More Regression

February 18, 2019

The expected value vector and the covariance matrix of jointly distributed random variables

Consider two jointly distributed random variables in the form of a vector $Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$. The bivariate distribution of Z has expected value vector

$$E(Z) = \begin{pmatrix} E(Z_1) \\ E(Z_2) \end{pmatrix}$$

and covariance matrix link

$$V(Z) = E\left[(Z - E(Z))(Z - E(Z))'\right] = E(ZZ') - E(Z)E(Z)' = \begin{pmatrix} \sigma^2_{Z_1} & \sigma_{Z_1,Z_2} \\ \sigma_{Z_1,Z_2} & \sigma^2_{Z_2} \end{pmatrix}$$

The expected value vector and the covariance matrix of jointly distributed random variables

Consider n jointly distributed random variables in the form of a vector $Z = \begin{pmatrix} Z_1 & Z_2 & \cdots & Z_n \end{pmatrix}'$. The multivariate distribution of Z has expected value vector

$$E(Z) = \begin{pmatrix} E(Z_1) \\ E(Z_2) \\ \vdots \\ E(Z_n) \end{pmatrix}$$

and covariance matrix

$$V(Z) = E\left[(Z - E(Z))(Z - E(Z))'\right] = \begin{pmatrix} \sigma^2_{Z_1} & \sigma_{Z_1,Z_2} & \cdots & \sigma_{Z_1,Z_n} \\ \sigma_{Z_2,Z_1} & \sigma^2_{Z_2} & \cdots & \sigma_{Z_2,Z_n} \\ \vdots & \vdots & \ddots & \cdots \\ \sigma_{Z_n,Z_1} & \sigma_{Z_n,Z_2} & \cdots & \sigma^2_{Z_n} \end{pmatrix}$$

The multivariate Normal distribution

The normal distribution has a general multivariate version, where the parameters are given by the expected value vector and the covariance matrix link. For two normal random variables link:

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N(E(Z), V(Z)).$$

The law of large numbers (LLN) and the central limit theorem (CLT) also applies to jointly distributed variables. If we take a random sample from a joint distribution $(Y_i, X_i),\ i = 1, \ldots, n$, then under some regularity conditions:

$$\begin{pmatrix} \bar{Y} \\ \bar{X} \end{pmatrix} \xrightarrow{p} \begin{pmatrix} E(Y) \\ E(X) \end{pmatrix}$$

and

$$\sqrt{n} \begin{pmatrix} \bar{Y} - \mu_Y \\ \bar{X} - \mu_X \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \sigma_{Y,X} \\ \sigma_{Y,X} & \sigma_X^2 \end{pmatrix} \right)$$

Homoskedasticity

In the small sample case we didn't only assume normality we also assumed $V(U|X) = \sigma^2$, this is called the homoskedasticity assumption **link to one example**. This assumption is very strong and in large samples this is not necessary.

Heteroskedasticity

If we don't impose the homoskedasticity assumption we allow for heteroskedasticity, i.e. that the variance of U may depend on X **link to one example**. In this case we have to estimate $V(\hat{\beta})$ to be able to perform inference about β.

link with both skedasticities

Heteroskedastic and auto-correlation robust inference

It is actually possible to estimate $V(\hat{\beta})$ when there is some dependence between $U_i, i = 1, \ldots, n$, but this is beyond the scope of the course.

Robust Hypothesis testing in large samples

Hypothesis in large samples can be conducted with the following T-statistic

$$T = \frac{(\hat{\beta}_k - \beta_k)}{\sqrt{\left(\hat{V}(\hat{\beta})\right)_{(k+1),(k+1)}/n}} \xrightarrow{d} N(0,1)$$

where we will use R to obtain $\hat{V}(\hat{\beta})$.

In large samples heteroskedastic-robust standard errors are correct even under homoskedasticity. But the opposite is not true.

Heteroskedastic robust T-testing in R

```
library(car)
data(Prestige)
#Load package for testing of the regression model
library(lmtest)
#Load package for robust covariance matrices
library(sandwich)
#Run regression
reg<-lm(prestige~income+women, data=Prestige)
#Non-robust results
summary(reg)
# Heteroskedasticity-robust estimate of V(beta_h)
V_b<-vcovHC(reg)
#Robust results
coeftest(reg,vcov=V_b)
```

Joint hypotheses

We want to test a joint hypothesis, e.g.

$$H_0 : \beta_1 = 0 \ \& \ \beta_2 = 0, \quad H_a : \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0$$

Consider

$$R = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \ \& \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \text{ then } R\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

then we can formulate the joint null hypothesis like

$$H_0 : R\beta = r, \text{ where } r = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Joint hypothesis testing in small samples

$$H_0 : R\beta = r \;\&\; H_a : \neg R\beta = r$$

If H_0 is true and $U_i \sim iidN(0, \sigma^2)$ (the latter is equivalent to assuming $U|X \sim N(0, \sigma^2)$ and you have a simple random sample), then it is possible to show

$$F = \frac{\left[(R\hat{\beta} - r)\right]' \left[R(\mathbf{X}'\mathbf{X})^{-1}R'\right]^{-1} \left[(R\hat{\beta} - r)\right] / Q}{S^2} \sim F_{Q,n-K-1}$$

where

- K is the number of regressors (X-variables)
- Q is the number of restrictions given by H_0; the number of rows of R

We can compute $\text{P-value} = P(F > \text{F-value})$ based on the F-distribution ($F_{Q,n-K-1}$) to test the joint hypothesis.

Joint hypothesis in large samples

By the CLT

$$\sqrt{n}(\hat{\beta} - \beta) \overset{a.}{\sim} N(\mathbf{0}, V(\hat{\beta}))$$

given this and that H_0 is true it is possible to show:

$$W = \left[\sqrt{n}(R\hat{\beta} - r)\right]' \left[RV(\hat{\beta})R'\right]^{-1} \left[\sqrt{n}(R\hat{\beta} - r)\right] \overset{a.}{\sim} \chi^2_Q.$$

With a large sample size we can estimate $V(\hat{\beta})$ such that

$$W = \left[\sqrt{n}(R\hat{\beta} - r)\right]' \left[R\hat{V}(\hat{\beta})R'\right]^{-1} \left[\sqrt{n}(R\hat{\beta} - r)\right] \overset{a.}{\sim} \chi^2_Q.$$

Robust Joint hypothesis in large samples: Wald test

Thus, in large samples we can use

$$W = n\left[(R\hat{\beta} - r)\right]' \left[R\hat{V}(\hat{\beta})R'\right]^{-1} \left[(R\hat{\beta} - r)\right] \overset{a.}{\sim} \chi^2_Q$$

as test statistic, it is valid both under homo- and heteroskedasticity and under non-normaly distributed U, in large samples. The statistic is often called Wald-statistic.

χ^2_Q is the chi-square distribution with parameter (degrees of freedom) Q. It is used when calculating P-values.

Standard Joint-test for regression

Consider

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$$

The standard joint test for regression analysis is the following:

$$H_0 : \beta_1 = 0 \ \& \ \beta_2 = 0, \quad H_a : \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0.$$

i.e. have at least one of the explanatory variables a significant effect on Y?

- Why not just look att the two t-tests for β_1 and β_2 and reject H_0 if at least one of them has $\text{P-value} < 0.05$?
  - Since the probability to reject a true joint H_0 generally is larger than 0.05 if you do this
  - If you increase the number of X-variables the probability to reject a true H_0 increases with the number of variables
  - With the joint test, if you reject when $\text{P-value} < 0.05$, the probability to reject a true H_0 will be 0.05 and nothing else

Example: Standard Joint-test for regression

```
#Continue with the Prestige data
#Run regression
reg<-lm(prestige~income+women, data=Prestige)
# The matrices for the joint hypothesis
#H_0: beta_1=beta_2=0
R<-rbind(c(0,1,0),c(0,0,1)); r<-c(0,0)

#The function linearHypothesis() is included in car
#The F-test you get by summary(reg)
linearHypothesis(reg, hypothesis.matrix=R, rhs=r)

# Heteroskedasticity-robust estimate of V(beta_h)
V_b<-vcovHC(reg)
#The heteroskedasticity-robust Joint Wald-test
linearHypothesis(reg, hypothesis.matrix=R, rhs=r,
test=c("Chisq"),vcov.=V_b)
```

Nonlinear modeling: Polynomials

Consider

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_r X^r + U$$

We can test if the function is linear

$$H_0 : \ \beta_2 = \beta_3 = \ldots = \beta_r = 0 \text{ vs.}$$

$$H_1 : \ \text{At least one } \beta_j \neq 0, \ j = 2, 3, \ldots, r$$

with help of a F-test or a Wald-test.

Exercise: Linear versus non-linear function

Use the *Prestige* data in the car package and compute the Wald-test for the joint hypothesis

$$H_0: \ \beta_4 = \beta_5 = 0 \text{ vs. } H_a: \ \text{At least one } \beta_j \neq 0, \ j = 4, 5$$

$$\begin{aligned} \text{prestige} = \beta_0 + \beta_1 \mathit{income} + \beta_2 \mathit{women} + \\ \beta_3 \mathit{education} + \beta_4 \mathit{education}^2 + \beta_5 \mathit{education}^3 + U \end{aligned}$$

Why is the null-hypothesis a hypothesis for linearity?

Other nonlinear models

The following models are example of non-linear specifications which can be modelled linearly

- $Y = exp(X\beta + U)$
- $Y = \prod_{k=1}^{K} X_k^{\beta_k} exp(\beta_0 + U)$
- $Y = (X\beta + U)^2$

given the following transformations

- $log(Y) = X\beta + U$
- $log(Y) = \beta_0 + \sum_{k=1}^{K} log(X_k)\beta_k + U$
- $\sqrt{Y} = X\beta + U$

where $X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$

R^2

Defined exactly like for the single regressor case

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{ESS}{TSS}$$

which also can be written as

$$1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{RSS}{TSS}.$$

- Important remark
  - If we add an extra regressor, X_+, RSS decreases, while TSS is unchanged

$$\Rightarrow R^2 \text{ increases}$$

Adjusted R^2

- The R^2 value will be larger for a model with 1000 regressors than a model with e.g. five.
- Is the model with 1000 regressors a better model?
  - Probably not. If we want to say something relevant about real-life it is difficult to have such a large model.
  - A model should be a simplification of reality that still explains the most important mechanisms.
- The adjusted R^2 ($\bar{R}^2$) is a penalization of R^2

$$\bar{R}^2 = 1 - \frac{n-1}{n-K-1}\frac{RSS}{TSS}$$

Compare $\bar{R}^2$ and R^2

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$\bar{R}^2 = 1 - \frac{n-1}{n-K-1}\frac{RSS}{TSS}$$

An additional regressor increases $\frac{n-1}{n-K-1} \Rightarrow \bar{R}^2 \downarrow$ and RSS decreases $\Rightarrow \bar{R}^2 \uparrow$ and $R^2 \uparrow$

Maximize the $\bar{R}^2$ value?

NO,

- We should have some knowledge of the research topic and be able to put up a model based theory and/or experience, and
- try to obtain a neat model that catches the main mechanisms we are interested in.
- However, there will always be secondary variables which we don't know if we should include or not and in this case a measure as $\bar{R}^2$ could give guidance.

Multicolinearity

Perfect multicolinearity

If one of the regressors is a linear combination of the other regressors we cannot compute the least-squares estimator. Consider

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U$$

if

$$X_1 = X_2 + 0.5 X_3$$

then

$$Y = \beta_0 + (\beta_1 + \beta_2) X_2 + (0.5\beta_1 + \beta_3) X_3 + U$$

and we can only identify $(\beta_1 + \beta_2)$ and $(0.5\beta_1 + \beta_3)$ not β_1, β_2 and β_3 separately. This make some intuitive sense and the mathematical definition for this problem is exact: $\mathbf{X}'\mathbf{X}$ is not invertible in cases with perfect multicollinearity ($\mathbf{X}'\mathbf{X}$ is 'singular')

Perfect multicollinearity and dummy variables

Example

$$D_1 = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{if male} \\ 0 & \text{otherwise} \end{cases}$$

$D_1 = 1 - D_2$ is a linear combination, we can only include one of D_1 and D_2 in the regression!

Imperfect multicollinearity

$$0 < \text{corr}(X_i, X_j) < 1$$

Imperfect multicollinearity does not disable computation, $\hat{\beta}$ is still unbiased, consistent and asymptotically normally distributed.

- Problem
  - When the multicollinearity is strong the variances of the beta-estimators are large.
- Intuition
  - When $\text{corr}(X_1, X_2)$ is large it is difficult to distinguish the unique effects of X_1 and X_2

Example imperfect multicollinearity

```
## Generate some data
   n<-50
   set.seed(7)
   X1<-runif(n,-5,5); X2<--0.5*X1+runif(n)
   U<-rnorm(n,0,sd=sqrt(1.5))
   Y<-0.5-1*X1+2*X2+U
   # Use lm to estimate
   reg<-lm(Y~X1+X2)
   summary(reg)
```

Remedies of Imperfect multicollinearity

Imperfect multicollinearity does not disable computation but causes large variances and standard errors of the estimators of the beta-coefficients

Some remedies are

- Increase sample size
- Are all regressors necessary, can you remove one X without the risk of 'omitted variable bias' in $\hat{\beta}$, i.e. without violating $E(U|X) = E(U)$?
- Use estimators with smaller variance than the LS estimator (e.g. the ridge estimator; not in this course)

Exercise: Continue example imperfect multicollinearity

1. If you remove X_1 and run the regression only with X_2 as regressor, what happens with the estimate of $\hat{\beta}_2$?
2. Do you think one can remove X_1 without violating $E(U|X_2) = E(U)$?
3. Assume we can obtain more data. Use the code from the previous example and set $n = 100$ and estimate again. Is X_1 significant now?

Outliers

- In small samples the results are affected a lot if there are outliers
  - Also asymptotic results can be affected by extreme outliers

Cook's Distance

- The Cook's distance is a way to identify influential observations in R
- Cook's distance is an F-statistic-like measure derived from the null hypotheses

$$H_0 : \hat{\beta} = \hat{\beta}_{-i}$$

for each observation, where $\hat{\beta}_{-i}$ is the estimator where the i^{th} observation has been removed. The distance is defined as follows

$$D_i = \frac{(\hat{\beta}_{-i} - \hat{\beta})' X' X (\hat{\beta}_{-i} - \hat{\beta})}{(K+1)S^2}$$

- The statistics is relatively intuitive: if the estimates doesn't change much when the observation is deleted D_i is small ($(\hat{\beta}_{-i} - \hat{\beta}) \approx \mathbf{0}$).
- Even in large samples the Cook distance doesn't converge to an F-distribution or any other limit distribution
  - It converges to zero

Cook's Distance and removal of outliers

- Since the Cook distance doesn't have a known distribution it is difficult to set an appropriate threshold for outliers
- One can try to remove the largest outliers and see if it affects the results
  - The estimates
  - the significant results
- If not, I would keep all "outliers"

Example: Diagnostics for Outliers

```
set.seed(7); n<-15
#I generate some X-varibles and construct some
#correlation between them
x1<-rexp(n);x2<-runif(n)+0.5*x1;x3<-runif(n)+x2-x1
#I generate Y through a linear model
Y<-0.5+2*x1+2*x2-2*x3+rchisq(n,df=2)#rnorm(n,0,1)
#I run the regression
reg2<-lm(Y~x1+x2+x3)
#Create graphical diagnostics
par(mfrow=c(1,2))
   plot(fitted(reg2),resid(reg2),xlab="Fitted values",
   ylab="Residuals")
plot(reg2,which=4)
```

Example: Remove Outlier (difficult in small samples)

According to the Cook's distances there seem to be one observation that has more influence than the other ones (Nr 7)

```
    data<-data.frame(Y,x1,x2,x3)
    data_sub<-data[-7,]
    reg3<-lm(Y~x1+x2+x3,data=data_sub)
par(mfrow=c(1,2))
    plot(fitted(reg3),resid(reg3),xlab="Fitted values",
    ylab="Residuals")
plot(reg3,which=4)
summary(reg2)
summary(reg3)
```