SCRIPT MOD1_1B: ASSUMPTIONS AND ROBUSTNESS, PART II

Set basic R-options upfront and load all required R packages:

```
> rm(list = ls(all = TRUE))#first, clear R's workspace
> options(prompt = "R> ", digits = 4)
R> options(continue=" ") #remove continuation prompt
R> setwd('C:/Klaus/AAEC5126/module1/')#R Sweaves to the default directory
R> options(width=60)#so R-chunks don't run over margin
R> set.seed(37)  #sets the random number generator so we can reproduce results
R> tic<-proc.time() #start stop watch
R> library("xtable")
```

1. Empirial part

1.1. Illustration of the role of assumptions in econometrics using simulated data - part B.

Now let's assume that hourly wages in your general population of interest follow a the same distribution for female and male workers with mean \$30 and std \$5. Lets' generate 1000 draws each from these distributions.

```
R> n<-1000 #number of draws
R> # female parameters
R> fmean<-30; #female mean
R> fstd<-5;#female standard deviation
R> fy<-matrix(rnorm(n,fmean,fstd),n)
R> #draw n observations for this normal density and place into an nx1 vector
R>
R> # male parameters
R> mmean<-30 #male mean
R> mstd<-5;#male standard deviation
R> my<-matrix(rnorm(n,mmean,mstd),n)
R> #draw n observations for this normal density and place into an nx1 vector
R>
```

Plot the two samples using kernel densities (think of it a s a smooth histogram). First, we need to generate density estimates for each data point.

```
R> ally<-rbind(fy,my)   #combine both data chunks for "naive" analysis below
R> fdens<-density(fy,kernel="epanechnikov",n=1000)
R> #get kernel density estimates, using Epanechnikov method
R> # and 1000 evaluation points
R> # note: this "n" is not our sample size from above,
R> # and only used internally by the "density" function
R> mdens<-density(my,kernel="epanechnikov",n=1000)
R> alldens<-density(ally,kernel="epanechnikov",n=2000)
R>
```

Now for the plot:

```
R> plot(fdens,type="l",main = "",xlab = "hourly wage ($)",ylab = "density",
 xlim=c(min(ally),max(ally)),ylim=c(0,0.15),lwd=2)
R> #main plotting command, start with female density plot
R> lines(mdens,col=2,lty=2,lwd=2)
R> #add male density with different line type & color
R> lines(alldens,col=3,lty=4,lwd=3)   # add combined density
R> labels<-c("female","male","all")
R> legend("topright", inset=.05,
 labels, lwd=1, lty=c(1,2,4), col=c(1,2,3))
R>
```

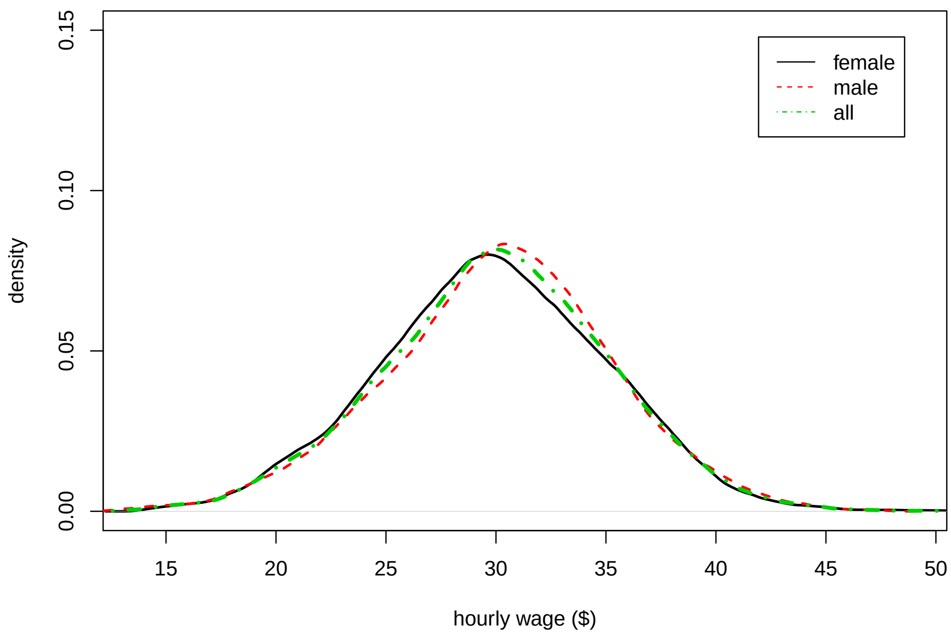

Figure 1. Wage plots

Now let's generate a table with basic descriptive statistics:

```
R> tt<-data.frame(col1=c("female","male","all"),
                 col2=c(mean(fy),mean(my),mean(ally)),
                 col3=c(sd(fy),sd(my),sd(ally)),
                 col4=c(min(fy),min(my),min(ally)),
                 col5=c(max(fy),max(my),max(ally)))
R> colnames(tt)<-c("population","mean","std","min","max")
R>
```

```
R> print(xtable(tt,caption="summary statistics for female, male,
 and overall sample"),include.rownames=FALSE,
 latex.environment="center", caption.placement="top",table.placement="!h")
```

TABLE 1. summary statistics for female, male, and overall sample

population	mean	std	min	max
female	29.91	5.08	15.69	49.07
male	30.17	5.09	13.56	45.90
all	30.04	5.09	13.56	49.07

```
R> #get rid of row counters, and center over decimal)
R>
```

1.1.1. *Estimation based on least general (most restrictive) assumptions.* Implicit assumptions:

(1) Both gender groups share the same expectation (mean) for wage (CORRECT)
(2) Both share the same standard deviation for wage (CORRECT)
(3) Wage follows a common normal density for both groups (CORRECT)

Estimation via OLS

```
R> y<-rbind(fy,my)
R> #stick both samples into one long vector - your dependent variable
R> X<-matrix(rep(1,length(y)));
R> #we just want to estimate the mean, so the only explanantory
R> # variable is a vector of ones
R> k<-ncol(X)
R> bols<-solve(((t(X))%*%X),(t(X)%*%y));# compute OLS estimator
R> e<-y-X%*%bols # Get residuals.
R> s2<-(t(e)%*%e)/(2*n-k) #get the regression error (estimated variance of "eps").
R> # note the "2*n" - that's our total sample size after combining male and female data
R> Vb<-s2[1,1]*solve((t(X))%*%X)
R> # get the estimated variance-covariance matrix of bols
R> se=sqrt(diag(Vb)) # get the standard erros for your coefficients;
R> tval=bols/se # get your t-values.
R>
```

Display results in a nice table:

```
R> tt<-data.frame(col1="constant",
                 col2=bols,
                 col3=se,
                 col4=tval)
R> colnames(tt)<-c("variable","estimate","s.e.","t")
R>
R> print(xtable(tt,caption="OLS output, Model 1"),include.rownames=FALSE,
 latex.environment="center", caption.placement="top",table.placement="!h")
```

Table 2. OLS output, Model 1

variable	estimate	s.e.	t
constant	30.04	0.11	264.18

```
R> #get rid of row counters, and center over decimal)
R>
```

The estimated standard deviation of the regression error is 5.09.

1.1.2. *Estimation based on milder (= more general) assumptions.* Implicit assumptions:

(1) Population means differ between the two groups (WRONG)
(2) Both share the same standard deviation for wage (CORRECT)
(3) Wage follows a separate normal density for each group (WRONG)

Estimation via OLS

```
R> y<-rbind(fy,my) #stick both samples into one long vector -
R> # your dependent variable
R> female<-matrix(c(rep(1,n),rep(0,n)))
R> male<-matrix(1-female)
R> X<-cbind(female,male)
R> k<-ncol(X)
R> bols<-solve(((t(X))%*%X),(t(X)%*%y));# compute OLS estimator
R> e<-y-X%*%bols # Get residuals.
R> s2<-(t(e)%*%e)/(2*n-k) #get the regression error (estimated variance of "eps").
R> # note the "2*n" - that's our total sample size after combining male and female data
R> Vb<-s2[1,1]*solve((t(X))%*%X)
R> # get the estimated variance-covariance matrix of bols
R> se=sqrt(diag(Vb)) # get the standard erros for your coefficients;
R> tval=bols/se # get your t-values.
R>
R> tt<-data.frame(col1=c("female","male"),
                col2=bols,
                col3=se,
                col4=tval)
R> colnames(tt)<-c("variable","estimate","s.e.","t")
R>
R> print(xtable(tt,caption="OLS output, Model 2"),include.rownames=FALSE,
 latex.environment="center", caption.placement="top",table.placement="!h")
R> #get rid of row counters, and center over decimal)
R>
```

The estimated standard deviation of the regression error is 5.09.

Table 3. OLS output, Model 2

variable	estimate	s.e.	t
female	29.91	0.16	185.99
male	30.17	0.16	187.65

1.1.3. *Estimation based on even more general assumptions.* Implicit assumptions:

(1) Population means differ between the two groups (WRONG)
(2) Standard deviations differ between the two groups (WRONG)
(3) Wage follows a separate normal density for each group (WRONG)

Estimation via FGLS

```
R> #capture residuls from previous OLS and derive estimates
R> #for group-specific variances
R> e<-y-X%*%bols
R> ef<-e[1:n] #female residuals
R> em=e[(n+1):length(e)] #male residuals
R> sig2f<-((t(ef))%*%ef)/n #estimate for female variance
R> sig2m=(t(em)%*%em)/n  #estimate for male variance
R> Om<-diag(c(sig2f[1,1]*rep(1,n),sig2m[1,1]*rep(1,n)))
R> #compose variance-covariance matrix
R> # for 2nd stage regression error
R>
R> bgls<-solve(((t(X))%*%solve(Om)%*%X),((t(X))%*%solve(Om)%*%y))
R> Vb<-solve((t(X))%*%solve(Om)%*%X);
R> # get the estimated variance-covariance matrix of bgls
R> se<-sqrt(diag(Vb)) # get the standard errors for your coefficients;
R> tval<-bols/se # get your t-values.
R>
R> tt<-data.frame(col1=c("female","male"),
                 col2=bols,
                 col3=se,
                 col4=tval)
R> colnames(tt)<-c("variable","estimate","s.e.","t")
R>
R> print(xtable(tt,caption="FGLS output, Model 3"),include.rownames=FALSE,
 latex.environment="center", caption.placement="top",table.placement="!h")
```

Table 4. FGLS output, Model 3

variable	estimate	s.e.	t
female	29.91	0.16	186.16
male	30.17	0.16	187.67

```
R> #get rid of row counters, and center over decimal)
R>
```

Estimated standard deviation of the regression error for female: 5.08.
Estimated standard deviation of the regression error for male: 5.08.

```
R> proc.time()-tic
   user  system elapsed 
   2.68    0.20    2.89 
```