SCRIPT MOD5S1: TREATMENT EFFECTS VIA REGRESSION: JOB TRAINING APPLICATION

1. Load and describe data

This script uses the job training data from **?**, as described in **?**, p. 928. The data are labeled "jtrain2" and comprise 445 observations on male workers, 118 of whom underwent some job training in the late 1970s. The outcome of interest are real earnings in 1978 (training occurred in 1975-1977). The data set is sorted by treatment (treated first, then untreated), and includes the following variables:

(1) train =1 if assigned to job training
(2) age =age in 1977
(3) educ =years of education
(4) black =1 if black
(5) hisp =1 if Hispanic
(6) married =1 if married
(7) nodegree =1 if no high school degree
(8) mosinex =No. months prior to Jan. 78 in experiment
(9) re74 =real earnings, 1974, $1000s
(10) re75 =real earnings, 1975, $1000s
(11) re78 =real earnings, 1978, $1000s
(12) unem74 =1 if unemployed all of 1974
(13) unem75 =1 if unemployed all of 1975
(14) unem78 =1 if unemployed all of 1978
(15) lre74 =log(re74); zero if re74 == 0
(16) lre75 =log(re75); zero if re75 == 0
(17) lre78 =log(re78); zero if re78 == 0
(18) agesq =age squared
(19) mostrn =months in training

Our goal will be to replicate the results reported in Wooldridge (2010), Table 21.1, p. 929 for the "jtrain2" sample. Because of rounding errors and (I suspect) subtle differences in STATA's and R's "inverse" functions, our standard errors will be a bit different from those given in Wooldridge's table.

The following loads in the data (including all variable names), and saves it immediately in R's internal ("rda") format:

```
R> data<- read.table('c:/Klaus/AAEC5126/R/data/jtrain2.txt',
 sep="\t", header=TRUE)
R> save(data, file = "c:/Klaus/AAEC5126/R/data/jtrain2.rda")
```

2. Checking for overlap

Assignment to treatment in the "jtrain2" data is random, so ignorability automatically holds. This also implies that $\tau_{ate} = \tau_{att}$ in the population. Estimates for the two effects may be differ, depending on estimation method. This is largely related to the overlap of regressors, which is somewhat problematic in this example.

We will check on overlap using the *normalized differences* suggested by **?**. The authors suggest that any normalized differences above 0.25 are worrisome, suggesting limited overlap across the two groups.

```
R> attach(data)
R> n<-nrow(data)
R> n1<-185
R> n0<-n-n1 #260 cases
R> y<-re78
R> X<-cbind(age,educ,black,hisp,married,re74,re75)
R> #
R> y1<-y[1:n1]
R> y0<-y[(n1+1):n]
R> my1<-mean(y1)
R> my0<-mean(y0)
R> sy1<-sd(y1)
R> sy0<-sd(y0)
R> ndiffy<-(my1-my0)/sqrt(sy1^2 + sy0^2)
R> #
R> X1<-X[1:n1,]
R> X0<-X[(n1+1):n,]
R> mX1<-colMeans(X1)
R> mX0<-colMeans(X0)
R> sX1<-apply(X1,2,sd)
R> sX0<-apply(X0,2,sd)
R> ndiffX<-as.vector((mX1-mX0)/sqrt(sX1^2 + sX0^2))
R> #
R> tt<-data.frame(col1=c("re78","age","educ","black","hisp",
       "married","re74","re75"),
 col2=c(ndiffy,ndiffX))
R> colnames(tt)<-c("variable","norm.diff")
```

All overlap scores are well below 0.25, so we can expect for ATE and ATT to be well-identified, using these explanatory variables.

3. Estimation via difference in means

```
R> m1<-(my1-my0)
R> sem1<-sqrt(sy1^2/(n1)+sy0^2/(n0))
R> tm1<-m1/sem1
```

Table 1. normalized differences

variable	norm.diff
re78	0.19
age	0.08
educ	0.10
black	0.03
hisp	-0.12
married	0.07
re74	-0.00
re75	0.06

```
R> #
R> m1ATT<-m1
R> sem1ATT<-sem1
R> tm1ATT<-tm1
```

4. Estimation via pooled regression adjustment

This procedure imposes the ex-ante restriction that the effect of covariates on outcome is identical for the treated and untreated. The treatment effect can then be estimated in straightforward fashion by including a "treatment dummy" in the regression of outcome on covariates, using the *entire sample.*

We follow **?** and use heteroskedasticiy-robust standard errors for this step.

```
R> X<-cbind(rep(1,n),train,age,educ,black,hisp,married,re74,re75)
R> bols<-solve((t(X)) %*% X) %*% (t(X) %*% y)
R> e<-as.vector(y-X%*%bols)
R> S<-diag(e^2)
R> Vb<-solve((t(X))%*%X) %*% t(X) %*% S %*% X %*% solve((t(X))%*%X)
R> se=sqrt(diag(Vb))
R> tval=bols/se
R> #
R> m2<-as.vector(bols[2])
R> sem2<-as.vector(se[2])
R> tm2<-as.vector(tval[2])
R> #
R> m2ATT<-m2
R> sem2ATT<-sem2
R> tm2ATT<-tm2
```

5. Estimation via regression adjustment using seperate equations

This approach allows for separate regression coefficients for the treated and untreated.

```
R> X<-cbind(rep(1,n),age,educ,black,hisp,married,re74,re75)
R> #eliminate train as a explanatory variable
```

```
R> y1<-as.matrix(y[1:n1])
R> y0<-as.matrix(y[(n1+1):n])
R> X1<-X[1:n1,]
R> X0<-X[(n1+1):n,]
R> #
R> b1<-solve((t(X1)) %*% X1) %*% (t(X1) %*% y1)
R> b0<-solve((t(X0)) %*% X0) %*% (t(X0) %*% y0)
R> #
R> #ATE
R> y1p<-X%*%b1 #create predictions for treated outcome for ALL observations
R> y0p<-X%*%b0 #create predictions for UNtreated outcome for ALL observations
R> m3<-mean(y1p-y0p)
R> #ATT
R> m3ATT<-mean(y1p[1:n1]-y0p[1:n1])
R> #
R> #
R> #run bootstrap to get s.e.'s (see Wooldridge, p. 918)
R> ################################
R> R<-1000 #number of bootstrap replications
R> out<-rep(0,R) #will collect ATE result for each replication
R> outATT<-rep(0,R) #will collect ATT result for each replication
R> com1<-cbind(y1,X1) #glue y1 and X1 together
R> com0<-cbind(y0,X0)
R> for (i in 1:R) {
   int1<-com1[sample(nrow(com1),n1,replace=TRUE), ]
   #sample n1 id's with replacement (this allows for multiple entries)
   y1r<-int1[,1]
   X1r<-int1[,2:dim(com1)[2]]
   b1<-solve((t(X1r)) %*% X1r) %*% (t(X1r) %*% y1r)
   #
   int0<-com0[sample(nrow(com0),n0,replace=TRUE), ]
   #sample n1 id's with replacement (this allows for multiple entries)
   y0r<-int0[,1]
   X0r<-int0[,2:dim(com0)[2]]
   b0<-solve((t(X0r)) %*% X0r) %*% (t(X0r) %*% y0r)
   #
   Xr<-rbind(X1r,X0r)
   y1rp<-Xr%*%b1
   y0rp<-Xr%*%b0
   #ATE
   out[i]<-mean(y1rp-y0rp)
   #ATT
   outATT[i]<-mean(y1rp[1:n1]-y0rp[1:n1])
 }
R> sem3<-sd(out)
R> tm3<-m3/sem3
R> #
```

```
R> sem3ATT<-sd(outATT)
R> tm3ATT<-m3ATT/sem3ATT
```

Table 2. Combined estimation results for ATE

estimator	estimate	s.e.	t-value
difference in means	1.794	0.671	2.674
pooled regression	1.683	0.651	2.583
separate regressions	1.633	0.655	2.493

Table 3. Combined estimation results for ATT

estimator	estimate	s.e.	t-value
difference in means	1.794	0.671	2.674
pooled regression	1.683	0.651	2.583
separate regressions	1.774	0.669	2.650

```
R> proc.time()-tic
   user  system elapsed
   1.58    0.25    1.83
```

References

Imbens, I. and Rubin, D. (forthcoming). *Causal Inference in Statistics and the Social Sciences*, Cambridge University Press.

Lalonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data, *American Economic Review* **76**: 604–620.

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*, MIT Press.