Roadrace Daten

80 Läufer sind 100 km in 10 Etappen gelaufen. Die Zeiten der Läufer für Etappe j sind in der Variable t_j enthalten. Das (angebliche) Alter der Läufer ist auch im Datensatz enthalten.

Einige Fragen

1. Gibt es besondere Läufer?

2. Gibt es Gruppen mit verschiedenen Laufmustern?

3. Sind die Zeiten für die verschiedenen Etappen miteinander korreliert/assoziiert?

4. Können die Zeiten für die späteren Etappen aus den Zeiten für die früheren hervorgesagt werden?

5. Bekommt man bessere Modelle, wenn man statt die Zeiten die Geschwindigkeiten analysiert?

6.6 Modellwerte und Diagnostiken in der Regression

6.6.1 Hebelwirkung

$$\widehat{y} = X\widehat{b}$$

$$\widehat{b} = (X'X)^{-1}X'y$$

$$\widehat{y} = X(X'X)^{-1}X'y$$

$$= Hy$$

H heißt der Hutmatrix und $H = X(X'X)^{-1}X'$.

H ist

- n x n
- symmetrisch ($h_{ij} = h_{ji}$)
- idempotent ($H^2 = H$)

und $0 \leq h_{ii} \leq 1$

$$\sum h_{ii} = p + 1$$

im Durchschnitt $h_{ii} = \frac{p+1}{n}$

Hebelwirkung $h_i = h_{ii}$ ist die Hebelwirkung der i.ten Beobachtung und mißt wie weit Beobachtung i von den anderen im X-Raum liegt.

6.6.2 Residuen in Multiple Regression

$$\epsilon_i \sim N(0, \sigma^2) \quad u.i.v.$$

$$e = y - X\widehat{b} = y - \widehat{y}$$

$$= y - X(X'X)^{-1}X'y$$

$$= (I - H)y$$

$$V(e) = (I - H)\sigma^2$$

$$V(e_i) = (1 - h_i)\sigma^2$$

d.h. die Varianz des Residuums einer Beobachtung mit größer Hebelwirkung wird klein sein. Um die Verteilung der Residuen zu standardisieren, betrachtet man studentisierte Residuen:

$$r_i = \frac{e_i}{s\sqrt{1 - h_i}} \sim t_{n-p-1}$$

"Student" weil man s statt σ benutzt.
Es gibt zwei Varianten, interne und externe. Bei den internen werden alle Daten benutzt. Bei den externen schätzt man σ^2 mit $s^2(i)$, "s^2 nicht i", d.h. ohne Beobachtung i.

6.6.3 Einfluß — Cook's D

Der Einfluß einer Beobachtung i, dieselbe Formel aber 3 verschiedene Interpretationen:

1. Unterschiede zwischen vorausgesagten Werten

$$D_i = \frac{(\hat{y}_{(i)} - \hat{y})'(\hat{y}_{(i)} - \hat{y})}{(p+1)s^2}$$

2. gewichtete Unterschiede zwischen Parameterschätzern

$$D_i = \frac{(\hat{b}_{(i)} - \hat{b})'(X'X)(\hat{b}_{(i)} - \hat{b})}{(p+1)s^2}$$

   weil $\hat{y} = X\hat{b}$ und $\hat{y}_{(i)} = X\hat{b}_{(i)}$

3. Beobachtung i allein

$$D_i = \frac{r_i^2}{(p+1)} \frac{h_i}{(1-h_i)}$$

   weil

$$\hat{b}_{(i)} - \hat{b} = -(X'X)^{-1}x_i'(1-h_i)^{-1}e_i$$

   Die Schwäche dieser Diagnostiken ist, dass sie sich immer nur auf eine Beobachtung beziehen.

6.7 Auswahl der erklärenden Variablen

Gegeben sind Y und $X_1, \ldots, X_p$.
Man könnte Modelle berechnen:

1. vorwärts schrittweise

2. rückwärts schrittweise

3. alle (!) mögliche Kombinationen (2^p)

Wenn man vorwärts rechnet, versucht man immer die nächstbeste erklärende Variable auszusuchen. Wenn man rückwärts rechnet, wirft man die Variable weg, die am wenigsten zum Modell beiträgt. Es gibt verschiedene Algorithmen dafür, aber alle sind theoretisch unzufriedenstellend, weil
(a) Tests werden mit denselben Daten durchgeführt
(b) das ”beste” Modell wird nicht unbedingt gefunden.

6.8 Warum ein F-Test im Regressionsoutput?

Betrachten wir die Tabelle

	QS	Freiheitsgrade	MQS
Modell	QS_M	p	$\frac{QS_M}{p}$
Residuen	$\sum(y_i - \hat{y}_i)^2$	$n-p-1$	$\frac{QS_R}{n-p-1}$
Insgesamt	$\sum(y_i - \bar{y})^2$	$n-1$	

$$QS_M = \sum(\hat{y}_i - \bar{y})^2$$

$$\frac{QS_R}{n-p-1} = s^2 \quad \text{ist unser Schätzer für } \sigma^2$$

Falls $b_i = 0 \quad \forall i \quad (H_0)$, dann wäre $\frac{QS_M}{p}$ auch ein Schätzer für σ^2. Dann ist die Statistik

$$F = \frac{QS_M/p}{QS_R/(n-p-1)}$$

eine Teststatistik für

$$H_0 : b_i = 0 \quad \forall i$$

mit einer $F_{(p,n-p-1)}$-Verteilung. Dieser Test testet das *Gesamtmodell* und ist deshalb fast immer signifikant.

6.9 Beurteilung eines Modells

1. R^2 (oder ein anderes globales Kriterium)

2. Residuenplots

3. Diagnostiken

   (a) Hebelwirkung

   (b) Einfluß

4. Ist das Modell sinnvoll?

Regression — Zusammenfassung

1. Theorie
   (a) gut für einfache Situationen
   (b) ”grau” für komplizierte Modelle

2. Praxis
   (a) Schwierigkeiten mit der Auswahl der Variablen
   (b) Einfluß von Ausreißern
   (c) Probleme mit der Interpretation der Ergebnisse
   (d) Sehr (sehr) oft angewandt