2 Binäre Regression (I)

Aufgabe 1

Der Datensatz `shuttle` beschreibt für die 23 Space Shuttle Flüge vor dem Challenger-Unglück 1986 die Temperatur ($°F$) zur Startzeit sowie das Auftreten bzw. Nicht-Auftreten einer thermischen Überbeanspruchung eines bestimmten Bauteils. Er enthält folgende Variablen:

`flight`	Nummer des Fluges
`temp`	Temperatur in $°F$
`td`	Thermische Überbeanspruchung (1 = Ja / 0 = Nein)

Laden Sie den Datensatz `shuttle` von der Vorlesungshomepage herunter. Öffnen Sie R, und lesen Sie den Datensatz mit dem Befehl `read.table()` ein. Erzeugen Sie eine zusätzliche Spalte `tempC`, welche die Temperatur in Grad Celsius angibt. Dabei gilt die Umrechnung $T_F = 1.8 \cdot T_C + 32$.

(a) Vergleichen Sie die Temperaturen, die bei `td` $= 1$ gemessen wurden mit jenen bei `td` $= 0$. Was ist an einer derartigen Analyse problematisch?

(b) Fitten Sie nun mittels der Funktion `lm()` ein lineares Modell mit Temperatur in $°F$ als Prädiktor. Erstellen Sie einen Plot, der die beobachteten Werte von `td` und `temp` zeigt, sowie die mit dem linearen Modell geschätzten Wahrscheinlichkeiten.

(c) Welche Parameter-Schätzer ergeben sich, wenn die Temperatur in $°C$ gemessen wird? Welche Auswirkungen hätte die gleichzeitige Aufnahme von `temp` und `tempC` ins Modell?

(d) Betrachten Sie die `summary` des linearen Modells und interpretieren Sie die auftretenden Zahlen. Ist der lineare Term signifikant? Warum ist der hier verwendete Test problematisch? Was spricht darüber hinaus gegen die Verwendung des linearen Modells?

(e) Fitten Sie nun mittels der Funktion `glm()` für binomialverteilten Response ein GLM mit Logit-, Probit- sowie komplementären Log-log-Link und betrachten Sie jeweils die `summary`. Erstellen Sie Plots analog zu (c). Welche Auswirkung hat der Übergang von $°C$ zu $°F$ (bzw. umgekehrt) im GLM (mit Begründung)?

(f) Wie kann der Steigungsparameter im Logit-Modell interpretiert werden?

(g) Berechnen sie für alle drei Linkfunktionen die Wahrscheinlichkeit einer thermischen Überbeanspruchung bei der am Tage des Challenger-Unglücks herrschenden Temperatur von $31\,°F$ (*Hinweis:* die Response-Funktion des komplementären Log-log-Link ist die Verteilungsfunktion der Minimum-Extremwertverteilung (Gompertz-Verteilung), $h(\eta) = 1 - \exp(-\exp(\eta))$).

(h) Für welche Temperatur beträgt diese Wahrscheinlichkeit jeweils 0.5?