Dr. Christof Luchsinger

Übungsblatt 10 zur Vorlesung

"Statistische Methoden"

Einfache Regression

Herausgabe des Übungsblattes: Woche 19, Abgabe der Lösungen: Woche 20 (bis Freitag, 1615 Uhr), Besprechung: Woche 21

Must

Aufgabe 49 [Simulation einer Regression]

a) Sei $x_i := i, 1 \leq i \leq 100$, eine feste Folge von erklärenden Daten (äquidistant). Es gelte

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

wobei $\epsilon \sim \mathcal{N}(0, 2)$ iid. Nehmen Sie $\beta_0 = 1$ und für β_1 nehmen Sie Ihre PN. Generieren Sie jetzt einen Vektor $(y_i)_{i=1}^{100}$.

b) Schätzen Sie jetzt β_0, β_1 und die Varianz von ϵ mit R, indem Sie eine OLS-Schätzung machen.

c) Vertauschen Sie die Rollen von x und y und machen Sie nochmals eine OLS-Regression. Vergleichen Sie die Resultate von b) und c).

Standard

Aufgabe 50 [Aufspaltung der Variation in den y] [3 Punkte]

Beweisen Sie mit der Notation aus 7.1.2, dass gilt

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Aufgabe 51 [Erwartungstreue Schätzer bei OLS] [2+1 Punkte]

Beweisen Sie, dass die Schätzer

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \quad \text{und} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

erwartungstreu sind. Tipps: Schätzer werden normalerweise auf der Ebene der Stichproben definiert. Um die Erwartungstreue zu überprüfen, müssen Sie auf die Ebene der Zufallsgrössen wechseln (y durch Y ersetzen) und mit Formel (7.1) arbeiten.

Aufgabe 52 [$y_i = \beta_0 + \epsilon_i$] [1+1+1 Punkte]

Das Modell

$$Y_i = \beta_0 + \epsilon_i$$

ist offenbar ein Spezialfall von (7.1). Berechnen Sie in diesem Modell die OLS-Schätzung von β_0 und die ML-Schätzungen von β_0 und $\sigma^2 := V[\epsilon_i]$.

Aufgabe 53 [Testen, ob $\beta_1 = 0$, $\beta_1 = PN$] [1+2 Punkte]

Testen Sie in der Situation von Aufgabe 49 (mit den dort erzeugten Daten), ob

a) $\beta_1 = 0$ und

b) $\beta_1 = PN$, wobei das PN das richtige (und uns in der Simulation ja bekannte) β_1 ist.

Wir geben hier kein Signifikanzniveau vor. Geben Sie den P-Wert an, d.h. sagen Sie, bis zu welchem Signifikanzniveau die $\mathcal{H}_0$-Hypothese noch aufrecht erhalten wird.

Honours

Aufgabe 54 [Test, ob $\beta_0 = 0$] [5 Punkte]

In 7.1.3 haben wir einen Test entwickelt, ob $\beta_1 = 0$ oder nicht. Entwickeln Sie jetzt mit analogen Überlegungen einen Test für die Frage ob $\beta_0 = 0$ oder nicht.

Übungsblatt 10 zur Vorlesung “Statistische Methoden”

Olivier Warin

22. Mai 2011

Aufgabe 49 [Simulation einer Regression]

```
> #Teilaufgabe a):
> PN <- 2
> N <- 100
> sigma <- sqrt(2)
> beta0 <- 1
> beta1 <- PN
> x <- 1:N
> y <- beta0 + beta1*x + rnorm(N,0,sigma)
>
> #Teilaufgabe b):
> #Zuerst von Hand:
> SSxx <- sum((x-mean(x))^2)
> SSxy <- sum((x-mean(x))*(y-mean(y)))
> beta1hut <- SSxy/SSxx
> beta0hut <- mean(y) - beta1hut*mean(x)
> yhut <- beta0hut + beta1hut*x
> sigma2hut <- 1/(N-2)*sum((y - yhut)^2)
> beta0hut
[1] 0.998394
> beta1hut
[1] 2.001701
> sigma2hut
[1] 2.204446
> sqrt(sigma2hut) #(weil R die Standardabweichung angibt)
[1] 1.484738
> #Nun noch direkt mit R:
> mydata <- data.frame(x,y)
> regr <- lm(y~x,mydata)
> summary(regr)

Call:
lm(formula = y ~ x, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-3.44556 -1.00851  0.02655  0.87074  3.76095

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.998394   0.299189   3.337  0.00120 **
x           2.001701   0.005144 389.168  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.485 on 98 degrees of freedom
Multiple R-squared: 0.9994,    Adjusted R-squared: 0.9993
```

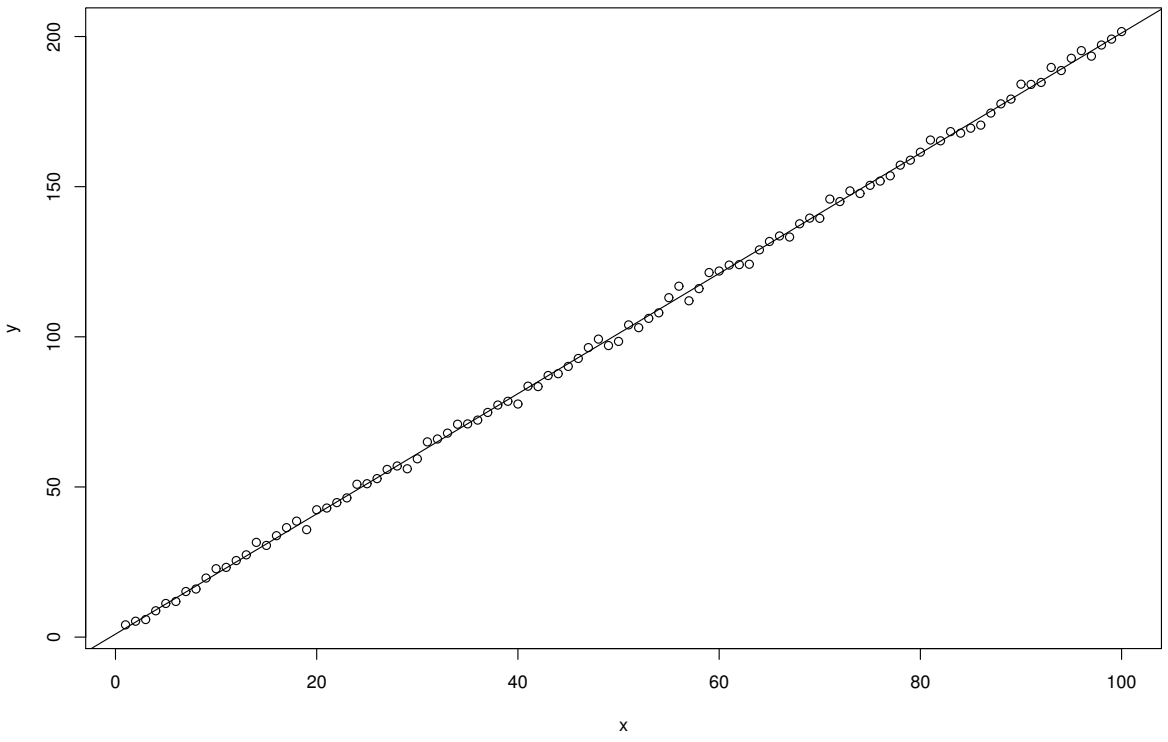
```
F-statistic: 1.515e+05 on 1 and 98 DF,  p-value: < 2.2e-16
> #Das Resultat ist also dasselbe wie von Hand> plot(x,y)
> abline(regr)
>
> #Teilaufgabe c):
> regrc <- lm(x~y,mydata)
> regrc

Call:
lm(formula = x ~ y, data = mydata)

Coefficients:
(Intercept)            y
    -0.4658       0.4993

> beta1hut_c <- sum((x-mean(x))*(y-mean(y)))/sum((y-mean(y))^2)
> beta0hut_c <- mean(x) - beta1hut_c*mean(y)
> beta0hut_c
[1] -0.4657943
> beta1hut_c
[1] 0.4992521
>
> #Zum Vergleich:
> -beta0hut/beta1hut
[1] -0.4987728
> 1/beta1hut
[1] 0.4995752
>
```

Diese R-Session hat noch den foglenden Plot erzeugt:



Aufgabe 50 [Aufspaltung der Variation in den y]
Behauptung: In den Notationen aus 7.1.2 gilt

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2.$$

Beweis: Nach 7.1.2 gilt

$$\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)\;=\;0\;\text{ und }\;\sum_{i=1}^{n}x_i(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)=0. \qquad (*)$$

Daraus schliessen wir

$$\begin{aligned}0&=\hat{\beta}_0\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)+\hat{\beta}_1\sum_{i=1}^{n}x_i(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)\;=\;\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)(\hat{\beta}_0+\hat{\beta}_1x_i)\\&=\sum_{i=1}^{n}(y_i-\hat{y}_i)\hat{y}_i\;=\;\sum_{i=1}^{n}(y_i\hat{y}_i-\hat{y}_i^2).\end{aligned}$$

Diese Gleichung werden wir gleich noch benutzen. Wenn wir noch beachten, dass aus $(*)$ sofort folgt $\sum_{i=1}^{n}\hat{y}_i=\sum_{i=1}^{n}y_i$, können wir nämlich schliessen

$$\begin{aligned}\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2&=\sum_{i=1}^{n}y_i^2+\sum_{i=1}^{n}\overline{y}^2-2\overline{y}\sum_{i=1}^{n}\hat{y}_i+2\sum_{i=1}^{n}\hat{y}_i^2-2\sum_{i=1}^{n}y_i\hat{y}_i\\&=\sum_{i=1}^{n}y_i^2+\sum_{i=1}^{n}\overline{y}^2-2\overline{y}\sum_{i=1}^{n}y_i-2\underbrace{\sum_{i=1}^{n}(y_i\hat{y}_i-\hat{y}_i^2)}_{=0}\\&=\sum_{i=1}^{n}(y_i+\overline{y}-2\overline{y}y_i)\;=\;\sum_{i=1}^{n}(y_i-\overline{y})^2.\end{aligned}$$

Dies beweist natürlich die Behauptung.

∎

Aufgabe 51 [Erwartungstreue Schätzer bei OLS]
Behauptung: Die Schätzer (in der Situation und den Notationen aus der Vorlesung)

$$\hat{\beta}_1\;=\;\frac{SS_{xy}}{SS_{xx}}\;\text{ und }\;\hat{\beta}_0\;=\;\overline{y}-\hat{\beta}_1\overline{x}$$

sind erwartungstreu.

Beweis: Da $E[Y_i]=E[\beta_0+\beta_1x_i+\varepsilon_i]=\beta_0+\beta_1x_i$ und $E[\overline{Y}]=\beta_0+\beta_1\overline{x}$ schliessen wir:

$$\begin{aligned}E[\hat{\beta}_1]=E\left[\frac{SS_{xY}}{SS_{xx}}\right]\;&=\;E\left[\frac{\sum_{i=1}^{n}(x_i-\overline{x})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]\\&=\frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sum_{i=1}^{n}(x_i-\overline{x})(E[Y_i]-E[\overline{Y}])\;=\;\frac{\sum_{i=1}^{n}(x_i-\overline{x})(\beta_0+\beta_1x_i-(\beta_0+\beta_1\overline{x}))}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\\&=\frac{\beta_1\sum_{i=1}^{n}(x_i-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\;=\;\beta_1.\end{aligned}$$

Dies bedeutet genau, dass $\hat{\beta}_1$ ein erwartungstreuer Schätzer von β_1 ist.

Es folgt:

$$E[\hat{\beta}_0]=E[\overline{Y}-\hat{\beta}_1\overline{x}]\;=\;E[\overline{Y}]-E[\hat{\beta}_1]\overline{x}\;=\;\beta_0+\beta_1\overline{x}-\beta_1\overline{x}\;=\;\beta_0.$$

Also ist $\hat{\beta}_0$ ein erwartungstreuer Schätzer von β_0.

■

Aufgabe 52 [$y_i = \beta_0 + \varepsilon_i$]
Wir vereinfachen das Modell von der Gleichung (7.1) auf

$$Y_i = \beta_0 + \varepsilon_i,$$

wobei $\varepsilon_1, \cdots, \varepsilon_n$ iid sind mit $\varepsilon_1 \sim \mathcal{N}(0, \sigma^2)$.

Nun bestimmen bestimmen wir in diesem Modell ein paar Schätzer:

- Zuerst wollen wir den OLS-Schätzer $\hat{\beta}_0^{\text{OLS}}$ von β_0 bestimmen. Die Schätzung $\hat{y}_i^{\text{OLS}}$ von y_i ist dann natürlich einfach durch $\hat{y}^{\text{OLS}} = \hat{\beta}_0^{\text{OLS}}$ gegeben. Dazu müssen wir die Funktion

$$\sum_{i=1}^{n}(y_i - \beta_0)^2 = n\beta_0^2 - 2n\beta_0\overline{y} + \sum_{i=1}^{n} y_i^2 \;=\; n(\beta_0 - \overline{y})^2 + \sum_{i=1}^{n} y_i^2 - \overline{y}^2$$

  bezüglich β_0 minimieren. In obiger Form können wir die Stelle, an der das Minimum angenommen wird, gleich ablesen und erhalten so

$$\hat{\beta}_0^{\text{OLS}} = \overline{y}.$$

- Jetzt möchten wir den ML-Schätzer $\hat{\beta}_0^{\text{ML}}$ von β_0 bestimmen. Dazu müssen wir uns die gemeinsame Dichtefunktion f_{β_0} von $Y_1, \ldots, Y_n$ bestimmen. Da $Y_i = \beta_0 + \varepsilon_i$, hat Y_i klar eine $\mathcal{N}(\beta_0, \sigma^2)$-Verteilung. Es folgt

$$f_{\beta_0}(y_1, \ldots, y_n) \overset{\text{iid}}{=} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \beta_0)^2\right).$$

  Da wir uns bei der ML-Schätzung nur für die Stelle des Maximums (bezüglich β_0) interessieren, können wir gerade so gut den Logarithmus davon betrachten:

$$\log f_{\beta_0}(y_1, \ldots, y_n) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0)^2.$$

  Für die Stelle des Maximums bezüglich β_0 sind die Summanden, die nicht von β_0 abhängen, natürlich irrelevant. Ausserdem spielt der Faktor $\frac{1}{2\sigma^2}$ (> 0) natürlich auch keine Rolle. Daher reicht es wenn wir die Maximumsstelle der Funktion

$$-\sum_{i=1}^{n}(y_i - \beta_0)^2$$

  bestimmen. Bis auf das Vorzeichein ist das genau die gleiche Funktion, wie diejenige die wir für $\hat{\beta}_0^{\text{OLS}}$ minimiert haben. Wegen dem anderen Vorzeichen wird natürlich aus dem Minimum ein Maximum. Wir schliesen also

$$\hat{\beta}_0^{\text{ML}} = \hat{\beta}_0^{\text{OLS}} \;=\; \overline{y}.$$

- Zum Schluss bestimmen wir noch den ML-Schätzer $\hat{\sigma^2}^{\text{ML}}$ von σ^2. Dazu müssen wir $f_{\sigma^2}(y_1, \ldots, y_n)$ bzw. $\log f_{\sigma^2}(y_1, \ldots, y_n)$ bezüglich σ^2 (und β_0) maximieren. Nach obigem gilt:

$$\log f_{\sigma^2}(y_1, \ldots, y_n) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma}\sum_{i=1}^{n}(y_i - \beta_0)^2.$$

  Der Summand, der nicht von σ^2 abhängt spielt hier keine Rolle. Ausserdem haben wir oben gezeigt, dass das Maximum bezüglich β_0 unabhänig von σ^2 an der Stelle $\beta_0 = \overline{y}$ angenommen wird. Aus Gründen der Bequemlichkeit multiplizieren wir die Funktion auch noch mit 2. Konkret müssen wir also noch die Funktion

$$-n\log(\sigma^2) - \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \overline{y})^2 \;=\; -n\log(\sigma) - \frac{S_{yy}}{\sigma^2}$$

bezüglich σ^2 minimieren. Ableiten nach σ^2 liefert:

$$-\frac{n}{\sigma^2} + \frac{S_{yy}}{(\sigma^2)^2}.$$

Ausserdem lautet die zweite Ableitung nach σ^2:

$$\frac{n}{(\sigma^2)^2} - 2\frac{S_{yy}}{(\sigma^2)^3}.$$

Die Nullstelle (bezüglich σ^2) von der ersten Ableitung lautet

$$\sigma^2 = \frac{S_{yy}}{n}.$$

Wenn wir dies in der zweiten Ableitung einsetzen erhalten wir:

$$-\frac{n^3}{S_{yy}} < 0,$$

also handelt es sich um ein Maximum. Somit kommen wir zu dem folgenden Ergebnis:

$$\hat{\sigma^2}^{\mathrm{ML}} = \frac{1}{n} S_{yy} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

Aufgabe 53 [Testen, ob $\beta_1 = 0$, $\beta_1 = \mathrm{PN}$]
Dies ist die Fortsetzung der R-Session von Aufgabe 49:

```
> #Teilaufabe a)
> t <- beta1hut/sqrt(sigma2hut/SSxx)
> pt(abs(t),N-2,lower.tail=F) #P-Wert
[1] 2.124216e-158
>
> #Teilaufgabe b)
> t <- (beta1hut - PN)/sqrt(sigma2hut/SSxx)
> pt(abs(t),N-2,lower.tail=F) #P-Wert
[1] 0.3707982
>
```

Wir haben dabei die Teststatistik aus der Vorlesung benutzt.

Aufgabe 54 [Test, ob $\beta_0 = 0$]
Wir wollen nun in Analogie zu 7.1.3 einen Test entwickeln, ob $\beta_0 = 0$ oder nicht. Dazu betrachen wir die Definition von $\hat{\beta}_0$:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}.$$

Wir wissen bereits aus 7.1.3, dass $\hat{\beta}_1$ eine $\mathcal{N}\left(\beta_1, \frac{\sigma^2}{SS_{xx}}\right)$-Verteilung hat. Ausserdem hat $\overline{Y}$ klar eine $\mathcal{N}(\beta_0 + \beta_1\overline{x}, \frac{\sigma^2}{n})$-Verteilung. Wir schliessen (siehe auch Aufgabe 51)

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{(SS_{xx} + \overline{x}^2 n)\sigma^2}{nS_{xx}}\right).$$

Wie in 7.1.3, benutzen wir jetzt die Schätzung $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ für σ^2. Damit erhalten wir die folgende Teststatistik

$$T = \frac{\hat{\beta}_0}{\sqrt{\frac{(SS_{xx} + \overline{x}^2 n)\hat{\sigma}^2}{nS_{xx}}}},$$

welche unter $\mathcal{H}_0$ (also $\beta_0 = 0$) eine t_{n-2}-Verteilung hat. Wir werden die $\mathcal{H}_0$-Hypothese verwerfen, sobald diese Teststatistik Werte annimmnt, welche weiter als ein bestimmter kritischer Wert von Null entfernt sind.

Bemerkung: In obiger Herleitung haben wir zwei Sachen benutzt, die eigentlich nicht offensichtlich sind:

- Bei der Bestimmung von der Verteilung von $\hat{\beta}_0$ haben wir Unabhängigkeit von $\bar{Y}$ und $\hat{\beta}_1$ benutzt. Denn sie Summe von zwei normalverteilten Zufallsgrössen ist im Allgemeinen nur dann normalverteilt, wenn die beiden Summanden unabhängig sind. Ausserdem haben wir die Unabhängigkeit (oder mindestens Unkorreliertheit) benutzt, um die Varianz zu berechnen.
- Das zweite ist das Einsetzen von $\hat{\sigma}^2$ statt σ^2. Denn auch hier ist a priori die Unabhängigkeit nicht klar und damit auch nicht die t_{n-2}-Verteilung.

Im freiwilligen Teil von Kapitel 7 werden wir diesen Test nochmals in einer etwas allgemeineren Situation sehen. Dort werden sich diese zwei Punkte noch klären.