STATISTIK 2 (107.325) WS 2014
COMPUTERSTATISTIK (107.258) WS 2014

Übung 9 **10. Jänner 2015**
Dutter

33. Betrachten Sie wieder die Daten `werner_bcd`, bereinigen diese (2 Ausreißer und fehlende Werte) und schätzen ein Modell, in dem Cholesterin von allen anderen Variablen (außer Patientennummer) linear abhängt.

    - Welche geschätzte Parameter sind von null signifikant verschieden (siehe summary-Tabelle)? Schätzen Sie ein 2. Modell mit nur den wesentlichen Variablen.
    - Bestimmen Sie aus den Einträgen in der 'summary-Tabelle' approximativ Konfidenzintervalle für die Parameter.
    - Wie können Sie aus den Konfidenzintervallen schnell die Signifikanz erkennen?

34. Prüfen Sie (grafisch) bei den obigen beiden Modellen die Verteilung der Residuen. Ist die Annahme der Normalverteilung gerechtfertigt?

    Gibt es beim Residuen/$\hat{y}$-Plot Auffälligkeiten? Welche Aussagen über das Modell könnte man mit diesem Plot machen?

35. Testen Sie formal (F-Test) das oben gefundene „ideale“ Modell gegen jenes mit allen sinnvollen 8 unabhängigen Variablen. Bitte um textliche Schlussfolgerung.

35a. **Vergleich: Prüfung über Extraquadratsummen bzw. Varianzvergleich:**

Schlussfolgerung: Die beiden Tests sind nicht äquivalent, weil die p-Werte verschieden sind. Die 0-Hypothesen sind ja auch verschieden.

Illustration:

$$\frac{(SS_\omega - SS_\Omega)/(p-q)}{SS_\Omega/(n-p)} = (\frac{SS_\omega}{SS_\Omega} - 1)/\frac{p-q}{n-p} \sim F_{p-q,n-p}$$

ist anders als bei var.test

$$\frac{SS_\omega/(n-q)}{SS_\Omega/(n-p)} \sim F_{n-q,n-p}$$

Dies ist aber noch kein Beweis, aber die obige, numerische Rechnung dürfte wohl genügen.

36. Verwenden Sie ein (relativ) großes Modell, nämlich wieder die Marathondaten von 2005 bis 2014 (siehe Übung 6) und betrachten die folgenden Möglichkeiten zur Erstellung eines linearen Modells:

    Löschen Sie zunächst alle Messungen, deren Zwischenzeit kleiner als eine halbe Stunde ist.

    - Endzeit = f(Zwischenzeit, Jahr)
      Bem.: 2006 hat keine Zwischenzeit, aber das macht nichts. Auswirkung nur in den Freiheitsgraden.

- Endzeit = f(Zwischenzeit, Jahr, Geschlecht)
- Endzeit = f(Zwischenzeit, Jahr, Geschlecht, Altersklasse), wobei einschränkend nur die Klassen c("H","30","35","40","45","50","55") verwendet werden sollten.

Bitte um Kommentare, auch sehr persönliche An(Auf)regungen.

Bitte, stellen Sie die Ausarbeitung in Form eines pdf-Files (nicht mehr als 3 Seiten) mit den Resultaten (Outputs plus textliche Kommentare) und Kurz-Listing des Programmkodes (Funktion) in die TUWEL-Seite

`https://tuwel.tuwien.ac.at/course/view.php?idnumber=107258-2014W`

bis zum 16. Jänner 2015, 23:45 Uhr.

Bitte den Namen des pdf-File folgendermaßen:

name_exer_9.pdf

wobei 'name' für den Familiennamen steht.

Empfehlenswert ist es, den Bericht mit „Sweave“ zu erstellen. (Vorlage siehe `http://www.statistik.tuwien.ac.at/public/dutt/vorles/` .)

Kreuzen Sie außerdem bitte im Internet jede Nummer des Übungsbeispiels an, das Sie dann in der Übungsstunde an der Tafel (mit Beamer-Unterstützung) vorrechnen wollen und können. Der Termin ist üblicherweise 2 Stunden vor der Übung, d.h. Mo., 12:00. Siehe `https://tuwel.tuwien.ac.at/mod/checkmark/submissions.php?id=196792` .