SCRIPT MOD3S2C: MODEL FIT AND PREDICTION

INSTRUCTOR: KLAUS MOELTNER

Load and describe data

This example is based on data from Ihlanfeldt and Taylor (2004), "Externality Effects of Small-Scale Hazardous Waste Sites: Evidence from Urban Commercial Property Markets", *Journal of Environmental Economics and Management*, vol. 47, no.1, pp. 117-39.

In this study, the authors examine the effect on commercial property values in the Atlanta, Georgia, area of nearby hazardous waste sites. We'll use a subset of the data that focuses on 395 apartment / condominium structures that were sold and bought by real estate companies between 1982 and 1998. All sales occurred after the official listing of a local hazardous waste site.

```
R> data<- read.table('c:/Klaus/AAEC5126/R/data/hedonics.txt', sep="\t", header=FALSE)
R> #
R> #assign variable names
R> names(data)[1]<-"price"
R> names(data)[2]<-"lnacres"
R> names(data)[3]<-"lnsqft"
R> names(data)[4]<-"age"
R> names(data)[5]<-"gradeab"
R> names(data)[6]<-"pkadeq"
R> names(data)[7]<-"vacant"
R> names(data)[8]<-"empden"
R> names(data)[9]<-"popden"
R> names(data)[10]<-"metro"
R> names(data)[11]<-"distair"
R> names(data)[12]<-"disthaz"
R> #
R> save(data, file = "c:/Klaus/AAEC5126/R/data/hedonics.rda")
R> attach(data)
```

Variable definitions:

TABLE 1. Variable description for property value data

pos.	variable	description
1	price	sales price in 2007 dollars
2	lnacres	log of (acreage of property)
3	lnsqft	log of (square footage, in 1000 feet)
4	age	age of property, in years
5	gradeab	1=propety received highest score from tax assessor
6	pkadeq	1=propert has adequate parking
7	vacant	percentage of vacant land in census tract
8	empden	employment density (workers /acre in census tract)
9	popden	population density in census tract (person / acre)
10	metro	1=within 1 mile of METRO station at time of sale
11	distair	distance to airport (miles)
12	disthaz	distance to hazardous waste site (miles)

MODEL 1

In this model, the dependent variable is scaled in units of 100,000.

```
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> y<-price/100000
R> #
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 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("constant","lnacres","lnsqft","age","gradeab","pkadeq","vacant",
 "empden","popden","metro","distair","disthaz"),
                col2=bols,
                col3=se,
                col4=tval)
R> colnames(tt)<-c("variable","estimate","s.e.","t")
```

TABLE 2. OLS output for Model 1

variable	estimate	s.e.	t
constant	-30.599	11.319	-2.703
lnacres	15.753	2.573	6.122
lnsqft	2.634	2.630	1.001
age	-0.099	0.097	-1.015
gradeab	11.751	8.113	1.448
pkadeq	3.522	4.519	0.779
vacant	-0.051	0.174	-0.295
empden	0.137	0.128	1.070
popden	-2.266	0.405	-5.600
metro	4.598	3.941	1.167
distair	8.320	0.621	13.400
disthaz	-0.220	0.445	-0.493

```
R> #Compute RMSE, MAE, THEIL-U
R> yhat<-X%*%bols
R> rmse<-sqrt(mean((y-yhat)^2))
R> mae<-mean(abs(y-yhat))
R> U<-rmse/sqrt(mean(y^2))
```

The RMSE for this model is 28.985.
The MAE for this model is 16.37.
The Theil-U statistic for this model is 0.607.

MODEL 2

In this model, the dependent variable is scaled in units of 1,000,000.

```
R> n<-nrow(data)
R> X<-cbind(rep(1,n),lnacres,lnsqft,age,gradeab,pkadeq,vacant,empden,popden,metro,distair,distl
R> k<-ncol(X)
R> y<-price/1000000
R> #
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 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("constant","lnacres","lnsqft","age","gradeab","pkadeq","vacant","empde
                  col2=bols,
                  col3=se,
                  col4=tval)
R> colnames(tt)<-c("variable","estimate","s.e.","t")
```

TABLE 3. OLS output for Model 2

variable	estimate	s.e.	t
constant	-3.060	1.132	-2.703
lnacres	1.575	0.257	6.122
lnsqft	0.263	0.263	1.001
age	-0.010	0.010	-1.015
gradeab	1.175	0.811	1.448
pkadeq	0.352	0.452	0.779
vacant	-0.005	0.017	-0.295
empden	0.014	0.013	1.070
popden	-0.227	0.040	-5.600
metro	0.460	0.394	1.167
distair	0.832	0.062	13.400
disthaz	-0.022	0.045	-0.493

```
R> #Compute RMSE, MAE, THEIL-U
R> yhat<-X%*%bols
R> rmse<-sqrt(mean((y-yhat)^2))
R> mae<-mean(abs(y-yhat))
R> U<-rmse/sqrt(mean(y^2))
```

The RMSE for this model is 2.899.
The MAE for this model is 1.637.
The Theil-U statistic for this model is 0.607.

PREDICTIONS FROM A LINEAR MODEL

Example 1. Hold all variables at their sample mean and set `hazdist` to 1.5 miles (for comparison, the sample mean of `hazdist` is 7.8).

```
R> xp<-matrix(c(colMeans(X[,1:(ncol(X)-1)]),1.5))
R> yp<-t(xp)%*%bols
R> Vyp<-t(xp) %*% Vb %*% xp
R> seyp<-sqrt(Vyp)
R> lo<-yp-1.96*seyp
R> hi<-yp+1.96*seyp
```

The predicted value of a property that is 1.5 miles from the hazardous waste site is (in million dollars) 1.401.
The lower bound of a 95% confidence interval for this estimate is 0.777.
The upper bound is 2.026

Example 2. Hold all variables at their sample mean and set `hazdist` to 10 miles (for comparison, the sample mean of `hazdist` is 7.8).

```
R> xp<-matrix(c(colMeans(X[,1:(ncol(X)-1)]),10))
R> yp<-t(xp)%*%bols
```

```
R> Vyp<-t(xp) %*% Vb %*% xp
R> seyp<-sqrt(Vyp)
R> lo<-yp-1.96*seyp
R> hi<-yp+1.96*seyp
```

The predicted value of a property that is 10 miles from the hazardous waste site is (in million dollars) 1.215.
The lower bound of a 95% confidence interval for this estimate is 0.868.
The upper bound is 1.561.

Nonlinear Predictions

Let's re-run the property value model with the dependent variable in log form (call it Model 3). While this is still a "linear regression model", in the sense that all parameters enter linearly, the predictions in *un-logged* form flowing from this model are inherently non-linear.

```
R> y<-log(price)
R> #
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 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("constant","lnacres","lnsqft","age","gradeab","pkadeq","vacant","empde
                  col2=bols,
                  col3=se,
                  col4=tval)
R> colnames(tt)<-c("variable","estimate","s.e.","t")
```

Table 4. OLS output for Model 3

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

Example 1 via Delta Method. Hold all variables at their sample mean and set `hazdist` to 1.5 miles (for comparison, the sample mean of `hazdist` is 7.8).

```
R> xp<-matrix(c(colMeans(X[,1:(ncol(X)-1)]),1.5))
R> yp<-exp(t(xp)%*%bols  + 0.5*s2)/1000000
R> # The exp(.) term is the conversion formula
R> # to go from log(price in dollars) back to "actual price in dollars"
R> # In statistical terms, we're switching from the normal to the log-normal
R> # distribution;
R> # The division by 1000000 returns prices in units of $1000000.
R> # This is optional, of course.
R> #
R> # Use Delta method to obtain the prediction variance
R> C<-t(yp[1,1]*xp)
R> Vyp<-C %*% Vb %*% t(C)
R> seyp<-sqrt(Vyp)
R> lo<-yp-1.96*seyp
R> hi<-yp+1.96*seyp
```

The predicted value of a property that is 1.5 miles from the hazardous waste site is (in million dollars) 0.293.
The lower bound of a 95% confidence interval for this estimate is 0.239.
The upper bound is 0.346

Example 2 via Delta Method. Hold all variables at their sample mean and set `hazdist` to 10 miles (for comparison, the sample mean of `hazdist` is 7.8).

```
R> xp<-matrix(c(colMeans(X[,1:(ncol(X)-1)]),10))
R> yp<-exp(t(xp)%*%bols  + 0.5*s2)/1000000
R> # The exp(.) term is the conversion formula
R> # to go from log(price in dollars) back to "actual price in dollars"
R> # In statistical terms, we're switching from the normal to the log-normal
R> # distribution;
R> # The division by 1000000 returns prices in units of $1000000.
R> # This is optional, of course.
R> #
R> # Use Delta method to obtain the prediction variance
R> C<-t(yp[1,1]*xp)
R> Vyp<-C %*% Vb %*% t(C)
R> seyp<-sqrt(Vyp)
R> lo<-yp-1.96*seyp
R> hi<-yp+1.96*seyp
```

The predicted value of a property that is 10 miles from the hazardous waste site is (in million dollars) 0.388.
The lower bound of a 95% confidence interval for this estimate is 0.349.
The upper bound is 0.428

Example 2 via Simulation. We will draw $R = 10000$ vectors of the OLS estimator from its empirical (or "sampling") distribution. For each draw, we compute the predicted value yp. We then examine the resulting simulated distribution of yp, for instance by extracting its mean and the 2.5^{th} and 97.5^{th} percentile. We use the latter two as empirical bounds for a 95% confidence Interval.

This procedure is similar in spirit to the Krinsky-Robb method illustrated in script `mod3_1d`. As before we'll try to do this without a loop, by taking all draws at the same time, and computing the function of interest simultaneously for all draws.

```
R> R<-10000; #number of repetitions
R> #Step 1: Draw R b's from empirical density
R> mubols<-matrix(bols)#turns bols into a 1xk vector
R> Varbols<-matrix(Vb,nrow=k)#turns Vb into a kxk matrix
R> bmat<-mvrnorm(n=R,mubols,Varbols)
R> bmat<-t(bmat)#change into kxR
R> #
R> #Step2: For each draw, compute function of interest
R> i<-matrix(1,1,R)  #needed for conformability
R> ypvec<-exp(t(xp)%*%bmat  + 0.5*s2[1,1]*i)/1000000
R> #
R> yp<-mean(ypvec)
R> lo<-quantile(ypvec,0.025)
R> hi<-quantile(ypvec,0.975)
```

The predicted value of a property that is 10 miles from the hazardous waste site is (in million dollars) 0.389.
The lower bound of a 95% confidence interval for this estimate is 0.351.
The upper bound is 0.43

```
R> proc.time()-tic
   user  system elapsed
   0.09    0.14    0.24
```