Übungszettel Logistische Regression, Exercise Sheet Logistic Regression

M.Psy.205, Dozent: Peter Zezula

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

30 Mai, 2022 21:39

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 (Rechtsklick > Speichern unter...).
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 8	Buchkapitel, das Schritt für Schritt erklärt, worum es geht, und wie man logistische Regressionen in R durchführt. **Große Empfehlung!**

Tipp der Woche

Wenn Sie Zeilenumbrüche in Ihrem Code verwenden, ist er für Sie und Andere deutlich besser lesbar. Sie können z.B. nach jeder Pipe (`%>%`) oder nach dem `+` in ggplot2-Befehlen sinnvolle Zeilenumbrüche einbauen.

Zögern Sie auch nicht, längere einzelne Befehle in mehrere Zeilen umzubrechen: Nach jedem Komma kann eine neue Zeile beginnen, wenn Sie es möchten.

Es kann sein, dass die Einrückung bei mehrzeiligen Befehlen irgendwann unordentlich wird. Wenn das passiert, markieren Sie den Code und drücken `strg` + `i` (windows), bzw. `cmd` + `i` (mac). Ihr Code wird dadurch autamtisch eingerückt.

1 Daten einlesen

1. Laden Sie das `tidyverse` und fügen Sie die entsprechende Code-Zeile an den Anfang Ihrer .Rmd-Datei ein.
2. Setzen Sie ein sinnvolles Arbeitsverzeichnis und fügen Sie die entsprechende Code-Zeile an den Anfang Ihrer .Rmd-Datei ein.
3. Laden Sie den Datensatz `stats_lectures` unter diesem Link herunter und speichern ihn in Ihrem Arbeitsverzeichnis, dort z.B. in einem Unterordner `data`.
4. Lesen Sie den Datensatz in R ein.

Erklärung der Variablen

Jede Zeile enthält Daten von einem/r StudentIn. Beobachtet wurde, ob die Studierenden an Statistik-Lehrveranstaltungen teilnehmen. Erhoben wurden dazu die Temperatur, wie sehr die Studis Sonne genießen und wie sehr sie Statistik genießen. Letzteres wurde mit einem Fragebogen gemessen, der aus 5 Items besteht.

Hinweis: Die Daten sind vollständig fiktiv und spiegeln natürlich auch nicht die Meinung der Lehrkräfte wieder. Dem Tutor ist nur gerade kein besseres Beispiel eingefallen.

Name	Bedeutung
attend_or_not	Hat zwei Ausprägungen: “course not attended” und “course attended”. Die Ausprägungen geben an, ob der/die Studi zur Lehrveranstaltung gegangen ist.
temperature	Die Temperatur. Höher ist heißer.
sun_joy	Wie sehr genießen die Studis Sonne. Höher ist stärkerer Genuss
stats_joy1 bis stats_joy5	Wie sehr genießen die Studis Statistik. Höher ist stärkerer Genuss.

2 Datenaufbereitung

1. Bilden Sie für jede Person im Datensatz den Mittelwert aus den 5 Items stats_joy1 bis stats_joy5 und fügen Sie diesen als neue Variable dem Datensatz hinzu. *Hinweis: Eine Google-Suche kann hier sehr hilfreich sein. Achten Sie auch auf Kleinigkeiten in den Befehlen, die Sie finden.* Falls Sie diese Aufgabe nicht lösen können, aber dennoch weiter machen möchten, finden Sie hier funktionierende Syntax, die Sie in Ihr Script kopieren können. Link
2. Nutzen Sie den Befehl `factor()`, um die Variable `attend_or_not` in einen Faktor umzuformen. Verwenden Sie das Argument `levels = c()`, um die Faktorstufen anzugeben. So stellen Sie sicher, dass Sie wissen, welche Kategorie die Baseline darstellt. Falls Sie diese Aufgabe nicht lösen können, aber dennoch weiter machen möchten, finden Sie hier funktionierende Syntax, die Sie in Ihr Script kopieren können. Link

3 Erste logistische Regression

1. Nutzen Sie den Befehl `glm()` mit dem Argument `family = binomial()`, um die Kursbesuche der Studierenden mit deren Genuss von Statistik vorherzusagen. Benutzen Sie dafür den Mittelwert, den Sie in der vorherigen Aufgabe gebildet haben, als Prädiktor.

2. Lassen Sie sich den Output mit `summary()` anzeigen. "Null deviance" gibt die Deviance-Statistik für das Null-Modell an, "Residual Deviance" für das Alternativmodell (user spezifiziertes). Was können Sie aus der Deviance ableiten? (Siehe Field, Kapitel 8.3.2: Assessing the model: the deviance statistic)
3. Bilden Sie mit `exp()` das Exponential zur Basis e für die Regressionskoeffizienten. Dies ist das Odds-Ratio. *Tipp: Mit `$coefficients` können Sie direkt auf die Koeffizienten zugreifen, wenn Sie den Namen, den Sie dem Modell gegeben haben, vor das $-Zeichen schreiben.*
4. Interpretieren Sie das Odds Ratio für den vorliegenden Fall. (Siehe Field, Kapitel 8.3.6: The odds ratio)
5. Wenden Sie `confint()` auf das Modell an, um die Konfidenzintervalle für die Prädiktoren zu erhalten. Bilden Sie auch hiervon das Exponential mit `exp()`, um die Konfidenzinternalle für die Odds-Ratios zu erhalten.
6. Erstellen Sie einen Plot, um Ihr Modell zu visualisieren.

    a) Wenden Sie den Befehl `fitted()` auf das Modell an, um die durch das Modell vorhergesagten Wahrscheinlichkeiten für Ihre Studierenden zu erhalten.
    b) Fügen Sie diese Werte Ihrem Datensatz hinzu.
    c) Nutzen Sie `gpplot`, um einen Plot zu erstellen. Verwenden Sie auf der X-Achse stats_joy und auf der y-Achse die in a) und b) erstellten vorhergesagten Wahrscheinlichkeiten.

4 Modellvergleich logistische Regression

1. Erstellen Sie drei weitere Modell mit dem `glm()`-Befehl:

    a) Fügen Sie den ersten Modell den Prädiktor temperature hinzu
    b) Fügen Sie dem Modell aus a) den Prädiktor sun_joy hinzu
    c) Fügen Sie dem Modell aus b) die Interaktion aus temperature und sun_joy als Prädiktor hinzu.

2. Testen Sie mit dem Befehl `anova()`, welches der Modelle am besten auf die Daten passt. Nutzen Sie die Option `test = "Chisq"`, um einen p-Wert zu erhalten.

    a) Sehen Sie sich mit `summary()` den Output für das beste Modell an.
    b) Welche Hypothesen würden Sie aus dieser Analyse ableiten?

3. Berechnen Sie für das beste Modell die Odds Ratios der Prädiktoren und deren Konfidenzintervalle.
4. Berechnen Sie für das beste Modell die drei Pseudo-R^2 Indices (Siehe Field, Kapitel 8.3.3: Assessing the model: R and R^2). Sie können dafür die unten stehende Funktion nutzen. *Mehr dazu in Field, R's Souls' Tip 8.2.*

    a) Kopieren Sie den Code und führen Sie in aus.
    b) Sie können nun die Funktion `logisticPseudoR2s()` auf ihr Modell anweden.

```
logisticPseudoR2s = function(LogModel) {
  dev = LogModel$deviance
  nullDev = LogModel$null.deviance
  modelN = length(LogModel$fitted.values)
  R.l = 1 - dev / nullDev
  R.cs = 1 - exp(-(nullDev - dev) / modelN)
  R.n = R.cs / (1 - (exp(-(nullDev / modelN))))

  cat("Pseudo R^2 for logistic regression\n")
  cat("Hosmer and Lemeshow R^2 ", round(R.l, 3), "\n")
  cat("Cox and Snell R^2 ", round(R.cs, 3),"\n")
  cat("Nagelkerke R^2 ", round(R.n, 3), "\n")
}
```

5 Rendern (knit)

Lassen Sie die Datei mit `Strg` + `Shift` + `K` (Windows) oder `Cmd` + `Shift` + `K` (Mac) rendern. Sie sollten nun im "Viewer" unten rechts eine "schön aufpolierte" Version ihrer Datei sehen. Falls das klappt: Herzlichen Glückwunsch! Ihr Code kann vollständig ohne Fehlermeldung gerendert werden. Falls nicht: Nur mut, das wird schon noch! Gehen Sie auf Fehlersuche! Ansonsten schaffen wir es ja in der Übung vielleicht gemeinsam.

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 as PDF

Some hints

1. Please try to solve this sheet in an .Rmd file. You can create one from scratch using `File > New file > R Markdown...`. You can delete the text beneath *Setup Chunk* (starting from line 11). Alternatively, you can download our template file unter this link (right click > save as...).
2. You'll find a lot of the important information on the website of this course
3. Please don't hesitate to search the web for help with this sheet. In fact, being able to effectively search the web for problem solutions is a very useful skill, even R pros work this way all the time! The best starting point for this is the R section on the programming site Stackoverflow
4. On the R Studio website, you'll find highly helpful cheat sheets for many of R topics. The base R cheat sheet might be a good starting point.

Ressources

Since this is a hands-on seminar, we won't be able to present each and every new command to you explicitly. Instead, you'll find here references to helpful ressources that you can use for completing this sheets.

Ressource	Description
Field, chapter 8	Book chapter explaining step by step the why and how of logistic regression in R. **Highly recommended!**

Hint of the week

You should use new lines in your code to make it better readable for yourself and for others. You could f. e. continue in a new line after every pipe (`%>%`) or after a `+` in a sequence of ggplot2 commands. Don't hesitate

to divide a complex and long command into several lines. After every comma (`,`) you can continue in a new line, if you like.

If your code looks scattered because of unregular indentation, you may mark a section of code and press `strg` + `i` (Windows), or `cmd` + `i`(Mac) to automatically indent your source code.

1 Read data

1. Load package `tidyverse` and insert the apropriate code at the beginning of your .Rmd document.
2. Change to an adequate working directory. Take care: The render process assumes, that your working directory is set to the location of your .Rmd document.
3. Load the data `stats_lectures` from this link and store it in your working directory or in a subfolder there. You might want to call it `data`.
4. Read the data in R. Alternatively read the data directly from the URL above.

Explanation of the variables

Each line has the data of one student. You find, whether the student took part in a statistic lession, the current temperature, to which extend the person enjoys sun and, on the other hand, statistics. This was measured using a 5 items questionnaire.

Hint: These are completely fictive data. They don't reflect the docents opinion. This is just an spontaneous idea of a student helper.

name	meaning
attend_or_not	dichotomous variable: "course not attended" and "course attended". The values code, whether the student attendet the course or not.
temperature	high values code high temperatures
sun_joy	the higher the value the higher the joy
stats_joy1 to stats_joy5	the higher the values the more the student enjoys statistics

2 Data preparation

1. Calculate the mean of items stats_joy1 to stats_joy5 and store it in the dataframe resp. tibble. *Hint: A google search might help. Take also care of the details of the commands, you find.* If you aren't able to solve this part of the task and want to continue, you find a working syntax for this part under https://md.psych.bio.uni-goettingen.de/mv/sheets/data/06_sheet_task2-1.R that you might integrate in your script.
2. Use the command `factor()` to transform variable `attend_or_not` into a factor. Use the argument `levels = c()` to specify factor levels. This is how you set the correct reference group or baseline. If you want to keep on and don't know how to do tis, you find a working syntax under https://md.psych.bio.uni-goettingen.de/mv/sheets/data/06_sheet_task2-2.R

3 First logistic regression

1. Use the command `glm()` with argument `family = binomial()` to predict, which students attend the statistics course and whether they enjoy it. The predictor is the mean, that you calculated recently.
2. Check the output using `summary()`. "Null deviance" is the deviance statistic for the null model, "residual deviance" is the same for the user defined alternative model. What can we learn from the deviance? C. f. Field (2012), chapter 8.3.2: Assessing the model: the deviance statistic.

3. Use `exp()` to calculate the exponet to base e for the regression coefficient. The result is the odds ratio. *Tip: `$coefficients` gives you direct access to the coefficients. As always you have to put the name of your result model and a $ before.*
4. Interprete the odds ratio for your example. (See Field (2012), chapter 8.3.6: The odds ratio)
5. Apply `confint()` to your result model to get confidence intervals for your predictors. Also use `exp()` to get the confidence intervals for the odds ratios.
6. Make a plot to visualize your model.

    a) Apply `fitted()` to your model to get the predicted probailities for your student subjects.
    b) Add these values to your data set.
    c) Use `ggplot()` to make a plot. Put `stats_joy` to the x-axes and on the y-axes the predicted probilities, you calculated in a) and b).

4 Model comparison in logistic regression

1. Adapt three more models using the command `glm()`:

    a) Add predictor `temperature`
    b) Add predictor `sun_joy`
    c) Add to the model in b) the interaction of `temperature` and `sun_joy`.

2. Compare the models adapted using the command `anova()`. Which of the models fit the data best, considering parsimonia. Use the parameter `test = "Chisq"` to get a p-value.

    a) Get the `summary()` of the best model.
    b) Which hypothesis would you deduct from this analysis.

3. Calculate the odds ratios and the confidence intervales for the parameters of the best model.
4. Calculate the pseudo R^2 for the best model (c. f. Field (2012), chapter 8.3.3: Assessing the model: R and R^2). You may use the function below for this. *Read more on that in Field (2012), R's Souls' Tip 8.2.*

    a) Copy the code and execute it
    b) You can apply `logisticPseudoR2s()` to your model.

```
logisticPseudoR2s = function(LogModel) {
  dev = LogModel$deviance
  nullDev = LogModel$null.deviance
  modelN = length(LogModel$fitted.values)
  R.l = 1 - dev / nullDev
  R.cs = 1 - exp(-(nullDev - dev) / modelN)
  R.n = R.cs / (1 - (exp(-(nullDev / modelN))))

  cat("Pseudo R^2 for logistic regression\n")
  cat("Hosmer and Lemeshow R^2 ", round(R.l, 3), "\n")
  cat("Cox and Snell R^2 ", round(R.cs, 3),"\n")
  cat("Nagelkerke R^2 ", round(R.n, 3), "\n")
}
```

5 Render (knit)

Render your document using `Strg` + `Shift` + `K` (Windows) oder `Cmd` + `Shift` + `K` (Mac). After that, you shoud see a nice looking version of your document. I that works out, congratulations! Your code could be rendered without any errors. If not, don't be frustrated. Fix your errors. We will help you in our exercise class.

Literature

Note: These sheets are based partially on exercises from the book *Dicovering Statistics Using R* (Field, Miles & Field, 2012). They've been modified for the porpuses of this seminar, and the R code was updated.

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

Version: 30 Mai, 2022 21:39