Modulprüfung

CAS Datenanalyse Modul B1

Anna Drewek

Datum: 27. Mai 2020

Zeit zur Bearbeitung:

2.5 Stunden

Hinweise:

- Die Prüfung findet „Open Book“ statt, d.h. beliebige schriftliche Hilfsmittel sind erlaubt.
- Die Antworten sind in **R-Markdown** abzugeben.
- Die Prüfungsfragen müssen im R-Markdown nicht wiederholt werden. Es ist ausreichend mittels der Aufgabennummer auf die Aufgabenstellung zu verweisen.
- In die Bewertung der Prüfung fliesst ein, ob sich das R-Markdown ohne Fehler kompilieren lässt. Von der Bewertung ausgenommen ist die “Darstellung” der Resultate.
- Benennen Sie Ihre R-Markdown Datei mit **Pruefung_Nachname.Rmd**.
- Am Prüfungsende senden Sie Ihre R-Markdown Datei zur Korrektur per E-Mail an die Dozentin (drew@zhaw.ch). Die E-Mail darf maximal 1 Minute nach Prüfungsende bei der Dozentin eintreffen.
- Erfragte Begründungen sollten in (kurzen) Sätzen formuliert werden und nachvollziehbar sein.

Datensätze

Die Datensätze für die Prüfung befinden sich auf Moodle im Ordner Prüfung. Laden Sie diese mit `load(Pruefung.rda")` in R. Es sollten nun `zueri` und `wine` zur Verfügung stehen, sowie die Funktion `resplot` für die Residuenanalyse.

Punkteverteilung und Note

	Aufgabe 1	Aufgabe 2	RMarkdown	**Total**	**Note**
Mögliche Punkte	12	14	2	28	
Erhaltene Punkte					

Viel Erfolg!

Aufgabe 1 (12 Punkte)

Der Datensatz `zueri` enthält Daten über freie Wohnungen in der Stadt Zürich, die in den ersten Maiwochen 2020 auf Homegate inseriert wurden. Erklärung zu den einzelnen Variablen:

`PLZ`	Postleitzahl
`Zimmer`	Anzahl Zimmer
`Miete`	monatliche Miete
`m2`	Grösse der Wohnung in Quadratmeter
`Objekttyp`	Wohnungstyp
`Balkon`	Gibt es einen Balkon?
`Stockwerk`	In welchem Stockwerk befindet sich die Wohnung?
`Renoviert`	Die Wohnung wurde innerhalb des letzten Jahres renoviert oder es handelt sich um eine Erstvermietung

(a) [2 Punkte]

Passen Sie ein einfaches lineares Regressionsmodell mit Zielgrösse `Miete` und erklärender Variable `m2` an. Geben Sie den geschätzten Achsenabschnitt und die geschätzte Steigung an und interpretieren diese Werte.

(b) [1 Punkt]

Gibt es einen auf 5% signifikanten Zusammenhang zwischen der monatlichen Miete und der Wohnungsgrösse? Begründen Sie Ihre Antwort.

(c) [2 Punkte]

Ein Freund von Ihnen sucht nach einer 100 m^2 grossen Wohnung in Zürich. Berechnen Sie mit Hilfe des angepassten Regressionsmodells die erwartete Miete und geben Sie zusätzlich noch ein 90% Prognoseintervall an. Formulieren Sie für Ihren Freund eine Antwort. Die Antwort soll die errechneten Werte möglichst einfach (d.h. ohne statistische Fachbegriffe) erklären.

(d) [2 Punkte]

Erweitern Sie das Modell um die Variable `Zimmer`. Welches Problem tritt hier auf? Wie könnte man es lösen?

(e) [3 Punkte]

Um das Modell zu verbessern, vergrössern wir nun das Modell mittels **schrittweiser** Variablenselektion mit **BIC**. Starten Sie mit dem Modell aus (a), welches nur die Miete m2 enthält. Benützen Sie alle Variablen im Datensatz **ausgenommen** `PLZ`. Wie viele Schritte werden bei der Variablenselektion durchgeführt? Welche Variablen enthält das final selektierte Modell?

(f) [2 Punkte]

Betrachten Sie das final selektierte Modell aus (e). Welche Variablen haben einen mietreduzierenden, welche einen mieterhöhenden Einfluss? Welche Effekte sind auf 5% signifikant?

Aufgabe 2 (14 Punkte)

Je grösser die Sehnsucht nach einem qualitativ hochwertigen Wein, desto tiefer muss in der Regel in die Tasche gegriffen werden. Mittels *Web-Scrapping* wurden Daten (von *WineEnthusiast*) von 1723 Weine mit Sorte (`variety`), Anbaugebiet (`country`, `province`), Preis (`price`) und Qualität (`points`) gesammelt. Die Qualität der Weine wurde von Sommeliers (`taster_name`) beurteilt.

(a) [2 Punkte]

Als erstes sind wir daran interessiert, ob zwischen der Qualität und dem Preis ein Zusammenhang besteht. Erstellen Sie hierfür ein Streudiagramm mit `points` auf der x-Achse und dem **logarithmiertem** `price` auf der y-Achse. Erkennen Sie einen Zusammenhang? Argumentieren Sie, warum für eine einfache lineare Regression die Logarithmus-Transformation von Preis sinnvoll erscheint.

(b) [4 Punkte]

Passen Sie eine Log-Response Regression (Zielgrösse logarithmierter Preis und erklärende Variable `points`) an die Daten an. Überprüfen Sie die Modelleignung mit den vier Standardgrafiken aus `plot` bzw. `resplot()`. Geben Sie zu jedem Graphen eine kurze Beschreibung und am Ende ein Fazit, ob Sie das Modell für geeignet halten.

(c) [2 Punkte]

Was kann man tun, falls die Residuenanalyse ergibt, dass die Modellvoraussetzungen nicht erfüllt sind? Geben Sie 2 Möglichkeiten an (Stichworte genügen).

(d) [2 Punkte]

Passen Sie nun das folgende multiple Regressionsmodell an:

```
wine$lprice <- log(wine$price)
wine$qpoints <- wine$points^2
fit.wine2 <- lm(lprice ~ points + taster_name + qpoints +
                variety + country, data = wine)
```

Wir möchten nun einen Schnäppchenwein identifizieren:

- Welchen Wein würden Sie gemäss den Residuen als grösstes Schnäppchen definieren? Begründen Sie Ihre Antwort.
- Woher stammt dieser Wein und wieviel kostet er?
- Was ist der vom Modell geschätzte Preis für diesen Wein?

(e) [2 Punkte]

Sind qualitativ gleichwertige Weine aus Österreich auf 5% signifikant teurer als jene aus Frankreich und Deutschland? Begründen Sie Ihre Antwort.

(f) [2 Punkte]

Ein Riesling des Weinguts Hirsch in Österreich erhält vom Sommelier Roger Voss 92 Punkte. Was ist der mittlere vorhergesagte Preis? Geben Sie ein Intervall an, in welchem Bereich sich der wahre Preis mit 90% Wahrscheinlichkeit befindet?

********** **ENDE** **********