Please read the discussion in your textbook. Be aware of the following points:

- Endogeneity occurs when there is a feedback relationship between one or more of the independent variables and the dependent variable.
- Endogeneity causes the OLS-estimated coefficients and standard errors to be biased.

Endogeneity was one of the first problems recognized in econometrics. Early attempts to estimate demand functions produced puzzling results: price either did not affect quantity, or did not affect it in the predicted way. It was then realized that the raw data only presented equilibrium quantity and price, and not a demand curve. Two equations must be estimated simultaneously--supply and demand--in order to convert equilibrium quantities and prices into a demand function.

One way to look at this is that price is determined by quantity, and quantity by price. Both are *endogenous*: that is, a change in the disturbance term will not only change quantity (the dependent variable) but also price (one of the independent variables).

The trick in such a case is to make price (the endogenous independent variable) *exogenous*: that is, make it independent of the disturbance term that influences quantity (the dependent variable). This trick is easily performed by two stage least squares.

In this exercise you will use data (*s:/teff/662/R/cop.dbf*) on world copper prices and quantities to estimate the demand function for copper. The variables are as follows:

PC	World price of copper
QC	World quantity of copper sold
PA	World price of Aluminum
Y	Income
X	World stocks of copper
YEAR	Time trend (proxies technological progress)

Start out by trying to estimate the demand and supply functions for copper. Be able to tell me why the independent variables belong in each equation, and what sign you would expect the coefficients to have.

```
library(foreign)
uu<-read.dbf("s:/teff/662/R/cop.dbf")

zD<-lm(QC~PC+Y+PA,data=uu)
summary(zD)
zS<-lm(QC~PC+X+YEAR,data=uu)
summary(zS)
```

Testing for endogeneity:

The most common test for endogeneity is the **Hausman test**. The following lines in R will perform the test.

```
#--Testing for endogeneity: Hausman test--
#--Regression on PC w/ Exog.regressors--
zP<-lm(PC~PA+X+Y+YEAR,data=uu)
PCfit<-zP$fitted.values
PCres<-zP$residuals

#--Add PCres to RHS of D function model--
zDh<-lm(QC~PC+Y+PA+PCres,data=uu)
#-- t-stat for PCres. H0: NO endogeneity--
summary(zDh)
```

What did you do? Essentially, the following two things:

1. Made price exogenous by creating a new price variable (PCfit). This variable is formed by regressing all the exogenous variables on price, and taking the fitted value as the new variable. The individual exogenous variables (as well as the new fitted value) are called *instrumental variables*, or *instruments*. Since the new variable is created from exogenous variables, it should not be correlated with the disturbance term, and can be considered exogenous.

2. Ran a regression for the demand function in which you included the residual left over when you created the new (exogenous) price variable. The Hausman test is simple: look at the t-statistic on the residual's coefficient. Your null hypothesis is that there is no endogeneity. If the t-statistic is high enough, you can reject the null hypothesis.

Correcting for Endogeneity:

Two stage least squares is the easiest way to correct for endogeneity. The procedure involves the following two steps:

1. Make any endogenous independent variable exogenous by regressing it against all the exogenous variables in the model (you did this above, creating PCfit).

2. Replace the endogenous independent variable with the new (exogenous) variable. Now when you run your model the parameters should not be biased (the standard errors and R^2 *will* be biased, but this is easily corrected).

```
#--Restimate, replacing PC with PCfit--
zDiv<-lm(QC~PCfit+Y+PA,data=uu)
summary(zDiv)
zSiv<-lm(QC~PCfit+X+YEAR,data=uu)
summary(zSiv)
```

The R command *ivreg* in the package *AER* will do all this automatically (giving the corrected standard errors and R^2). Run the following lines:

```
ivD<-ivreg(QC~PC+Y+PA|Y+PA+X+YEAR,data=uu)
summary(ivD)
ivS<-ivreg(QC~PC+X+YEAR|Y+PA+X+YEAR,data=uu)
summary(ivS)
```

A few questions that always come up when dealing with endogeneity, as well as some partial answers:

Q: How do I know which variables are endogenous?

A: It is not necessary to test every variable. Start from theory and plausible deduction, and select those which appear most likely to be involved in a feedback relationship with the dependent variable. Test these, using the Hausman test.

Q: When creating instrumental variables, where do I get my exogenous variables?

A: First, take all the exogenous variables in the model. Then add others. In most cases, the easiest variables to add are temporally lagged values of the endogenous variables (it makes sense that these lagged values would be exogenous, doesn't it?).

Q: How many exogenous variables do I need to create an instrumental variable?

A: It is possible to have too few; having too many is not a serious problem. Count the number of instrumental variables you have put in your equation as replacements for endogenous variables (call this number I). Then count the number of other exogenous variables appearing in your equation (call this number E). In creating your instruments, you will need at least I+E-1 exogenous variables.

Q: How can I tell if my instrumental variable is a satisfactory exogenous replacement for an endogenous variable?

A: You should think carefully about the exogenous variables you use to create your instrumental variable. Do they act as near proxies for the endogenous variable? Take a look at the R-squared from the regression creating your instruments; it should be reasonably high, let's say above 0.50.

HOMEWORK

1) In three weeks, you must bring to class a first draft of your paper. I will read it and give you some feedback the day of the exam.

2) Redo *one* of the following homework projects:

   a) County-level U.S. Production Function, considering that KAP probably exhibits endogeneity.

   b) U.S. national consumption function, considering that income is probably endogeneous.

   c) The school district test score data, considering that some independent variables might be endogenous.

   Don't do too much work: simply do the Hausman tests, replace the endogenous variable with an instrument when necessary, and note in your write-up how your results changed.

```
#--Monte Carlo: Endogeneity ("S:/teff/662/R/r09mc.R")--
rm(list=ls(all=TRUE))
#--Set path to your own directory--
setwd("S:/teff/662/R/")
options(echo=TRUE)
library(foreign)
library(AER)
library(gplots)

estcoef<-NULL
corvals<-NULL
nobs<-50
truecoef<-6
for (i in 1:5000){
err<-33*rnorm(nobs)
x<-33*rnorm(nobs)
q<-(err+x)/2
q<-33*scale(q)
x<-33*scale(x)
y1<-x*truecoef+err
y2<-q*truecoef+err
qe<-cbind(q,err)
xe<-cbind(x,err)
corvals<-rbind(corvals,cbind(cor(xe)[1,2],cor(qe)[1,2]))
z1<-summary(lm(y1~x))
cf1<-z1$coefficients[2,1]
se1<-z1$coefficients[2,2]
z2<-summary(lm(y2~q))
cf2<-z2$coefficients[2,1]
se2<-z2$coefficients[2,2]
estcoef<-rbind(estcoef,cbind(cf1,se1,cf2,se2))
}
estcoef<-data.frame(estcoef)
names(estcoef)<-c("ExogBeta","ExogSE","EndogBeta","EndogSE")

#--take mean and standard deviation of results--
#--compare mean of SE with SD of estimated coefficient (should be equal)--
round(apply(estcoef,2,mean),6)
round(apply(estcoef,2,sd),6)

#--correlation of variable with error term--
apply(corvals,2,range)

#--plot histograms of results: cor(et,et-1)--
layout(matrix(1:2,1,2))
hist(corvals[,1],breaks=50,xlim=c(-1,1),main="Exogenous")
abline(v=0,col="red",lty=2,lwd=2)
hist(corvals[,2],breaks=50,xlim=c(-1,1),main="Endogenous")
abline(v=0,col="red",lty=2,lwd=2)
layout(1)

#--plot histograms of results: estimated coefficients and standard errors--
layout(matrix(1:4,2,2,byrow=TRUE))
xlm<-range(estcoef[,c("ExogBeta","EndogBeta")])
hist(estcoef$ExogBeta,breaks=50,main="Exogenous",xlim=xlm,xlab="Coefficients; true: red; mean:
blue")
abline(v=truecoef,col="red",lty=2,lwd=2)
abline(v=mean(estcoef$ExogBeta),col="blue",lty=3,lwd=2)
hist(estcoef$EndogBeta,breaks=50,main="Endogenous",xlim=xlm,xlab="Coefficients; true: red;
mean: blue")
abline(v=truecoef,col="red",lty=2,lwd=2)
abline(v=mean(estcoef$EndogBeta),col="blue",lty=3,lwd=2)
xlm<-range(estcoef[,c("ExogSE","EndogSE")])
hist(estcoef$ExogSE,breaks=30,main="Exogenous",xlim=xlm,xlab="Standard Errors; true: red;
mean: blue")
abline(v=sd(estcoef$ExogBeta),col="red",lty=2,lwd=2)
abline(v=mean(estcoef$ExogSE),col="blue",lty=3,lwd=2)
hist(estcoef$EndogSE,breaks=30,main="Endogenous",xlim=xlm,xlab="Standard Errors; true: red;
mean: blue")
abline(v=sd(estcoef$EndogBeta),col="red",lty=2,lwd=2)
abline(v=mean(estcoef$EndogSE),col="blue",lty=3,lwd=2)
layout(1)
```