Arbeitsblatt 5

Multiple lineare Regression (Vielfalt)

Aufgabe 1: Einkommen

Der Datensatz `salary.dat` enthält die Zielvariable y, die den Jahreslohn eines Angestellten in einer Firma beschreibt. Als kontinuierliche erklärende Variable dienen die Variable `experience`, welche die Berufserfahrung (in Anzahl Jahren) enthält und die kategorielle Variable `education` mit den Ausprägungen 1 = Berufslehre, 2 = Maturität und 3 = Hochschulabschluss.

(a)

Verschaffen Sie sich einen Überblick über die Daten. Erstellen Sie hierfür ein Streudiagramm mit Jahreslohn gegen Berufserfahrung und färben Sie die Punkte in Abhängigkeit der Ausbildung unterschiedlich ein. Kommentieren Sie den Plot.

(b)

Passen Sie ein multiples lineares Modell mit den erklärenden Variablen `experience` und `education` an. Stellen Sie dabei sicher, dass die Variable `education` eine Faktorvariable ist. Welche und wie viele Dummy-Variablen werden von R zur Modellierung der Variable `education` verwendet?

(c)

Geben Sie die angepassten Gleichungen für Berufslehre, Maturität und Hochschulabschluss an.

(d)

Zeichnen Sie das geschätzte Modell ins Streudiagramm aus Teilaufgabe (a) ein.

(e)

Haben die erklärenden Variablen einen auf 5% signifikanten Einfluss? Beurteilen Sie die Güte des Modells mit R^2.

(f)

Was ist der Unterschied im Jahreslohn zwischen einer Person mit 20 Jahren Berufserfahrung und Berufslehre und einer Person mit ebenfalls 20 Jahren Berufserfahrung und einem Hochschulabschluss? Ist der Unterschied auf 5% signifikant?

Aufgabe 2: Preis von Diamenten

Die wesentlichen Einflussfaktoren, die den Preis von Diamanten bestimmen, sind das Gewicht, die Reinheit, die Farbe und der Schliff. Im Englischen spricht man von den vier C's: Carat, Clarity, Colour and Cut. Diesem Sachverhalt wollen wir anhand von 307 runden Diamantsteinen nachgehen, die am 18. Februar 2000 in Singapur gehandelt wurden. Die Preisangaben sind in Singapur-Dollars und das Gewicht in Karat (ein Karat entspricht 0.2 Gramm). Die Daten dazu sind in `diamant.dat` gespeichert.

(a)

Als erstes betrachten wir nur den Einfluss des Gewichts auf den Preis. Logarithmieren Sie die beiden Variablen `Carat` und `Preis` und passen ein quadratisches Polynom an, d.h.

$$\log(\text{Price}_i) = \beta_0 + \beta_1 \log(\text{Carat}_i) + \beta_2 (\log(\text{Carat}_i))^2 + E_i$$

Wie lautet das "rücktransformierte" Modell, d.h. $\text{Price}_i = \ldots$? Weshalb wird von Beginn an mit Logarithmus transformiert?

(b)

Erstellen Sie ein Streudiagramm mit dem angepassten Modell.

(c)

Ist der quadratische Term wesentlich zur Modellierung dieser Daten?

(d)

Passen Sie nun ein etwas komplexeres Regressionsmodell an, in dem Sie `Colour`, `Clarity` und `CBody` zum quadratischen Polynom hinzufügen. Um wie viele Parameter wird das Modell erweitert? Bringen die zusätzlichen Variablen eine Verbesserung?

(e)

Ein Käufer möchte einen runden 0.4 karätigen Diamanten mit GIA Zertifikat von Reinheit "IF" kaufen. Bei der Farbe ist er sich noch etwas unschlüssig zwischen "D" und "E". Geben Sie ihm, ein 95%-Prognoseintervall für den jeweiligen Preis.