Ulrich Pötter Wintersemester 2007/2008
Statistische Techniken der Sozialforschung I

Arbeitsblatt 6

Ein wahrscheinlichkeitstheoretisches Modell der linearen Regression ist

$$Y =_d \alpha + X_1\beta_1 + X_2\beta_2 + \ldots + X_p\beta_p + \epsilon$$
$$\epsilon \perp\!\!\!\perp (X_1, X_2, \ldots, X_p), \quad \mathbb{E}(\epsilon) = 0, \quad \mathbb{V}(\epsilon) =: \sigma^2 < \infty$$

oder kurz:

$$Y =_d X\beta + \epsilon, \quad \epsilon \perp\!\!\!\perp X, \quad \mathbb{V}(\epsilon) =: \sigma^2 < \infty$$

$(\beta_1, \ldots \beta_p)$ und σ^2 sind *Parameter* des Modells. Anstelle von Zufallsvariablen kann auch die bedingte Verteilungsfunktion benutzt werden, um das Modell zu beschreiben:

$$\begin{aligned}\Pr(Y \leq y \,|\, X = x) &= \Pr(Y - x\beta \leq y - x\beta \,|\, X = x) \\ &= \Pr(\epsilon \leq y - x\beta) = F_\epsilon(y - x\beta)\end{aligned}$$

1) Simulation

Simulieren Sie 50 Beobachtungen des Modells

$$Y := 0.2 - 0.5X_1 + 0.2X_2 + \epsilon$$

wobei ϵ normalverteilt mit Erwartungswert 0 und Varianz 1 sei (`rnorm(50,0,1)`), X_1 sei gleichverteilt auf dem Intervall (0,2) (`runif(50,0,2)`) und X_2 sei normalverteilt mit Erwartungswert 1 und Varianz 1. X_1, X_2 und ϵ sollen paarweise unabhängig sein.

a) Wiederholen Sie diese Simulation 1000 mal und halten Sie die Ergebnisse der 1000 Regressionen (sowohl die geschätzten Parameter $\hat{\beta}, \hat{\sigma}$ als auch die geschätzten Standardfehler der Parameter) in einer 1000×8 Matrix fest.

Hinweis: Zunächst sollte eine Matrix erzeugt werden, die die Simulationsergebnisse aufnimmt, etwa `erg <- matrix(NA,nrow=1000,ncol=8)`. Dann sollte ein Startwert für den Zufallsgenerator festgelegt werden, so dass die Ergebnisse reproduzierbar sind (etwa durch: `set.seed(123)`). Dann können in einer Schleife (Syntax: `for (i in 1:1000) ...`) die Zufallsvariablen simuliert werden und die Ergebnisse linearer Regressionen (`ergi <- lm(y ~ x1 + x2)`) berechnet werden. Auf die geschätzten Parameter der linearen Regressionen kann mit der Funktion `coef(ergi)` zugegriffen werden, `vcov(erg1)` liefert die geschätzte Varianz-Kovarianz-Matrix der geschätzten Parameter, ihre geschätzten Standardfehler erhält man also durch `sqrt(diag(vcov(ergi)))`. Schließlich können die Ergebnisse jeder der 1000 Simulationen in die Matrix `erg` unter Verwendung des Laufindexes `i` eingetragen werden (z.B. `erg[i,1:3] <- coef(ergi)` etc.).

b) Zeichnen Sie einen Dichteschätzer der geschätzten 1000 Koeffizienten $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2$. *Hinweis:* Z.B. `plot(density(erg[,2]))`.

c) *Konfidenzintervalle:* Ein Konfidenzintervall ist ein Intervall, dessen Ober- und Untergrenzen aus den Daten berechnet werden kann (also eine Statistik ist), so dass für alle Werte eines Parameters (etwa β_1) und ein gewähltes Konfidenzniveau α gilt:

$$\Pr_{\beta_1}(\beta_1 \in (\hat{L}, \hat{U})) \geq 1 - \alpha$$

Wäre $\hat{\beta}_1$ normal verteilt und die Standardabweichung der Statistik ($\sigma(\hat{\beta}_1)$) bekannt, dann wäre tatsächlich

$$\Pr_{\beta_1}(\beta_1 \in (\hat{\beta}_1 - 1.96\sigma(\beta_1), \hat{\beta}_1 + 1.96\sigma(\beta_1))) \geq 0.95$$

Berechnen Sie dieses Intervall für alle 1000 Simulationen, indem Sie die geschätze Standardabweichung $\hat{\sigma}(\hat{\beta}_1)$ anstelle von $\sigma(\beta_1)$ verwenden. Zählen Sie dann, wie oft der Parameter β_1 in diesen Intervallen liegt.

d) Zeichnen Sie einen Dichteschätzer der Statistik $\hat{\sigma}(\hat{\beta}_1)$.

e) *Tests:* Eine Möglichkeit, eine Hypothese über einen der Parameter, etwa $\beta_1 = \beta_1^0$, zu testen, besteht darin, die Statistik

$$\frac{|\hat{\beta}_1 - \beta_1^0|}{\hat{\sigma}(\hat{\beta}_1)}$$

zu betrachten. Diese Zahl sollte groß werden, wenn tatsächlich β_1 stark von β_1^0 abweicht. Wäre wieder die Standardabweichung der Statistik ($\sigma(\hat{\beta}_1)$) bekannt, dann wäre der Fehler 1. Art

$$\Pr_{\beta_1^0}\left(\frac{|\hat{\beta}_1 - \beta_1^0|}{\sigma(\hat{\beta}_1)} \geq 1.96\right) \leq 0.05$$

Berechnen Sie, wie oft unter den 1000 Simulationen tatsächlich

$$\frac{|\hat{\beta}_1 + 0.5|}{\hat{\sigma}(\hat{\beta}_1)} \leq 1.96$$

ist.

Da auch $\hat{\sigma}(\hat{\beta}_1)$ variiert, liegt es nahe, die Verteilung des Bruchs besser anzunähern und dann anstelle des 0.975 Quantils der Normalverteilung das 0.975 Quantil der Verteilung des Bruchs zu verwenden.

Berechnen Sie zunächst das 0.975 Quantil der t-Verteilung mit 50 - 3 = 47 Freiheitsgraden. *Hinweis:* `qt(...)` berechnet die Quantile der t-Verteilung.

Berechnen Sie dann mit diesen Quantilen anstelle des Wertes 1.96, wie häufig

$$\frac{|\hat{\beta}_1 + 0.5|}{\hat{\sigma}(\hat{\beta}_1)}$$

diese neue Schwelle unterschreitet.

f) Berechnen Sie mit diesen Quantilen auch erneut Konfidenzintervalle für β_1 und geben Sie an, wie häufig der Parameter von diesen Intervallen überdeckt wird.

g) Benutzen Sie nur die Ergebnisse der letzten Simulation `ergi` und versuchen Sie, die Plots zu reproduzieren, die durch `plot(ergi)` erzeugt werden. *Hinweis:* Viele der verwendeten Größen sind unter dem Titel `influence.measures` im Paket `stats` zusammengefasst.

h) `dfbeta(ergi)` berechnet die Veränderung der Werte der Regressionskoeffizienten, wenn je eine der 50 simulierten Fälle aus der Berechnung der Regression ausgeschlossen werden.

Berechnen Sie die Varianz-Kovarianz Matrix dieser Veränderungen. Berechnen Sie auch die Standardabweichungen getrennt für die drei Regressionsparameter, multiplizieren Sie sie mit $\sqrt{n} = \sqrt{50}$ bzw. mit $\sqrt{n-3}$ und vergleichen Sie diese Zahlen mit den geschätzten Standardabweichungen $\hat{\sigma}(\hat{\beta}_1)$.