Economics 468

October 26, 2015 **R. Davidson**

Midterm Examination

Do not be upset if this exam seems too long for you! Do not waste time on questions for which you do not see how to obtain the answer. Rather, answer as much as you can. Everything you do will be taken into account.

1. If two random variables X_1 and X_2 are statistically independent, show that $\mathrm{E}(X_1 \mid X_2) = \mathrm{E}(X_1)$.

The **covariance** of two random variables X_1 and X_2, which is often written as $\mathrm{cov}(X_1, X_2)$, is defined as the expectation of the product of $X_1 - \mathrm{E}(X_1)$ and $X_2 - \mathrm{E}(X_2)$. Consider a random variable X_1 with mean zero. Show that the covariance of X_1 and any other random variable X_2, whether it has mean zero or not, is just the expectation of the product of X_1 and X_2.

Show that the variance of the random variable $X_1 - \mathrm{E}(X_1 \mid X_2)$ cannot be greater than the variance of X_1, and that the two variances are equal if X_1 and X_2 are independent.

Let a random variable X_1 be distributed as $\mathrm{N}(0, 1)$. Now suppose that a second random variable, X_2, is constructed as the product of X_1 and an independent random variable Z, which equals 1 with probability $1/2$ and -1 with probability $1/2$. What is the (marginal) distribution of X_2? What is the covariance between X_1 and X_2? What is the distribution of X_1 conditional on X_2?

2. Show that the t^{th} residual from running the regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \alpha \boldsymbol{e}_t + \boldsymbol{u},$$

is zero. Here $\boldsymbol{e}_t$ is the unit basis vector, all of whose elements are zero, except for element t, which is equal to one. Use this fact to demonstrate that, as a result of omitting observation t, the t^{th} residual from the regression $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ changes by an amount

$$\hat{u}_t \frac{h_t}{1 - h_t},$$

where h_t is the (t, t)-diagonal element of the orthogonal projection matrix $\boldsymbol{P}_{\boldsymbol{X}}$.

Show that the leverage measure h_t is the square of the cosine of the angle between the unit basis vector $\boldsymbol{e}_t$ and its projection on to the span $\mathcal{S}(\boldsymbol{X})$ of the regressors.

3. Consider the following linear regression:

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u},$$

where $\boldsymbol{y}$ is $n \times 1$, $\boldsymbol{X}_1$ is $n \times k_1$, and $\boldsymbol{X}_2$ is $n \times k_2$. Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ be the OLS parameter estimates from running this regression.

Now consider the following regressions, all to be estimated by OLS:

(a) $\boldsymbol{y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$;
(b) $\boldsymbol{P}_1\boldsymbol{y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$;
(c) $\boldsymbol{P}_1\boldsymbol{y} = \boldsymbol{P}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$;
(d) $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$;
(e) $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$;
(f) $\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$;
(g) $\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$;
(h) $\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$;
(i) $\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$;
(j) $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$.

Here $\boldsymbol{P}_1$ projects orthogonally on to the span of $\boldsymbol{X}_1$, and $\boldsymbol{M}_1 = \mathbf{I} - \boldsymbol{P}_1$. For which of the above regressions are the estimates of $\boldsymbol{\beta}_2$ the same as for the original regression? Why? For which are the residuals the same? Why?

4. Show that the difference between the unrestricted estimator $\tilde{\boldsymbol{\beta}}$ of the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\mathbf{0}, \sigma^2\mathbf{I}),$$

and the restricted estimator $\hat{\boldsymbol{\beta}}$ of the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\mathbf{0}, \sigma^2\mathbf{I}),$$

is given by

$$\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}.$$

Hint: In order to prove this result, it is easiest to premultiply the difference by $\boldsymbol{X}^\top\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X}$.

Suppose that the random variable z follows the $\text{N}(0,1)$ density. If z is a test statistic used in a two-tailed test, the corresponding P value is shown to be $p(z) \equiv 2\big(1 - \Phi(|z|)\big)$. Show that $F_p(\cdot)$, the CDF of $p(z)$, is the CDF of the uniform distribution on $[0,1]$. In other words, show that

$$F_p(x) = x \quad \text{for all } x \in [0,1].$$

5. Two econometricians obtained data on the sale prices and various characteristics of 546 houses sold in Windsor, Ontario in 1987. The characteristics are:

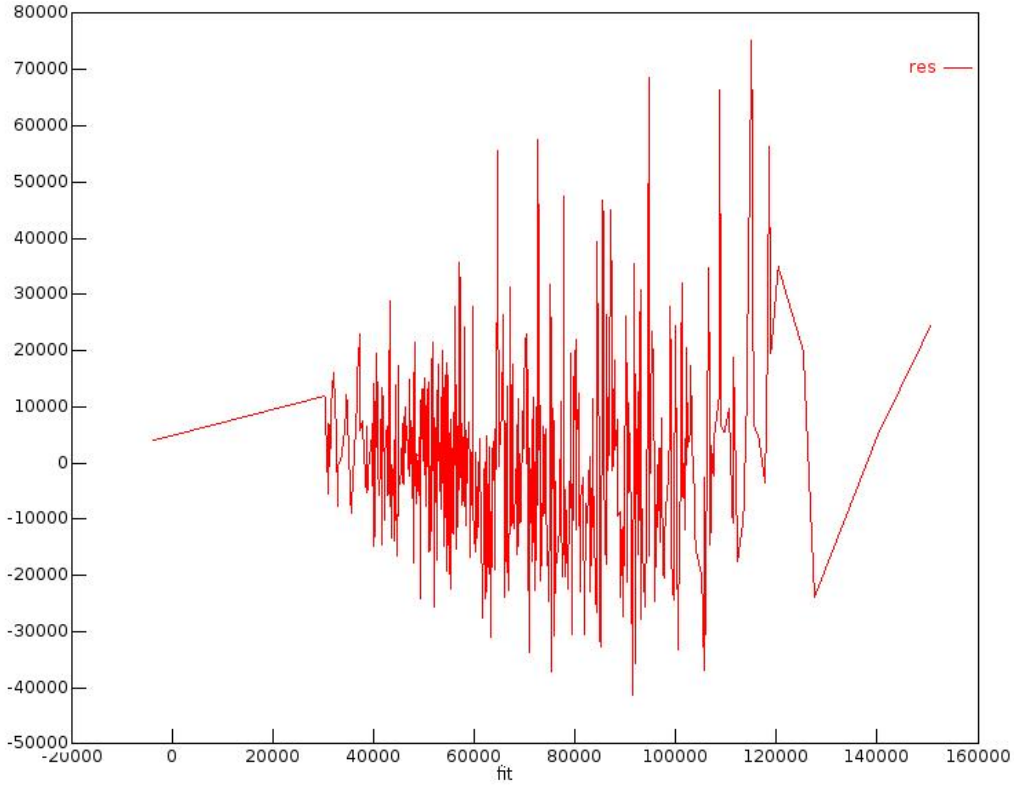
`ls`	= the lot size of a property in square feet
`bdms`	= the number of bedrooms
`baths`	= the number of full bathrooms
`sty`	= the number of stories excluding basement
`drv`	= 1 if the house has a driveway
`rec`	= 1 if the house has a recreational room
`ffin`	= 1 if the house has a full finished basement
`ghw`	= 1 if the house uses gas for hot water heating
`ca`	= 1 if there is central air conditioning
`gar`	= the number of garage places
`reg`	= 1 if the house is located in the preferred neighbourhood of the city

Note that all of the characteristics except the first three are indicator, or dummy, variables, that can take on only the values of zero or one. The sale prices (the variable `value`) were regressed on a constant and all of the characteristics. The results of the regression are as follows:

```
Ordinary Least Squares:

Variable    Parameter estimate  Standard error  T statistic

constant    -3890.912944        3316.422971      -1.173226
ls              3.535404           0.348470      10.145510
bdms         1775.929742        1034.454519       1.716779
baths       14418.793223        1483.883283       9.716932
sty          6571.231722         921.765886       7.128960
drv          6575.674659        2029.200053       3.240526
rec          4573.381783        1892.691281       2.416338
ffin         5510.572018        1582.046133       3.483193
ghw         12831.052275        3205.286305       4.003091
ca          12431.136062        1552.170690       8.008872
gar          4240.271919         836.987405       5.066112
reg          9469.694699        1662.941237       5.694546

Number of observations = 546   Number of estimated parameters = 12
Mean of dependent variable = 67929.289377
Sum of squared residuals = 1.260551e+11
Explained sum of squares = 2.785264e+12
Estimate of residual variance
  (with d.f. correction) = 2.360582e+08
Mean of squared residuals = 2.308701e+08
Standard error of regression = 15364.186869
R squared (uncentred) = 0.956702  (centred) = 0.678319
```

This type of regression is called a **hedonic regression**. It attempts to estimate the values that consumers place on various characteristics of a good, in this case a house. By how much, on average, does the price of a house increase when the number of full bathrooms increases by one?

After running the regression, the econometricians plotted the residuals against the fitted values, and obtained the plot shown on the next page. Why does this plot suggests that the regression is not well specified? Can you make suggestions as to how a better fitting regression might be obtained?


res
80000
70000
60000
50000
40000
30000
20000
10000
0
-10000
-20000
-30000
-40000
-50000
-20000
0
20000
40000
60000
80000
100000
120000
140000
160000
fit