Übungszettel: Einfache lineare Regression (03)

M.Psy.205, Dozent: Dr. Peter Zezula

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

06 Mai, 2021 23:11

Deutsche Version

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 (7.1 - 7.5, 7.9)	Buchkapitel, das Schritt für Schritt erklärt, worum es geht, und wie man Regressionen in R durchführt. **Große Empfehlung!**
R for Data Science	Einsteiger-Buch von R-Gott Hadley Wickham. Hier wird topaktuell in die Arbeit mit R, insbesondere zur Datenaufbereitung und Visualisierung, eingeführt.
R Tutorial	Schritt-für-Schritt Einführung in das Arbeiten mit R von Christian Treffenstädt. Nützlich, falls Sie grundlegende Dinge noch einmal nachschlagen möchten

Tipp der Woche

Mit `strg` + `alt` + `c` (Windows) oder `cmd` + `alt` + `c` (Mac) können Sie direkt den Code-Chunk ausführen, in dem sich Ihr Cursor gerade befindet. Mit `strg` + `alt` + `n` (Windows) oder `cmd` + `alt` + `n` (Mac) führen Sie direkt den nächsten Chunk aus.

1) Daten einlesen

1. Setzen Sie ein sinnvolles Arbeitsverzeichnis für den Übungszettel (in der Regel der Ordner, in dem Ihre .Rmd liegt). Aber Vorsicht: Beim Rendern (Knit) geht RStudio davon aus, dass das Working-Directory das ist, in dem auch die .Rmd-Datei liegt. Dies ist besonders wichtig, wenn es um relative Links geht.
2. Laden Sie den Datensatz starwars.csv herunter (*Rechtsklick > Ziel speichern unter* oder *Rechtsklick > Verknüpfte Datei laden*) und speichern Sie ihn in Ihrem Arbeitsverzeichnis (idealerweise haben Sie noch den Ordner vom letzten Übungszettel - speichern Sie den Datensatz im Unterordner /data).
3. Laden Sie die Pakete des `tidyverse` und fügen Sie eine entsprechende Code-Zeile an den Beginn Ihres Dokuments ein.
4. Lesen Sie den Datensatz `starwars.csv` unter dem Namen `sw_data` in R ein.

2) Regressionsmodell

0. Schlagen Sie für Erklärungen zur Verwendung von `lm()` in Kapitel 7.4.2 und zur Interpretation des Outputs in Kapitel 7.5 von *Discovering Statistics Using R* (Field, 2012) nach.
1. Erstellen Sie ein Regressions-Modell namens `m_height`, in dem Sie das **Gewicht** durch die **Größe** der Personen im Datensatz vorhersagen. Nutzen Sie dafür die FUnktion `lm()`.
2. Lassen Sie sich eine Zusammenfassung der Analyse mit `summary()` anzeigen.
3. Schreiben Sie mit den Werten aus dem Output aus `summary()` die Regressionsgleichung auf.
4. Interpretieren Sie die Regression
    a. Passt das Modell auf die Daten?
    b. Ist Größe ein signifikanter Prädiktor für Gewicht?

3) Scatterplot

1. Erstellen Sie mit den ggplot-Befehlen, die Sie beim letzten Mal kennengelernt haben, einen Scatterplot. Auf der x-Achse sollte die Größe der Personen stehen, auf der y-Achse das Gewicht (`mass`). Hinweis: Die Werte in `mass` sind in Kilogramm angegeben.
2. Fügen Sie dem Plot eine Regressionsgerade hinzu.
3. Fügen Sie dem Plot einen aussagekräftigen Titel und y-, sowie x-Achsenbeschriftungen hinzu.
4. Geben Sie dem Plot ein *Theme*, das ihn Publikationsreif aussehen lässt.
5. Speichern Sie den Plot als .png Datei in Ihrem Arbeitsverzeichnis.

4) Regression überprüfen

Der Plot erzeugt einen Verdacht: Wird die Regression von einem einzigen Extremwert verzerrt? Wir wollen dem weiter auf den Grund gehen.

1. Wenden Sie die Funktion `plot()` auf das Regressionsmodell an, das Sie oben erzeugt haben. Folgen Sie nun den Anweisungen in der Konsole. Dort sollten Sie die Aufforderung "Drücke Eingabetaste für den nächsten Plot:" sehen. Insgesamt werden Ihnen nacheinander vier Plots angezeigt.

2. Lesen Sie diesen kurzen Artiekl. Finden Sie die folgenden Dinge für jeden Plot heraus:

    a. Welche Annahmen der Regressionsanalyse (siehe Field, 2012; Kap. 7.7.2.1) können Sie anhand des Plots überprüfen?
    b. Wie sollte der jeweilige Plot aussehen, wenn alles in Ordnung ist?
    c. Welche Muster in den Plots deuten potentiell auf Probleme mit dem Modell hin? Wenn Sie mehr erfahren möchten, oder zusätzliche Informationen benötigen, ist das Kapitel 7.7.1 in *Field (2012): Discovering Statistics Using R* hilfreich.

3. Schauen Sie sich nun noch einmal die vier Plots an. Was fällt Ihnen auf?

4. Identifizieren Sie, zu welcher Person die auffällige Beobachtung gehört. (Hinweis: In den diagnostischen Plots steht neben Extremwerten in der Regel die zugehörige Zeile im Datensatz.)

5) Korrigierte Regression

1. Erstellen Sie eine Kopie von `sw_data` namens `sw_data_ex`, in der Sie den problematischen Fall ausschließen, den Sie oben identifiziert haben.

2. Führen Sie erneut eine Regressionsanalyse durch und lassen Sie sich den Output mit `summary()` anzeigen.

3. Erstellen Sie, wie oben, einen Scatterplot mit Regressionslinie, basierend auf den neuen Daten. (Tipp: Sie können hier sehr viel Code wiederverwenden.)

4. Vergleichen Sie die erste und die zweite Regressionsanalyse in Bezug auf

    a. Anteil erklärter Varianz (Multiples R^2)
    b. F-Test für Modellfit
    c. t-Test für den Koeffizienten $\hat{\beta}_1$

5. Interpretieren Sie die Ergebnisse der Analyse.

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 Version

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 donloading 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 (7.1 - 7.5, 7.9)	Book chapter with a step for step introduction to simple regression and how to do it in R. **Recommendation!**
R for Data Science	Textbook with an introduction to R
Peters Simple Regression Pages	Peters unit on simple regression. A resource to find running examples.
R Tutorial	A step by step introduction to working with R. authored by Christian Treffenstädt. Useful as a reference for basic stuff.

Tip of the week

You can run directly the code chunk, where your cursor is currently in by using shortcut: `strg` + `alt` + `c` (Windows) or `cmd` + `alt` + `c` (Mac). You can run the following chunk with: `strg` + `alt` + `n` (Windows) or `cmd` + `alt` + `n` (Mac).

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. Load the data starwars.csv and store it in your working directory. You might still have the folder you used for your last sheet - then store the data in a data subdirectory.
3. Assure, that the packages of `tidyverse` are loaded. Insert a code line for that in the beginning of your Rmd-file.
4. Read datafile `starwars.csv` and store it as a data object named `sw_data`.

2) Regression model

0. See chapter 7.4.2 of *Discovering Statistics Using R* (Field, 2012) for the usage of `lm()` and chapter 7.5 for the interpretation of the results.

1. Create a regression model named `m_height` where you predict **mass** by **height**. Use function `lm()` for that.
2. Take a look at the results using `summary()`.
3. Write down a regression equation using the results.
4. Interprete the results of your model:
    a. How good are your data fitted by the model?
    b. Is heigth a significant predictor of weight?

3) Scatterplot

1. Create a scatterplot using the ggplot commands you already know. Show height on the x-axis and weight on the y-axis. Hint: the values in `mass` are kg.
2. Add a regression line.
3. Add meaningful titles and axis labels.
4. Give your plot a *theme* to make it compliant with publication rules.
5. Store your plot in format .png in your working directory.

4) Check regression quality

From the plot we suspect, that only one single outlier might have had too much influence on our results. We want to clarify that.

1. Apply `plot()` to the above generated regression model. Follow the hints on the console. There you should find "press enter for the next plot". You will see four plots in total, one by one.
2. Read this short article. Find out:
    a. Which preconditions of regression can be checked with these diagnostic plots? (see Field, 2012; chap. 7.7.2.1)
    b. What should each plot look like if everything looks good?
    c. What patterns indicate potential problems in the fitted model? If you want to learn more or if you need additional information please refer to chapter 7.7.1 of *Field (2012): Discovering Statistics Using R* hilfreich.
3. Take another look to the four plots. What is special?
4. Identify the outlier (hint: You may find the corresponding line in the data in the diagnostic plots)

5) Modified regression

1. Make a copy of `sw_data` named `sw_data_ex` where you exclude the problematic observation.
2. Repeat your regression analysis and check the output of `summary()`.
3. Make a scatterplot, like above, based on the new data. Hint: you may reuse a lot of the code above.
4. Compare the first and the second regressin analysis with respect to
    a. percentage of explained variance (multiple R^2)
    b. F-test for model fit
    c. t-test for the coefficient $\hat{\beta}_1$
5. Interprete the results of your new analysis.

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.