STATISTIK 2 (107.325) WS 2010
COMPUTERSTATISTIK (107.258) WS 2010

Übung 9 **15. Dezember 2010**
Dutter

33. Betrachten Sie die cissik-Daten (File '`cissik_split.txt`' auf `http://www.statistik.tuwien.ac.at/public/dutt/vorles/`).
Kurzbeschreibung: Experiment bezüglich Produktion von gasförmigem Stickstoff im menschlichen Körper. Die gemessenen Werte hängen vermutlich von der Art der Diät und weiteren Faktoren (Zeitpunkt (m/a) ...) ab.
Gibt es durchschnittliche Unterschiede bezüglich den Faktoren 'Personen', 'Zeitpunkt' (morgens/abends) bzw. bei der 'Art der Diät'? Gibt es Wechselwirkungen?

    Interpretieren Sie nun die 'Person' als Faktor mit zufälligen Effekten.

    Hinweis: Man könnte einfach mit festen Effekten arbeiten und dann händisch laut Skriptum die Teststatistiken berechnen. In der Funktion 'aov' kann allerdings in diesem Fall 'Person' auch als Fehlerterm interpretiert und in die Modellformel eingegeben werden (`+ Error(Person)`).[1]

34. Verwenden Sie die Daten des Vienna City Marathons im Jahr 2010. Betrachten Sie die einfache, lineare (?) Abhängigkeit der Endzeit von der Zwischenzeit. Zeichnen Sie die Werte, die geschätzte Gerade, den Konfidenzbereich für die Gerade und den Toleranzbereich für weitere (zukünftige, unabhängige) Beobachtungen (in Form von Hyperbeln).[2][3]

    Bereinigen Sie die Daten des Vienna City Marathons, indem Sie die Nullen aus den Zwischenzeiten entfernen und weiters alle offensichtlich unsinnigen Werte wie solche, wo z.B. das Verhältnis von Endzeit zu Zwischenzeit kleiner als 1.8 ist. Führen Sie sonst die gleichen Rechnungen wie vorhin durch und diskutieren Sie die Ergebnisse.

35. Betrachten Sie wieder die Daten `werner_bcd`, bereinigen sie (2 Ausreißer und fehlende Werte) und betrachten die einfache, lineare Abhängigkeit von Cholesterin vom Alter (ohne Klasseneinteilung).

    Sind die Parameter der linearen Abhängigkeit (der Modellgeraden) signifikant?

---

[1] Fortgeschrittene können auch die Funktion `lmer` aus dem Paket `lme4` versuchen. (Vorsicht: Verwendung von S4-Klassen.)

[2] Ein Konfidenzintervall mit der Konfidenzzahl α für den Mittelwert $\mu_{y.x}$ an der Stelle x erhält man mit der Formel

$$\hat{y}_x - t_{n-2;1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} < \mu_{y.x} < \hat{y}_x + t_{n-2;1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}}.$$

Ein Toleranzintervall ist folgendermaßen:

$$\hat{y}_x - t_{n-2;1-\frac{\alpha}{2}} S \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} < y < \hat{y}_x + t_{n-2;1-\frac{\alpha}{2}} S \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} \quad .$$

[3] R: siehe auch `predict`.

Um wieviel ändert sich statistisch (d.h. mit Angabe der ungefähren Genauigkeit von +/- entsprechend einem ungefähren Konfidenzintervall) der Cholesterin-Wert pro Jahr?

Zeichnen Sie die Werte, die geschätzte Gerade, den Konfidenzbereich für die Gerade und den Toleranzbereich für weitere (zukünftige, unabhängige) Beobachtungen (in Form von Hyperbeln).

36. Betrachten Sie als Modell Cholesterin linear abhängend von allen anderen Variablen. Welche Variablen sind in erster Linie von Einfluss und welche nicht?

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?id=2604`

bis incl. 9. Jänner 2011.

Bitte den Namen des pdf-File folgendermaßen:

name_exer_9.pdf

wobei 'name' für den Familiennamen steht. Bitte im pdf-File Name, Datum und Seitennumerierung nicht vergessen!

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