperature and enjoyment while taking a shower. You can find the raw data at: https://https://md.psych.bio.uni-goettingen.de/mv/data/div/dusch_data.csv
You are chief of the research department and your employees bring you the following calculation:

```
shower_data <- readr::read_csv("https://md.psych.bio.uni-goettingen.de/mv/data/div/shower_data.csv")
```

```
## Rows: 200 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): temperatur, genuss
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
enjoyment <- shower_data$genuss
temperature <- shower_data$temperatur
temp_model <- lm(enjoyment ~ temperature)
summary(temp_model)
```

```
##
## Call:
## lm(formula = enjoyment ~ temperature)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -143.811   -7.291    7.326   19.296   45.700
##
## Coefficients:
```

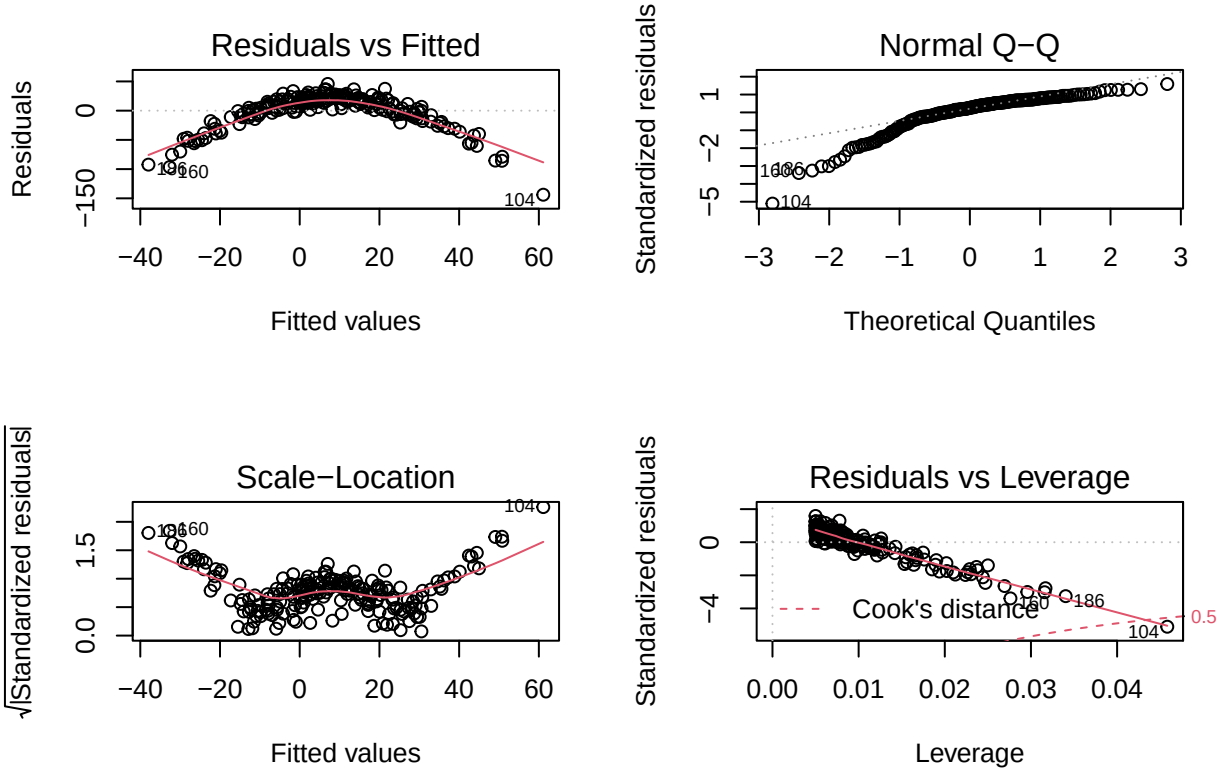
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.7374     2.4260  -1.953   0.0523 .
## temperature   5.7487     0.6235   9.220   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.85 on 198 degrees of freedom
## Multiple R-squared:  0.3004, Adjusted R-squared:  0.2968
## F-statistic:    85 on 1 and 198 DF,  p-value: < 2.2e-16
```

Comment of an employee:

> The results of the analysis show clear evidence, that enjoyment has a linear relation to water temperature. The model can explain $R^2 = 0.3003517\%$, of the variance of enjoyment and the model fits the data highly significant with $R^2 = 0.3003517$ $F(1,198) = 84.9993445$ and the regression coefficient is also highly significabt with $t(198) = 9.2195089$; $p < 0.001$. We propose a press anouncement in which we recommend people to take a shower as hot as possible to maximize their enjoyment.

You are interested, because the results are surprising for you. To get more into detail, you ask for the diagnostic plots.

```
par(mfrow = c(2,2)) # This line results in a compacter presentation
plot(temp_model)
```



Questions

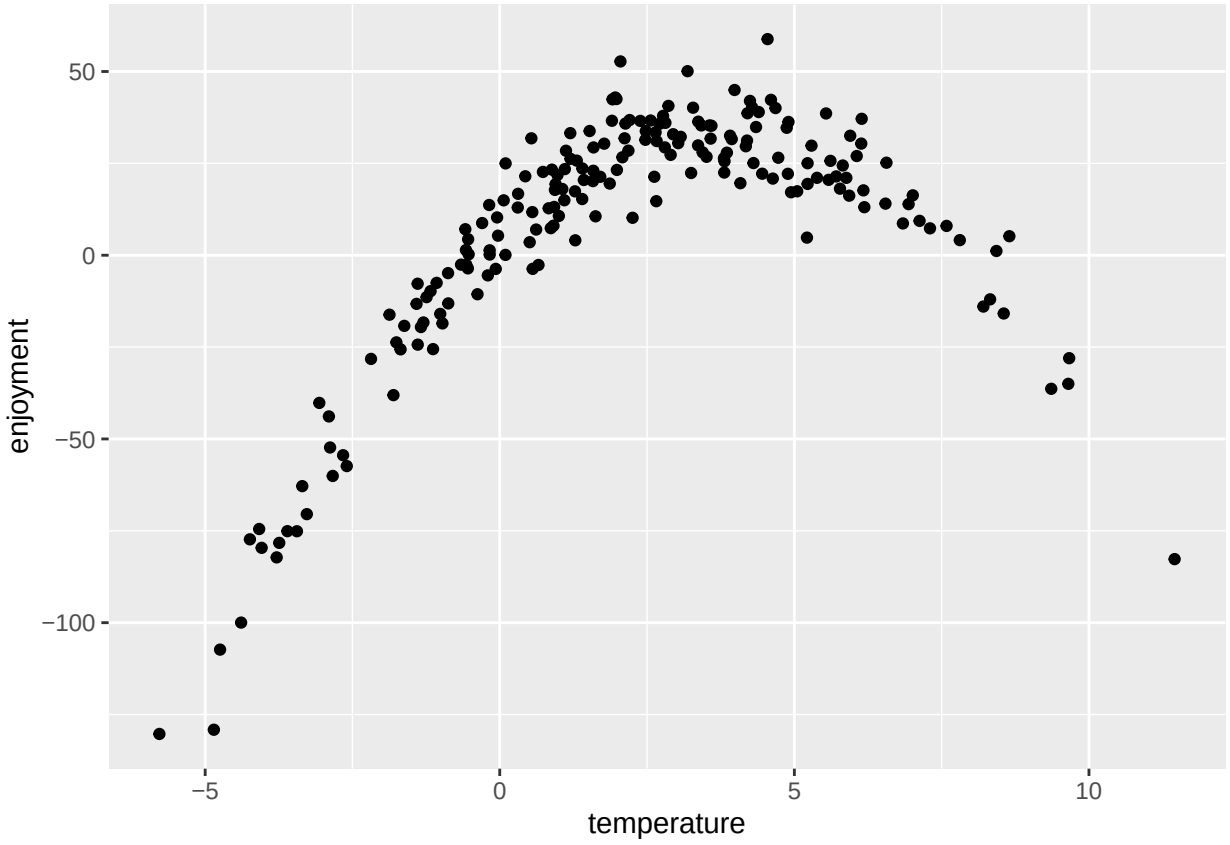
1. Make up your mind: Is the anaysis of your employee o.k.?
2. Formulate a new alternative hypothesis about the relation between water temperature and enjoyment while taking a shower. Tip: The diagnostic plots might be useful.
3. Run your alternative in R and compare your results with the original suggestion of your employee.

Solution

Subtask 1 In the first diagnostic plot we can see a clear pattern with the form of an arc, this indicates that there is no linear relation between predictor and outcome. The rest of the plots are also somewhat alarming. In the second one the deviation of the line from a straight one seems to be considerable, so our errors don't seem to be normally distributed. Moreover the pattern of plot 3 seems to indicate a violation of homoskedacity.

Subtask 2 The diagnostic plots indicate an quadratic relation between predictor and outcome. We take a closer look at that by plotting the data:

```
temp_plot <- ggplot(shower_data, aes(x = temperature, y = enjoyment))
temp_plot + geom_point()
```



Indeed, we seem to have a U-shaped relation between water temperature and enjoyment while taking a shower: Too low as well as too high temperatures seem to lower enjoyment, the most enjoyable temperatures seem to be the ones in between.

Subtask 3 We can include a quadratic term in our regression model. Therefore we specify our model like this:

```
temp_model_2 <- lm(enjoyment ~ temperature + I(temperature^2), data = shower_data)
```

If we specify `I()` in a regression model, the calculations are made before the results are used to fit the model. Thus we can specify the squared `temperature` directly in the code of the linear model formula. Of course we could have done that manually:

```
temperature_sq <- temperature^2
temp_model_2 <- lm(enjoyment ~ temperature + temperature_sq, data = shower_data)
```

We compare the models using `anova()`

```
anova(temp_model, temp_model_2)
```

```
## Analysis of Variance Table
##
## Model 1: enjoyment ~ temperature
## Model 2: enjoyment ~ temperature + temperature_sq
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    198 164781
## 2    197  14380  1    150401 2060.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The second model can explain more variance as the first one. So this is the model, we should interprete. We get the output and the standardized coefficients.

```
summary(temp_model_2) # Output
```

```
##
## Call:
## lm(formula = enjoyment ~ temperature + temperature_sq, data = shower_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.2866  -5.7581  -0.4822   5.7685  26.7778
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.19091    0.75776    8.17 3.66e-14 ***
## temperature    14.50442    0.26704   54.32  < 2e-16 ***
## temperature_sq -1.94005    0.04274  -45.39  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.544 on 197 degrees of freedom
## Multiple R-squared:  0.9389, Adjusted R-squared:  0.9383
## F-statistic:  1515 on 2 and 197 DF,  p-value: < 2.2e-16
```

```
lm.beta(temp_model_2) # Standardized coefficients
```

```
##
## Call:
## lm(formula = enjoyment ~ temperature + temperature_sq, data = shower_data)
##
## Standardized Coefficients::
##    (Intercept)    temperature temperature_sq
##             NA       1.382751      -1.155564
```

The output tells us, that we can explain more than 94% of the variance of enjoyment. The model fits the data significantly with R^2 = .94, F(2, 197) = 1384, p < .001. The linear coefficient (β = 1.38, t(197) = 52.07, p < .001) as well as the quadratic one (β = -1.14, t(197) = -42.80, p < .001) are significantly different from 0.

Literature

Annotation: This exercise sheet bases in part on exercises, that you can find in the textbook *Dicovering Statistics Using R* (Field, Miles & Field, 2012). They were modified for the purpose of this sheet and the R-code was actualized.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. London: SAGE Publications Ltd.