SCRIPT MOD4S1A: OV BIAS IN OLS REGRESSION

INSTRUCTOR: KLAUS MOELTNER

OLS on full model

We will use the hedonic property value data from mod3s2c with the dependent variable in log-form. Let's assume this full specification is indeed the "true" model.

```
R> load("c:/Klaus/AAEC5126/R/data/hedonics.rda")
R> attach(data)
R> y<-log(price)
R> n<-nrow(data)
R> X<-cbind(rep(1,n),lnacres,lnsqft,age,gradeab,pkadeq,
 vacant,empden,popden,metro,distair,disthaz)
R> k<-ncol(X)
R> #
R> #OLS:
R> bols<-solve((t(X)) %*% X) %*% (t(X) %*% y)# compute OLS estimator
R> e<-y-X%*%bols # Get residuals.
R> SSR<-(t(e)%*%e)#sum of squared residuals - should be minimized
R> s2<-(t(e)%*%e)/(n-k) #get the regression error (estimated variance of "eps").
R> Vb<-s2[1,1]*solve((t(X))%*%X) # get the estimated
R> #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> bfull<-bols #capture for later
R> kfull<-k
R> #
R> ttols<-data.frame(col1=c("constant","lnacres","lnsqft","age","gradeab",
 "pkadeq","vacant","empden","popden","metro","distair","disthaz"),
                col2=bols,
                col3=se,
                col4=tval)
R> colnames(ttols)<-c("variable","estimate","s.e.","t")


R> ttolsx<- xtable(ttols,caption="OLS output (full model)")
R> digits(ttolsx)<-3
R> print(ttolsx,include.rownames=FALSE,
 latex.environment="center", caption.placement="top",table.placement="H")
```

Table 1. OLS output (full model)

variable	estimate	s.e.	t
constant	9.905	0.332	29.878
lnacres	0.372	0.075	4.934
lnsqft	0.595	0.077	7.719
age	0.002	0.003	0.763
gradeab	0.716	0.238	3.012
pkadeq	0.025	0.132	0.190
vacant	-0.004	0.005	-0.799
empden	0.015	0.004	4.112
popden	-0.003	0.012	-0.220
metro	0.488	0.115	4.231
distair	0.108	0.018	5.951
disthaz	0.033	0.013	2.550

The estimated variance for the regression error is 0.743.

OLS on flawed model

Let's omit `distair`, a variable we know has a significant nonzero effect on `log(price)`.

```
R> load("c:/Klaus/AAEC5126/R/data/hedonics.rda")
R> attach(data)
R> y<-log(price)
R> n<-nrow(data)
R> X<-cbind(rep(1,n),lnacres,lnsqft,age,gradeab,pkadeq,
 vacant,empden,popden,metro,disthaz)
R> k<-ncol(X)
R> #
R> #OLS:
R> bols<-solve((t(X)) %*% X) %*% (t(X) %*% y)# compute OLS estimator
R> e<-y-X%*%bols # Get residuals.
R> SSR<-(t(e)%*%e)#sum of squared residuals - should be minimized
R> s2<-(t(e)%*%e)/(n-k) #get the regression error (estimated variance of "eps").
R> Vb<-s2[1,1]*solve((t(X))%*%X) # get the estimated
R> #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> bflaw<-bols #capture for later
R> #
R> #
R> ttols<-data.frame(col1=c("constant","lnacres","lnsqft","age","gradeab",
 "pkadeq","vacant","empden","popden","metro","disthaz"),
                col2=bols,
                col3=se,
                col4=tval)
R> colnames(ttols)<-c("variable","estimate","s.e.","t")
```

```
R> ttolsx<- xtable(ttols,caption="OLS output (flawed model)")
R> digits(ttolsx)<-3
R> print(ttolsx,include.rownames=FALSE,
 latex.environment="center", caption.placement="top",table.placement="H")
```

Table 2. OLS output (flawed model)

variable	estimate	s.e.	t
constant	10.654	0.320	33.269
lnacres	0.445	0.078	5.728
lnsqft	0.595	0.080	7.400
age	0.004	0.003	1.319
gradeab	0.837	0.247	3.386
pkadeq	-0.045	0.138	-0.326
vacant	0.006	0.005	1.143
empden	0.013	0.004	3.284
popden	0.039	0.010	3.936
metro	0.474	0.120	3.934
disthaz	0.020	0.013	1.524

The estimated variance for the regression error is 0.81.

Examine OV Regression Matrix

Clearly, the flawed model produces somewhat different results. We are mainly concerned about three effects:

(1) Changes in significance
(2) If significance is unchanged, changes in sign and / or magnitude

With respect to the first concern, we note that the coefficient for `popden` is now significant at the 1% level, with a positive effect on log(price), and the coefficient for `disthaz` has dimisihed in magnitude and lost significance.

With respect to the second concern, we observe that the coefficients for `lnacres` and `gradeab` have increased substantially, and the coefficients for `metro` and `empden` have decreased slightly.

As we know from the lecture, the formula for the OV bias is $\mathbf{P}_{1.2} * \boldsymbol{\beta}_2$, where $\mathbf{P}_{1.2}$ is the coefficient matrix of a regression of the included variables on the omitted variable(s), and $\boldsymbol{\beta}_2$ is the coefficient vector that captures the true effect of the omitted variable(s) on the dependent variable in the original regression.

Let's first compute the OV regression matrix. This will give us an idea of the magnitude and direction of bias for each of the included coefficients. From the full model we know that $\boldsymbol{\beta}_2 = \beta_{distair} = 0.108$. Thus, we know that the sign of a given element of $\mathbf{P}_{1.2}$ will indicate the direction of the bias for the corresponding included variable. The magnitude of the bias will be 0.108 times the magnitude of the $\mathbf{P}_{1.2}$ element.

```
R> X1<-cbind(rep(1,n),lnacres,lnsqft,age,gradeab,pkadeq,
 vacant,empden,popden,metro,disthaz)
R> X2<-distair
R> P12<- solve((t(X1)) %*% X1) %*% (t(X1) %*% X2)
R> #
R> tt<-data.frame(col1=c("constant","lnacres","lnsqft","age","gradeab",
 "pkadeq","vacant","empden","popden","metro","disthaz"),
                col2=P12)
R> colnames(tt)<-c("variable", "P12 coeff.")
```

```
R> ttx<- xtable(tt,caption="P12 output")
R> digits(ttx)<-3
R> print(ttx,include.rownames=FALSE,
 latex.environment="center", caption.placement="top",table.placement="H")
```

Table 3. P12 output

variable	P12 coeff.
constant	6.913
lnacres	0.673
lnsqft	0.004
age	0.016
gradeab	1.118
pkadeq	-0.647
vacant	0.091
empden	-0.025
popden	0.386
metro	-0.134
disthaz	-0.118

These results are perfectly consistent with our observed changes in sign and magnitude of the coefficients for the included variables. For example, the link between the omitted variable and `lnsqft` is virtually zero. Not surprisingly, we don't see any changes for that variable in the flawed model. In contrast, the $\mathbf{P}_{1.2}$ coefficient for `gradeab` is comparatively large and positive. This explains the measurable increase in the coefficient for `gradeab` in the flawed model compared to the full model.

Intuitively, this would indicate that properties with a larger distance from the airport also tend to have higher quality grades. One could try to confirm this by re-examinig the raw data. It's these kinds of "linkages" that drive omitted variable problems.

Also note that `distair` has a strong negative link with `pkadeq`, implying that parking becomes increasingly less adequate as one moves away from the airport. However, we're not ovely concerned about this bias since `pkadeq` is not a significant regressor in the full model.

For a full comparison, let's compute the exact bias for each regressor in the flawed model, and then compare all results in a single table:

```
R> tt<-data.frame(col1=c("constant","lnacres","lnsqft","age","gradeab",
 "pkadeq","vacant","empden","popden","metro","disthaz"),
                col2=c(bfull[1:(kfull-2)],bfull[kfull]),
                col3=bflaw,
                col4=P12,
                col5=bfull[kfull-1]*P12)
R> colnames(tt)<-c("variable","correct","flawed","P12","bias")
```

```
R> ttx<- xtable(tt,caption="Comparison of Results")
R> digits(ttx)<-3
R> print(ttx,include.rownames=FALSE,
 latex.environment="center", caption.placement="top",table.placement="H")
```

Table 4. Comparison of Results

variable	correct	flawed	P12	bias
constant	9.905	10.654	6.913	0.748
lnacres	0.372	0.445	0.673	0.073
lnsqft	0.595	0.595	0.004	0.000
age	0.002	0.004	0.016	0.002
gradeab	0.716	0.837	1.118	0.121
pkadeq	0.025	-0.045	-0.647	-0.070
vacant	-0.004	0.006	0.091	0.010
empden	0.015	0.013	-0.025	-0.003
popden	-0.003	0.039	0.386	0.042
metro	0.488	0.474	-0.134	-0.014
disthaz	0.033	0.020	-0.118	-0.013

```
R> proc.time()-tic
   user  system elapsed
   0.21    0.05    0.30
```