ASMNW - Lösung 4

Peter von Rohr

2016-05-08

Kontrollfrage 1

Eine Voraussetzung für die Verwendung von Least Squares zur Schätzung der Parameter ist, dass die Matrix $\mathbf{X}$ vollen Kolonnenrang hat. Welche Beziehung resultiert aus dieser Voraussetzung für die Beziehung zwischen den Anzahl Parametern p und die Anzahl Beobachtungen n?

Lösung

Es muss gelten: $p < n$, d.h. die Anzahl Parameter muss kleiner sein als die Anzahl Beobachtungen

Kontrollfrage 2

Abgesehen von Kolonnenrang, wie lauten die fünf weiteren Bedingungen für das Verwenden von multiplen linearen Regressionen

Lösung

1. Lineares Modell ist korrekt $\rightarrow E(\epsilon) = \mathbf{0}$
2. Die Werte in $\mathbf{X}$ sind exakt
3. Die Varianz der Fehler ist konstant ("Homoskedazidität") für alle Beobachtungen $\rightarrow Var(\epsilon) = \mathbf{I} * \sigma^2$
4. Die Fehler sind unkorreliert
5. Weitere Eigenschaften folgen, falls die Fehler normal verteilt sind

Kontrollfrage 3

Falls die Bedingung der konstanten Varianz nicht erfüllt ist, hatten wir gezeigt, dass **Generalised Least Squares** verwendet werden kann. Bei generalised least square nehmen wir an, dass

$$var(\epsilon) = \Sigma$$

Diese Co-Varianzmatrix Σ wird mit der Cholesky-Zerlegung und das Produkt $\Sigma = \mathbf{C}\mathbf{C}^T$ zerlegt. Die Zielgrössen $\mathbf{y}$ und die erklärenden Variablen in $\mathbf{X}$ werden mit der Matrix C^{-1} transformiert. Daraus resultieren dann

$$\begin{aligned} \tilde{\mathbf{y}} &= \mathbf{C}^{-1}\mathbf{y} \\ \tilde{\mathbf{X}} &= \mathbf{C}^{-1}\mathbf{X} \end{aligned} \tag{1}$$

Gegeben unser ursprüngliches lineares Modell

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{2}$$

Wie sieht die Beziehung zwischen den Grössen $\tilde{\mathbf{y}}$ und $\tilde{\mathbf{X}}$ aus, insbesondere handelt es sich bei dieser Beziehung wieder um ein lineares Modell und wenn ja, wie sieht dieses aus?

Lösung

Aufgrund der Transformationen in Gleichung (1) folgt, dass

$$\begin{aligned} \mathbf{C} * \tilde{\mathbf{y}} &= \mathbf{y} \\ \mathbf{C} * \tilde{\mathbf{X}} &= \mathbf{X} \end{aligned} \tag{3}$$

Setzen wir die Beziehungen aus Gleichung (3) in unser lineares Modell (2) ein, dann erhalten wir

$$\mathbf{C} * \tilde{\mathbf{y}} = \mathbf{C} * \tilde{\mathbf{X}}\beta + \epsilon \tag{4}$$

Durch Multiplikation von links beider Seiten in Gleichung (4) mit $\mathbf{C}^{-1}$ erhalten wir

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\beta + \mathbf{C}^{-1}\epsilon \tag{5}$$

Ersetzen wir analog zu $\mathbf{y}$ und $\mathbf{X}$, $\mathbf{C}^{-1}\epsilon$ durch $\tilde{\epsilon}$, dann resultiert wieder ein lineares Modell der Form:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\beta + \tilde{\epsilon}$$

Aufgabe 1

Für unseren Datensatz aus der Vorlesung mit den Zunahmen und dem Geburtsgewicht (siehe http://charlotte-ngs.github.io/GELASM/w12/gain_data.csv) nehmen wir an, dass die Reste ϵ nicht mehr konstant und unkorreliert sind, sondern dass folgende Covarianzmatrix gefunden wurde.

$$\Sigma = \left[\begin{array}{ccccc} 1.44 & 0.12 & 0.08 & 0.09 & 0.12 \\ 0.12 & 2.26 & 0.12 & 0.10 & 0.09 \\ 0.08 & 0.12 & 3.25 & 0.18 & 0.16 \\ 0.09 & 0.10 & 0.18 & 2.91 & 0.14 \\ 0.12 & 0.09 & 0.16 & 0.14 & 2.27 \end{array}\right]$$

Die Matrix ist verfügbar als CSV-Datei unter: http://charlotte-ngs.github.io/GELASM/w12/covar_sigma.csv verfügbar.

Ihre Aufgabe

Schätzen sie die Regressionskoeffizienten unter Berücksichtigung der Covarianzmatrix Σ.

Hinweise

Die Funktion `chol()` macht die Cholesky-Faktorisierung in R. Das Resultat von `chol()` ist eine obere rechte Dreiecksmatrix. Dies entspricht der Matrix $\mathbf{C}^T$ in der Vorlesung. Die Inverse einer Matrix lässt sich mit der Funktion `inv()` berechnen.

Lösung

Als erstes lesen wir die Daten und die Varianz-Kovarianzmatrix von den entsprechenden CSV-Files ein. Das Dataframe mit der Covarianzmatrix konvertieren wir in eine Matrix.

```r
dfGainData <- read.csv2(file = "http://charlotte-ngs.github.io/GELASM/w12/gain_data.csv",
                        stringsAsFactors = FALSE)
dfCovSigma <- read.csv2(file = "http://charlotte-ngs.github.io/GELASM/w12/covar_sigma.csv",
                        stringsAsFactors = FALSE)
mCovSigma <- as.matrix(dfCovSigma)
```

Die Covarianzmatrix wird anhand der Cholesky-Faktorisierung zerlegt.

```r
matC <- t(chol(mCovSigma))
```

Die Inverse von `matC` wird verwendet um die Zielgrösse `PWG` und die erklärenden Variablen zu transformieren.

```r
matCInv <- solve(matC)
vPwgTilde <- matCInv %*% dfGainData[,"PWG"]
vWwgTilde <- matCInv %*% dfGainData[,"WWG"]
vBwTilde <- matCInv %*% dfGainData[,"BW"]
```

Mit den transformierten Grössen können wir das lineare Modell anpassen

```r
dfGainTilde <- data.frame(PWGTilde = vPwgTilde, WWGTilde = vWwgTilde, BWTilde = vBwTilde)
summary(lmGainTilde <- lm(PWGTilde ~ ., data = dfGainTilde))
```

```
##
## Call:
## lm(formula = PWGTilde ~ ., data = dfGainTilde)
##
## Residuals:
##          1          2          3          4          5
##  2.315e-02  7.164e-05  1.812e-02  4.711e-03 -4.606e-02
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.128409   0.114748   1.119    0.379
## WWGTilde    0.990485   0.718601   1.378    0.302
## BWTilde     0.008741   0.022956   0.381    0.740
##
## Residual standard error: 0.03878 on 2 degrees of freedom
## Multiple R-squared:  0.9853, Adjusted R-squared:  0.9707
## F-statistic:  67.2 on 2 and 2 DF,  p-value: 0.01466
```