Prof. Antony Unwin, Alexander Pilhöfer
Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Institut für Mathematik
Universität Augsburg
http://stats.math.uni-augsburg.de/

Statistik I
Übungsblatt 11

Abgabe: *Dienstag 03. Juli 2012, bis spätestens 12.00 Uhr*; Briefkasten: Statistik I oder per email an die Übungsleiter

Die Aufgaben können auch in 2er-Gruppen bearbeitet und abgegeben werden!

1. (a) Nehmen Sie zu folgender Aussage Stellung: **(1P)**

   "Lineare Regressionsmodelle können nicht verwendet werden, um nicht-lineare Zusammenhänge zwischen zwei Variablen zu modellieren."

   (b) Nennen sie jeweils mindestens eine Möglichkeit, wie sie Verteilungen folgender Variablenkombinationen graphisch abbilden können: stetig, kategoriell, kat/stetig, stetig/stetig, kat/kat. **(1P)**

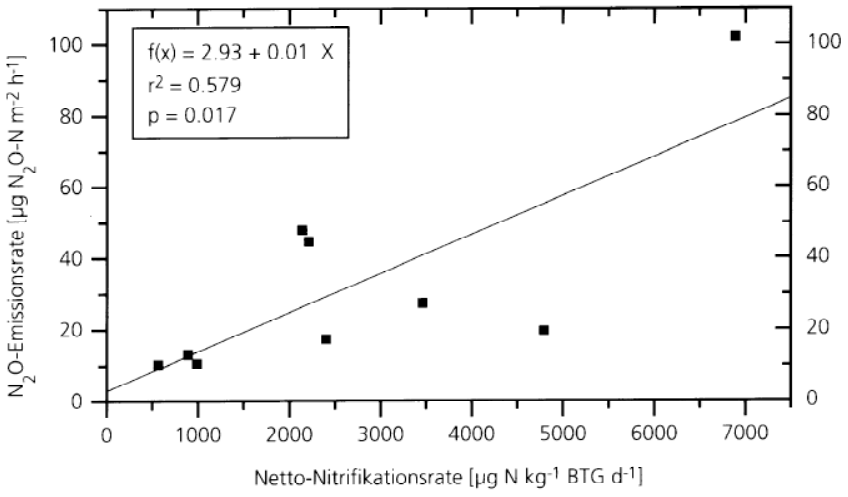
   (c) Warum plottet man zur Überprüfung einer Regression die Residuen gegen die vorhergesagten Werte und nicht gegen die tatsächlichen Werte? **(1P)**

   (d) Wie kann man im R-Output für lineare Modelle (*summary(lm())*) die Anzahl Fälle der verwendeten Daten ablesen? **(1P)**

2. **Regression (5P)**

   Unter der abgebildeten Graphik war der folgende Text abgedruckt:

   Korrelationsanalyse [=Regressionsanalyse] zwischen den unter Einsatz der neuen Methode bestimmten Netto-Nitrifikationsraten und der [...] N_2O-Emissionsrate. Ergebnis der Korrelationsanalyse zwischen N_2O-Emissionsrate und Netto-Nitrifikationsrate des Buchenbestandes im Höglwald.



   (a) Beurteilen Sie die Güte der Regression. Hätten Sie hier eine Regression empfohlen?

(b) Die Steigung der Geraden ist fast null, aber statistisch (einigermaßen) signifikant. Erläutern Sie einem Laien den Unterschied zwischen der statistischen Signifikanz eines Parameters und der Frage, ob ein Parameter verschwindet.

(c) Bei einer weiteren Messung wurde eine Netto-Nitrifikationsrate von 200 gemessen. Welche NO_2-Emissionsrate sagt das Modell vorher und würden Sie dem Wert Vertrauen schenken?

(d) Zeigen Sie, dass für den Stichprobenkorrelationskoeffizienten r gilt, dass $|r| = \sqrt{R^2}$. Berechnen Sie die Korrelation der beiden Variablen.

3. **Sonographie (5P)** Betrachtet werden die Variablen THQ (Thoraxquerschnitt) und FL (Oberschenkellänge) im Datensatz *Sonographie*. Man interessiert sich für deren Zusammenhang und hat ein lineares Regressionsmodell berechnet und dieses in die folgende Grafik eingezeichnet:.

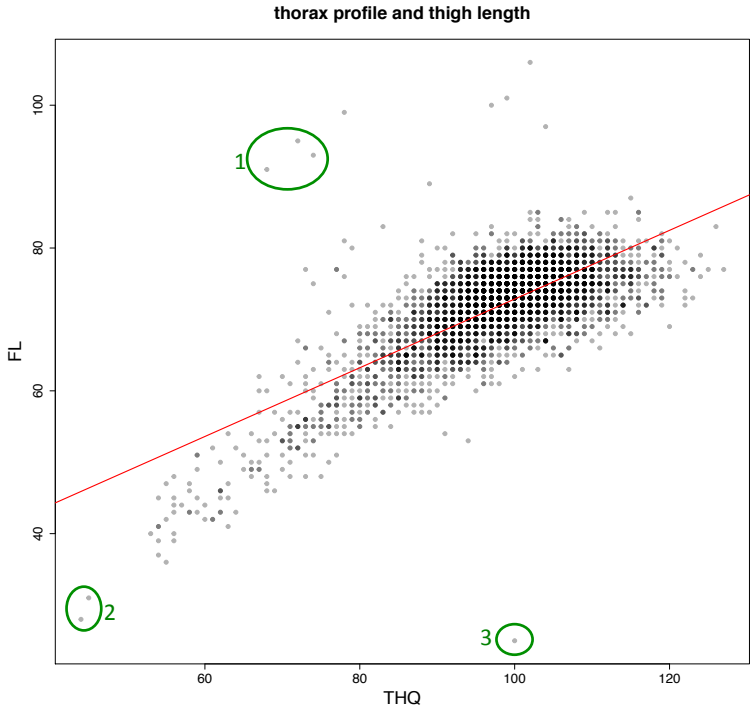


Abbildung 1:

(a) Betrachten Sie obige Grafik. Halten Sie das lineare Modell hier für sinnvoll? Welche Annahmen des Modells sind erfüllt und welche nicht? Gehen Sie auf etwaige Verletzungen der Annahmen präzise ein.

(b) Welche Annahmen wären nach einer monotonen Transformation einer der Variablen möglicherweise erfüllt, welche nicht?

(c) Betrachten Sie die grün markierten Punkte. Geben Sie eine Einschätzung darüber ab, welche der Punkte

- einen hohen Einfluss (Cook's D)
- eine hohe Hebelwirkung

haben.

(d) Wie würde sich die Modellkurve verändern, nähme man die Punkte aus dem Datensatz?

(e) Würde sich die Modellkurve stark ändern, wenn man alle Beobachtungen mit $THQ < 80$ entfernte?

4. **Cars (5P)**

   Betrachtet wird der Datensatz `Cars`. Modelliert werden soll der Verbrauch `City Miles Per Gallon` (kurz `CMPG`).

   (a) Führen Sie eine Regression durch für `CMPG` versus `Weight` und für `CMPG` versus `Horsepower`. Stellen Sie die Zusammenhänge in Scatterplots dar, in die Sie die Regressionslinien einzeichnen. Sind Sie mit den Modellen zufrieden? Gibt es Ausreißer; sind Verletzungen der Annahmen sichtbar?

   (b) Erzeugen Sie Scatterplots von den Residuen versus vorhergesagten Werten. Beschreiben und interpretieren Sie den Unterschied in den Plots.

   (c) Transformieren Sie `CMPG` zu `1/CMPG` und führen Sie die Regression `1/CMPG` versus `Weight` durch. Erzeugen Sie einen Scatterplot der Residuen versus vorhergesagten Werte. Wie interpretieren Sie das Resultat? Welche Probleme erwarten Sie aufgrund der Transformation, und wie kann man diese behandeln?

   (d) In `R` gibt es Funktionen (`residuals`, `rstandard` und `rstudent`) zur Berechnung der folgenden Residuentypen für Modelle:

   - residuals (Residuen),
   - standardised residuals (standardisierte Residuen),
   - studentised residuals (studentisierte Residuen).

   Was sind die Unterschiede zwischen den drei Formen? Wann sind sie nützlich? Welche Form von Residuen würden Sie für den Datensatz `Cars` (vgl. Aufgabe 1) verwenden? Welche Verteilung sollen die standardisierten Residuen haben? Überprüfen Sie das mittels QQ-Plots für Ihre Modelle zum `Cars` Datensatz aus Aufgabe 1.

5. $\chi^2 - Müsli$ **(5P)**

   Auf der Webseite *mymuesli.com* können Sich Kunden selbst Müslis aus einer großen Auswahl von Ingredienzien zusammenstellen. Zum fünften Geburtstag der Firma wurde untersucht, welche Kombinationen am beliebtesten sind.

   Betrachten Sie den Beitrag unter folgendem Link:

   `http://moritz.stefaner.eu/projects/musli-ingredient-network/`

   - Sind die grafischen Darstellungen gelungen? Analysieren und kritisieren Sie diese auf gewohnte Weise.
   - Welche Schlüsse kann man aus den Grafiken ziehen?
   - Wie würden Sie selbst die Daten visualisieren?
   - Wäre ein χ^2-Test im Falle der letzten Matrix angebracht?

6. **Zusatzübung: Multiple Regression (0P)**

   Diese Aufgabe gehört nicht offiziell zum Übungsblatt und dient der eigenständigen Vorbereitung.

   Modellieren Sie im Datensatz *Sonographie* den Kopfumfang des Kindes (*head*) in einem linearen Regressionsmodell mit den Kovariablen *BIP* (biparietaler Durchmesser), *THQ* (Thoraxquerschnitt) und *FL* (Oberschenkellänge). Schränken Sie den Datensatz auf die Beobachtungen ein, die in allen vier Variablen einen Wert besitzen.

   (a) Berechnen Sie die Modelle $head \sim THQ + FL$ und $head \sim THQ$. Sind die Koeffizienten für THQ gleich? Warum?

   (b) Plotten Sie $head \sim THQ$. Zeichnen Sie die Regressionsgerade ein. Zeichnen Sie ebenfalls die Regressionsgeraden des ersten Modells für feste Werte von FL ein. Als feste Werte wählen Sie zB die 5% , 10% , ... , 95% Quantile von FL.

   (c) Berechnen Sie das Modell mit allen drei Kovariablen. Können Sie das Modell verbessern, indem Sie Transformationen wie `I(THQ^2)` hinzuziehen?

   (d) Was sind die Nullhypothesen, Verteilungen und Annahmen bei den Parametertests und dem F-Test?