Übungszettel Multiple Regression, Exercise Sheet Multiple Regression

M.Psy.205, Dozent: Peter Zezula

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

25 Mai, 2022 08:57

German

Links

Übungszettel als PDF-Datei zum Drucken

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.

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?

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?

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?

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
```

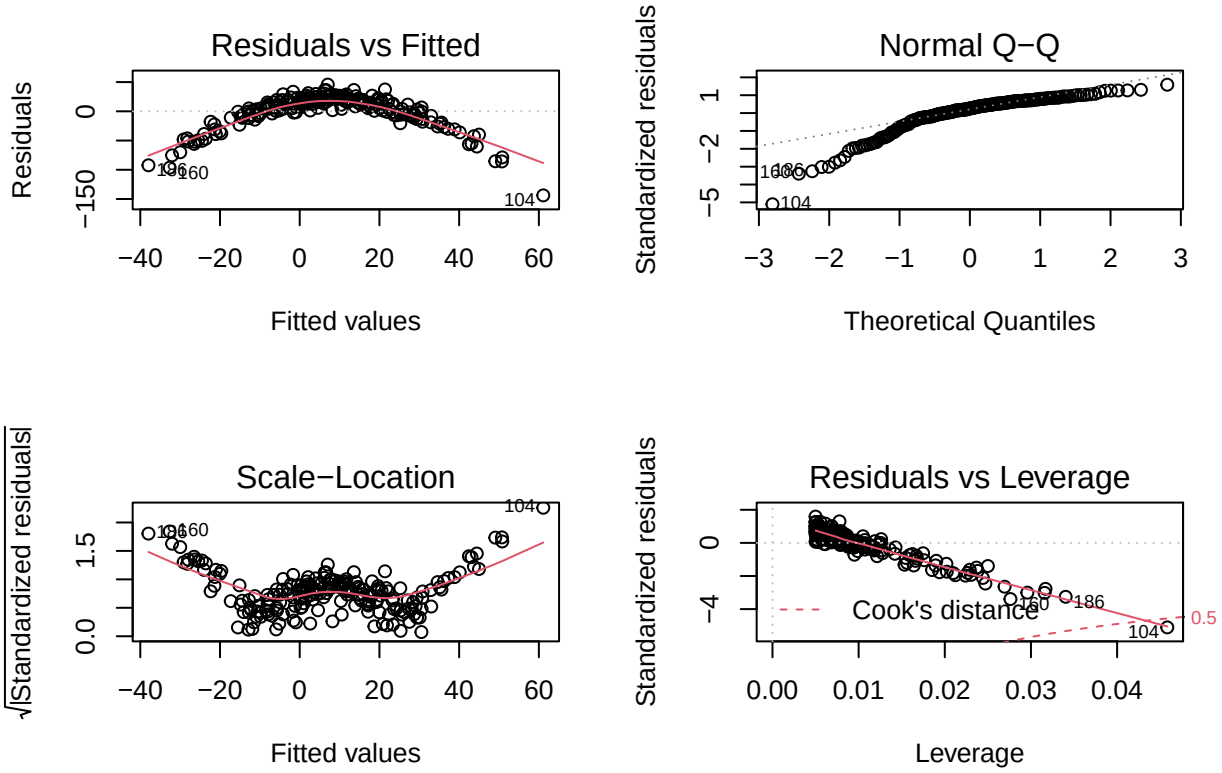
Kommentar der Mitarbeiter:

> 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.

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

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.

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?

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?

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?

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.
```

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

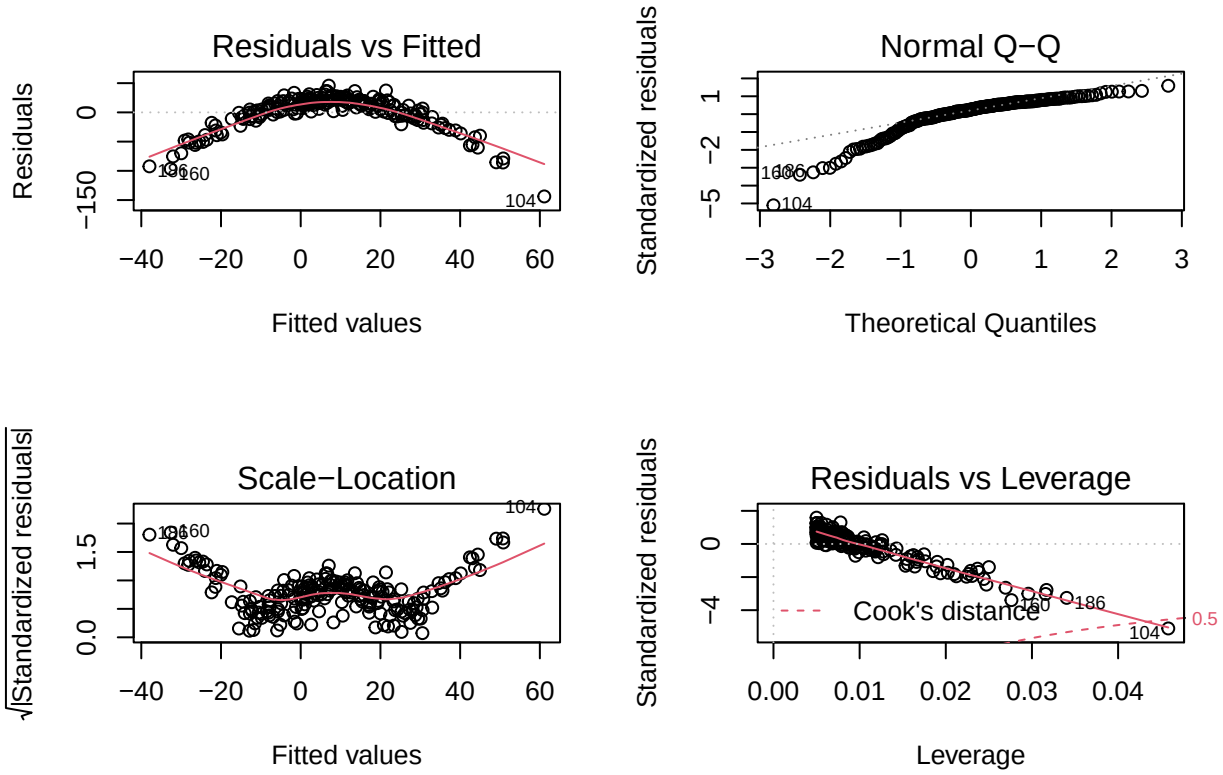
```
##
## Call:
## lm(formula = enjoyment ~ temperature)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -143.811   -7.291    7.326   19.296   45.700
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.7374     2.4260  -1.953   0.0523 .
## temperature   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.

```r
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.

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.