Ulrich Pötter Wintersemester 2007/2008
Statistische Techniken der Sozialforschung I

Arbeitsblatt 4

1) Monte Carlo

Wenn man n unabhängige und identisch verteilte Zufallsvariable auf einem Rechner erzeugen könnte, legt das starke Gesetz der großen Zahlen es nahe, Erwartungswerte durch Mittelwerte der Zufallsvariablen anzunähern. Kann man Zufallsvariable mit der Verteilungsfunktion $F(.)$ erzeugen, kann man $E_F(X) = \int x\,dF(x) = \int xf(x)\,dx$ durch den Mittelwert $M(X) = 1/n \sum_{\omega\in\Omega} X(\omega)$ annähern. Entsprechend lassen sich auch Erwartungswerte für Funktionen von X annähern. Ist $E_F(g(X)) = \int g(x)\,dF(x) = \int g(x)f(x)\,dx$ für eine Funktion $g(.)$, dann kann $1/n \sum g(X(\omega))$ als Näherung benutzt werden. Insbesondere kann man für $g(.)$ die Indikatorfunktion $I[x \in A]$ wählen, die den Wert 1 annimmt, wenn $x \in A$ ist, und den Wert 0 sonst. Man erhält dann eine Näherung für Wahrscheinlichkeiten: $E_F(I[X \in A]) = \int I[x \in A]\,dF(x) = \int_A dF(x) = \Pr(X \in A)$ kann durch $1/n \sum I[X(\omega) \in A]$ angenähert werden.

2) Aufgaben

a) Berechnen Sie 1000 Zahlen $x_i, i = 1, 2, \ldots, 1000$ durch $x_i = ia - \lfloor ia \rfloor$ für ein "irrationales" $a \in (0, 1)$. Dabei bezeichnet $\lfloor x \rfloor$ die größte ganze Zahl, die nicht größer als x ist. `floor(x)` ist der entsprechende Operator in `R`. Ist die Folge 1-distributed?

b) Berechnen Sie für die oben erzeugten Zahlen die Kovarianz zwischen x_i und x_{i-1} (`cov()` ist die entsprechende `R` Funktion). Ist die Folge 2-distributed?

c) Benutzen Sie diese Zahlen zur Berechnung des Integrals $\int_0^1 x\,dx$. Benutzen Sie als nächstes die Funktion `runif()` zur Erzeugung von 1000 Pseudo-Zufallsvariablen und berechnen Sie wieder $\int_0^1 x\,dx$.

3) Zufallsvariable mit vorgegebener Verteilungsfunktion

Sei U eine gleichverteilte Zufallsvariable, also eine Zufallsvariable mit der Verteilungsfunktion $F_U(x) = 0$ für $x \leq 0$, $F_U(x) = x$ für $x \in (0, 1)$ sowie $F_U(x) = 1$ für $x \geq 1$. Ist F eine beliebige stetige Verteilungsfunktion, dann ist $X = F^{-1}(U)$ eine Zufallsvariable mit der Verteilungsfunktion F:

$$\begin{aligned} \Pr(X \leq x) &= \Pr(F^{-1}(U) \leq x) = \Pr(F(F^{-1}(U)) \leq F(x)) \\ &= \Pr(U \leq F(x)) = F(x) \end{aligned}$$

4) Aufgaben

a) Erzeugen Sie 1000 Zufallsvariable mit der Verteilungsfunktion $F(x) = \frac{x}{0.01+x}$ für $x \geq 0$. Berechnen Sie Mittelwert und Varianz dieser Zufallsvariablen.

b) Erzeugen Sie 1000 Zufallsvariable mit der Verteilungsfunktion $F(x) = \frac{\exp(x)}{1+\exp(x)}$ für $x \in \mathbb{R}$. Berechnen Sie Mittelwert und Varianz. Zeichnen Sie auch die Verteilungsfunktion und Vergleichen Sie sie mit der vorgegebenen Verteilungsfunktion (die Verteilungsfunktion kann mit `plot(ecdf(X))` gezeichnet werden, `curve(f, from= , to= , add=T)` zeichnet Funktionen zwischen `from` und `to`. Ist `add=T`, dann wird das Bild zum vorherigen Plot hinzugefügt).

c) Erzeugen Sie 1000 Zufallsvariable mit der Verteilungsfunktion $F(x) = 1 - \exp(-\exp(x))$ für $x \in \mathbb{R}$. Berechnen Sie Mittelwert und Varianz und zeichnen Sie die Verteilungsfunktion.

5) PSID: Einkommen

Der Zusammenhang zwischen individuellem (Arbeits-) Einkommen (`I11110_2`) und Alter (`D11101_2`), Geschlecht (`D11102LL`), Schuljahren (`D11109_2`) und Arbeitszeit (Stunden je Jahr) (`E11101_2`) soll mit der 2003 Welle der PSID-Studie untersucht werden. Es sollen nur Beschäftigte `E11102_2==1` der Hauptstichprobe (`X11104LL==11`) betrachtet werden, die auch zur Stichprobe gerechnet werden (`X11103_2==1`).

```
X <- read.spss("../../cnef/psid/pequiv_2003.sav",use=F,to=T)
X <- subset(X, E11102_2==1&X11104LL==11&X11103_2==1,
            c(X11101LL,X11102_200,D11102LL,
              D11101_2,D11109_2,E11101_2,I11110_2))
```

Wieviele Fälle bleiben übrig? In wievielen Fällen fehlen Angaben? Sind die Bereiche der Variablen plausibel? Wieviele Beschäftigte gibt es je Haushalt? Wieviele Haushalte sind in dieser Teilstichprobe?

6) Lineare Regression

Ein lineares Modell (multiple Regression mit der Methode der kleinsten Quadrate) wird mit der Funktion `lm()` berechnet:

```
attach(X)
erg <- lm(I11110_2 ~ D11101_2 + D11102LL + D11109_2 + E11101_2)
summary(erg)
```

Die `lm()`-Funktion muss mindestens eine Formel enthalten, die das zu schätzende Modell beschreibt: Links von dem Zeichen `~` steht der Name der abhängigen Variablen, rechts davon die Namen der unabhängigen („erklärenden“) Variablen, die durch `+` verbunden werden.

`summary(erg)` erzeugt eine Zusammenfassung der Ergebnisse der Regression mit den geschätzten Koeffizienten, deren Standardfehler, t-Werten und deren beobachtetes Signifikanzniveau. Außerdem werden R^2 Werte und weitere Fitstatistiken ausgegeben.

Das Ergebnis eines Aufrufes von `lm()` ist eine Liste, die u.a. die folgenden Elemente enthält: `coefficients` (die Regressionskoeffizienten), `residuals` (die Residuen des Modells), `fitted.values` (die vorhergesagten Werte). Die entsprechenden Elemente können also etwa durch `erg$coefficients` angesprochen werden.

Oft ist es aber einfacher, anstelle der Listenelemente entsprechende Funktionen zu verwenden, um auf Teilergebnisse zuzugreifen. Für alle Regressionsmodelle gibt es die folgenden Funktionen: `coef(erg)`, `resid(erg)`, `fitted(erg)`, `summary(erg)`, `vcov(erg)` (die geschätzte Kovarianzmatrix der geschätzten Parameter. Sie ist nicht direkt in der Liste der Ergebnisse (`erg`) enthalten):

```
coef(erg)
kov <- vcov(erg)
stdfehler <- sqrt(diag(kov));stdfehler
```

7) Diagnostische Plots

`plot(erg)` erzeugt vier diagnostische Plots, die es erlauben sollen, Probleme des Modells aufzuzeigen. Das erste Bild zeigt ein Scatterplot der Residuen geordnet nach den vorhergesagten Werten der abhängigen Variablen. Insbesondere sollten sich Hinweise auf Heteroskedastie sowie Ausreißer und Abweichungen von der linearen Form des Kovariableneinflusses finden lassen.

Das nächste Bild zeigt ein Q-Q Diagramm (Quantil-Quantil Diagramm) der Residuen im Vergleich mit den Quantilen einer Normalverteilung.

Bild 3 zeichnet die Wurzel der standardisierten Residuen gegen die vorhergesagten Werte. Das sollte die Form einer möglichen Heteroskedastie zeigen.

Im letzten Bild werden die standardisierten Residuen gegen den „Leverage“ der Beobachtungen (den „Abstand“ der unabhängigen Variablen der Beobachtungen von ihrem Mittelwert) abgetragen. Damit sollten sich auch Ausreißer in den Kovariablen identifizieren lassen.

8) Fehlende Werte

`complete.cases(X)` erzeugt einen logischen Vektor der Länge `dim(X)[1]`, dessen Elemente den Wert `TRUE` genau dann haben, wenn die entsprechende Zeile eines `data.frame` (oder einer Matrix oder eines Vektors) keine fehlenden Werte (`NA` oder `NaN`) enthält.

`na.omit(X)` entfernt alle Zeilen aus einem `data.frame` oder einer Matrix (und aus Vektoren), die fehlende Werte enthalten.

```
X1 <- na.omit(X)
dim(X1)
```

Die Zeilennummern (und deren Namen), die durch `na.omit` ausgeschlossen werden, werden als Attribut des erzeugten Objekts mit dem Namen `na.action` gespeichert. Sie können durch

```
om1 <- attr(X1,"na.action")
```

angesprochen werden.

`na.exclude(X)` entfernt ebenfalls alle Zeilen mit fehlenden Werten. Beide Befehle ergeben numerisch identische Ergebnisse und halten beide die Zeilennummern der ausgeschlossenen Zeilen fest:

```
X2 <- na.exclude(X)
all(X1-X2==0)  # TRUE
om2 <- attr(X2,"na.action")
class(om1); class(om2)
```

Alle multivariaten statistischen Prozeduren verwenden zunächst `na.omit` oder `na.exclude`, bevor sie ihre Ergebnisse berechnen.

Der Unterschied besteht in der Behandlung fehlender Werte bei der Berechnung von Residuen und vorhergesagten Werten insbesondere in Regressionsmodellen: Bei `na.exclude` werden die ausgeschlossenen Fälle an der „richtigen“ Stelle wieder in die Vektoren der Residuen bzw. der vorhergesagten Werte aufgenommen.

Die Voreinstellung zur Behandlung fehlender Werte ist durch den Wert von `options("na.action")` gegeben und kann durch `options(na.action="na.exlude")` bzw. durch `options(na.action="na.omit")` verändert werden.