Übungen zur Vorlesung

Datenanalyse - Dr. Terveer, Vogt, Pohl

Sommersemester 2022 Blatt 10 28.06.2022

Aufgabe 27 (Zweistichprobenfall der Regression in R)

	x	y
1	1	1
2	2	1
3	3	4
4	4	4
5	5	6
6	6	6
7	7	8
8	8	8
9	9	9

Gegeben sei der links stehende Datensatz:

a) Lesen Sie die Daten in R ein und skizzieren Sie ein Streudiagramm

b) Fügen Sie die Regressionsgerade dem Schaubild hinzu.

c) Rufen Sie die Zusammenfassung der Regression via `summary(lm(...))` auf und interpretieren Sie den Output.

Aufgabe 28 (R (Corona-Datensatz) – Multiple Lineare Regression) Normalerweise bereitet Ihnen Ihre Tutorin Datensätze vor, an denen Sie Ihre R-Kenntnisse vertiefen können. Leider kommt die Tutorin auch gerade ein bisschen in Stress und hat deswegen vergessen, die fehlenden Datenpunkte aus dem Datensatz zu ersetzen. Zum Glück haben Sie gerade die Lineare Regression durchgenommen und wollen deswegen versuchen, mit dieser die fehlenden Datenpunkte zu ersetzen.

a) Lesen Sie den Datensatz der bestätigten COVID19-Neuinfektionen `covid_19_daily_ reports_06-20-2022.csv`[6] in `R` ein. Lesen Sie auch den `meta.csv`-Datensatz ein.

b) Fügen Sie beide Datensätze in ein Dataframe zusammen, sodass keine Spalte doppelt auftaucht. Berechnen Sie die relativen bestätigten Fallzahlen.

c) Teilen Sie den Datensatz in zwei Datensätze auf: Einen, bei dem keine Daten fehlen und einen, der die Zeilen mit den fehlenden Datenpunkten enthält. Nennen Sie den vollständigen Datensatz `training` und jenen mit den fehlenden Datenpunkten `testing`.

d) Führen Sie eine Multiple Lineare Regression auf dem `training`-Datensatz durch, die die relativen bestätigten Fallzahlen erklären soll. Wählen Sie immer 2-3 Variablen als Regressoren aus und interpretieren Sie den Output der Regression. Welche Variable scheint den größten Einfluss auf die bestätigten Fallzahlen zu haben?

e) Nutzen Sie Ihre Erkenntnisse aus Aufgabe d) und dann die Funktion `predict.lm` um die fehlenden Daten im `testing`-Datensatz zu bestimmen.

f) Ihre vergessliche Tutorin hat den richtigen Datensatz doch noch gefunden. Laden Sie den `covid_19_daily_ reports_06-20-2022_COMPLETE.csv`-Datensatz. Plotten Sie erst den `training` Datensatz (d.h. die relativen bestätigten Fallzahlen und die Variable, welche Sie für die Regression genutzt haben). Fügen Sie dann eine rote Regressionsgerade hinzu. Fügen Sie dann die vorhergesagten Punkte vom `testing`-Datensatz hinzu. Schließlich, stellen Sie auch die „wahren“ Datenpunkte aus dem `COMPLETE.`-Datensatz dar.

[6] https://github.com/CSSEGISandData/COVID-19