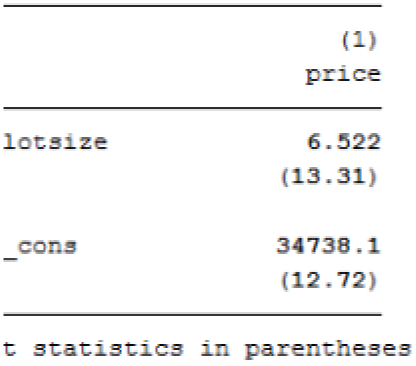
C528 Data Predictive Analytics and Business Strategy
Summer 2018
Homework 2, Team Purple

Benjamin Frazier
Sathya Thiruvengadam
Bassi Sidibe
Zach Toelle

Problem 1

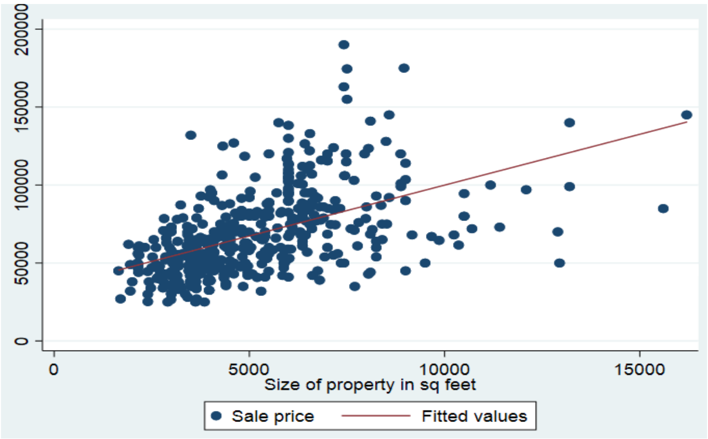
1.

Using the *HousePrice.dta* dataset, a scatter plot of the following basic model is displayed below with a regression line overlaid regressing *lotsize* against *price*.



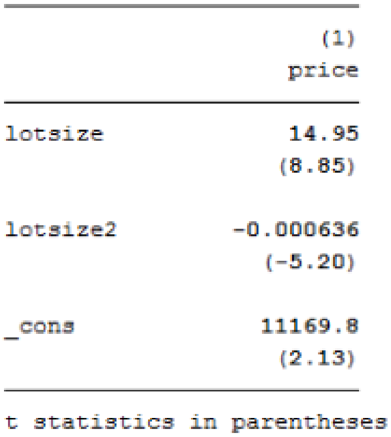
```
----------------------------
                      (1)
                    price
----------------------------
lotsize             6.522
                  (13.31)

_cons             34738.1
                  (12.72)
----------------------------
t statistics in parentheses
```



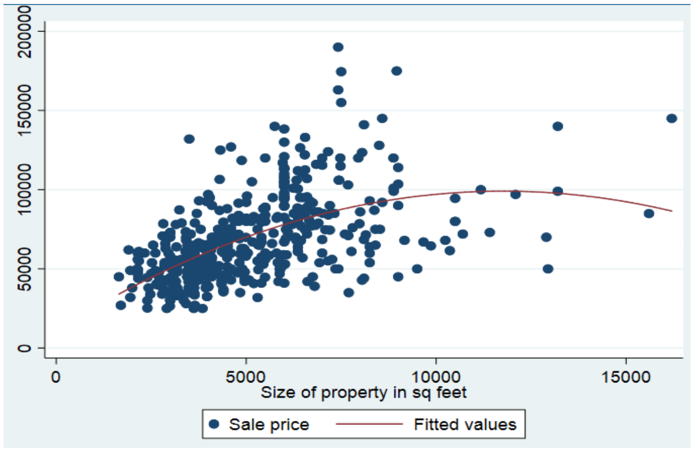
2.

To adjust this model we can use a quadratic functional form. The regression line on the scatter plot below reflects this change. The quadratic form may do a better predicting house prices as we see properties that begin to decrease in price as the square footage increases to the largest amounts. This may be due to the idea of there being too much maintenance involved in these large properties. When buying property an acreage is typically seen as a large amount of space for any family, and anything beyond that would be costly or time consuming to upkeep. Our model predicts that prices begin to peak and then decline after two acres, which supports this theory.



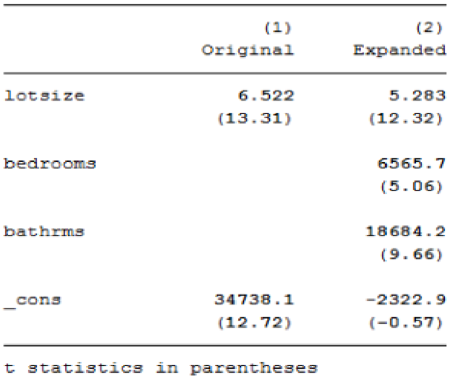
	(1) price
lotsize	14.95
	(8.85)
lotsize2	-0.000636
	(-5.20)
_cons	11169.8
	(2.13)

t statistics in parentheses



3.

We can see some improvements to the *lotsize* coefficient by including number of bedrooms and bathrooms as variables, both of which are statistically significant. This is because by introducing these variables we are reducing omitted-variable bias. When bedrooms and bathrooms are not included in the model, their effects on price are attributed to the included variable (*lotsize).*



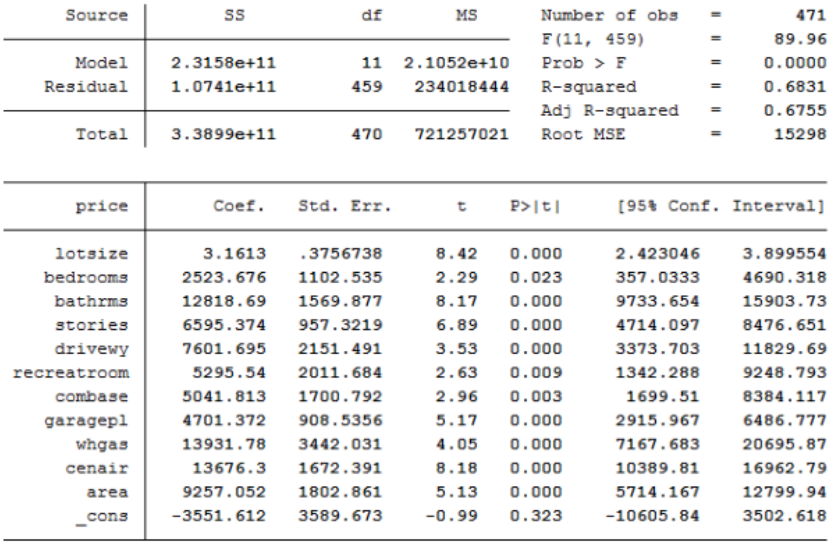
	(1) Original	(2) Expanded
lotsize	6.522	5.283
	(13.31)	(12.32)
bedrooms		6565.7
		(5.06)
bathrms		18684.2
		(9.66)
_cons	34738.1	-2322.9
	(12.72)	(-0.57)

t statistics in parentheses

The original model estimate gives 6.52233 as the coefficient for *lotsize* prior to including the two new variables. The expanded model gives a *lotsize* coefficient of 5.28284. Since the number of bedrooms and bathrooms are positively associated with *price*, that attribute is removed from *lotsize*, giving it a smaller coefficient.

4.

The model with the rest of the variables in the dataset included in the model is shown below. All the variables have a positive coefficient, meaning that including or increasing these factors in a property increases the price. Looking at B_1, the coefficient for *lotsize*, we see that it has a smaller coefficient here (3.1613) than in the original model (6.52233) and the model expanded with bedrooms and bathrooms (5.28284). This is due to the previously omitted variables no longer attributing their effect on *price* to *lotsize*.



```
      Source |       SS           df       MS      Number of obs   =       471
-------------+----------------------------------   F(11, 459)      =     89.96
       Model |  2.3158e+11        11  2.1052e+10   Prob > F        =    0.0000
    Residual |  1.0741e+11       459   234018444   R-squared       =    0.6831
-------------+----------------------------------   Adj R-squared   =    0.6755
       Total |  3.3899e+11       470   721257021   Root MSE        =     15298

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lotsize |     3.1613   .3756738     8.42   0.000     2.423046    3.899554
    bedrooms |   2523.676   1102.535     2.29   0.023     357.0333    4690.318
     bathrms |   12818.69   1569.877     8.17   0.000     9733.654    15903.73
     stories |   6595.374   957.3219     6.89   0.000     4714.097    8476.651
     drivewy |   7601.695   2151.491     3.53   0.000     3373.703    11829.69
 recreatroom |    5295.54   2011.684     2.63   0.009     1342.288    9248.793
     combase |   5041.813   1700.792     2.96   0.003      1699.51    8384.117
    garagepl |   4701.372   908.5356     5.17   0.000     2915.967    6486.777
       whgas |   13931.78   3442.031     4.05   0.000     7167.683    20695.87
      cenair |    13676.3   1672.391     8.18   0.000     10389.81    16962.79
        area |   9257.052   1802.861     5.13   0.000     5714.167    12799.94
       _cons |  -3551.612   3589.673    -0.99   0.323    -10605.84    3502.618
------------------------------------------------------------------------------
```

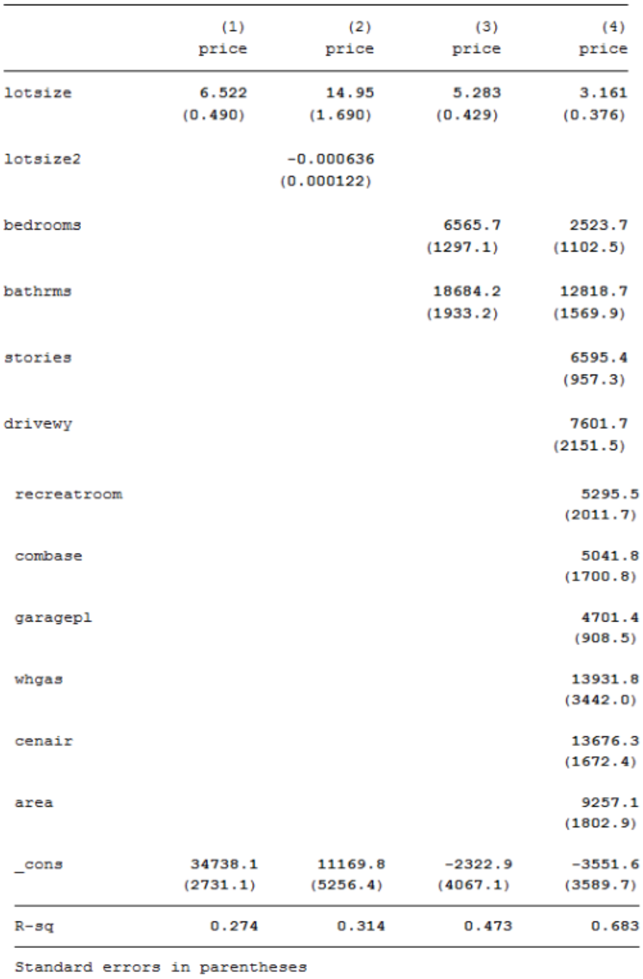
All Variable Model

5.

In Question 4, omitted variables may still be causing endogeneity in *lotsize*. One possible variable is relative tax rate. This variable causes endogeneity because higher relative tax rates tend to decrease price, and lotsize is a determining factor in your taxes. A second potential omitted variable is whether there is a pool on the lot in addition to a home. A large lotsize may have a higher possibility of having a pool included, which will affect the overall price.

6.

A table comparing the coefficient estimates, standard errors, and R^2 of the discussed models is shown below. Model 1 includes only *lotsize*, Model 2 includes only *lotsize* in a quadratic form, Model 3 includes *bedrooms* and *bathrms*, and Model 4 includes all variables in the dataset.



	(1) price	(2) price	(3) price	(4) price
lotsize	6.522 (0.490)	14.95 (1.690)	5.283 (0.429)	3.161 (0.376)
lotsize2		-0.000636 (0.000122)		
bedrooms			6565.7 (1297.1)	2523.7 (1102.5)
bathrms			18684.2 (1933.2)	12818.7 (1569.9)
stories				6595.4 (957.3)
drivewy				7601.7 (2151.5)
recreatroom				5295.5 (2011.7)
combase				5041.8 (1700.8)
garagepl				4701.4 (908.5)
whgas				13931.8 (3442.0)
cenair				13676.3 (1672.4)
area				9257.1 (1802.9)
_cons	34738.1 (2731.1)	11169.8 (5256.4)	-2322.9 (4067.1)	-3551.6 (3589.7)
R-sq	0.274	0.314	0.473	0.683

Standard errors in parentheses

Problem 2

1.

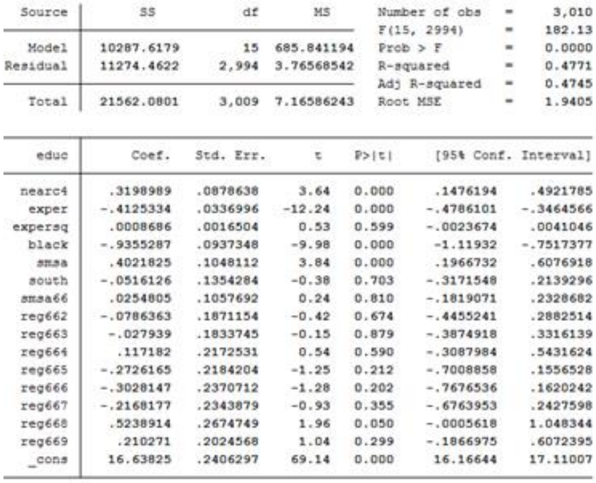
In this model (OLS column in the table below) the education coefficient is .0746933, and the standard error is .0034983. This implies that an additional year of schooling increases wages by 7.4%. However, this model does not give a causal effect of education on wages. This is because there have been many studies that support the hypothesis that both your income and education are both primarily driven by your parents income. This would mean that educ is endogenous.

2.

If we wish to use *nearc4* as an instrumental variable, it must be both relevant and exogenous. It is plausible that growing up near a 4 year college would not have a direct impact on wages, and that it would usefully explain education. Many high schools in "college towns" have a reputation for being highly educated.

3.

The results of the first stage regression of education on nearc4 and the other controls are shown below. Based on the results, nearc4 has a positive effect on education, as it holds a statistically significant p-value (.000 < .05).



```
      Source |       SS           df       MS      Number of obs   =     3,010
-------------+----------------------------------   F(15, 2994)     =    182.13
       Model |  10287.6179        15  685.841194   Prob > F        =    0.0000
    Residual |  11274.4622     2,994  3.76568542   R-squared       =    0.4771
-------------+----------------------------------   Adj R-squared   =    0.4745
       Total |  21562.0801     3,009  7.16586243   Root MSE        =    1.9405

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      nearc4 |   .3198989   .0878638     3.64   0.000     .1476194    .4921785
       exper |  -.4125334   .0336996   -12.24   0.000    -.4786101   -.3464566
     expersq |   .0008686   .0016504     0.53   0.599    -.0023674    .0041046
       black |  -.9355287   .0937348    -9.98   0.000     -1.11932   -.7517377
        smsa |   .4021825   .1048112     3.84   0.000     .1966732    .6076918
       south |  -.0516126   .1354284    -0.38   0.703    -.3171548    .2139296
      smsa66 |   .0254805   .1057692     0.24   0.810    -.1819071    .2328682
      reg662 |  -.0786363   .1871154    -0.42   0.674    -.4455241    .2882514
      reg663 |   -.027939   .1833745    -0.15   0.879    -.3874918    .3316139
      reg664 |    .117182   .2172531     0.54   0.590    -.3087984    .5431624
      reg665 |  -.2726165   .2184204    -1.25   0.212    -.7008858    .1556528
      reg666 |  -.3028147   .2370712    -1.28   0.202    -.7676536    .1620242
      reg667 |  -.2168177   .2343879    -0.93   0.355    -.6763953    .2427598
      reg668 |   .5238914   .2674749     1.96   0.050    -.0005618    1.048344
      reg669 |    .210271   .2024568     1.04   0.299    -.1866975    .6072395
       _cons |   16.63825   .2406297    69.14   0.000     16.16644    17.11007
------------------------------------------------------------------------------
```

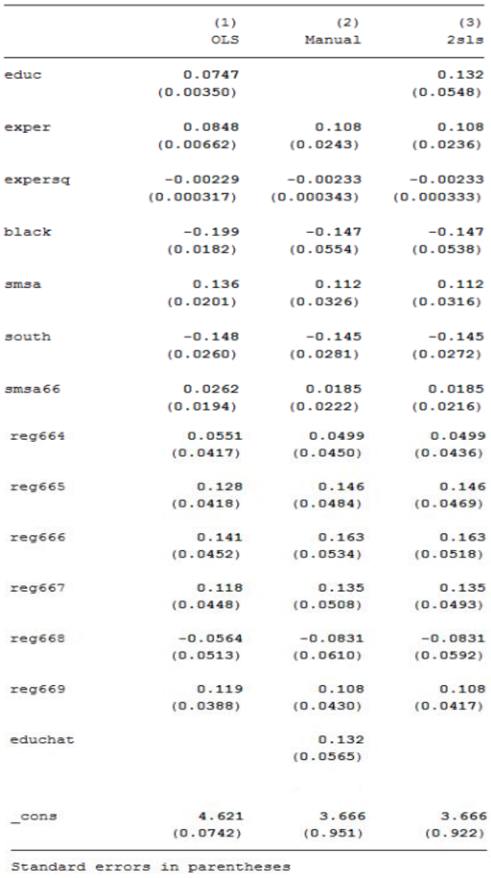
4.

In this second stage regression of education on nearc4 and the other controls (Manual column in the table below), we find that the coefficient for education has decreased. The *educhat* variable has a coefficient of .1315037, indicating that every year of education increases wages by 13.15%. This is less than the increase related to education estimated in the model found in the OLS column below. The p-value of .020 indicates that it is statistically significant at the 95% level. This implies that nearc4 had an impact being used as an instrumental variable.

5.

Running the ivregress 2sls command (2sls column in the table below), the education coefficient is .1315036 with a standard error of .0548174. This estimate is similar to the manual calculation found in Question Four, except it has a slightly smaller standard error (coefficient of .1315037 and standard error .0565104). Comparing these coefficients to the model in the OLS column, we see a much higher coefficient, and a larger standard deviation (coefficient of .0746933, and standard error of .0034983).

With this in mind, we estimate that the returns to wage from education are closer to the coefficient found in this current model,. with a 13.15% increase in wages per year of education.



	(1)	(2)	(3)
	OLS	Manual	2sls
educ	0.0747		0.132
	(0.00350)		(0.0548)
exper	0.0848	0.108	0.108
	(0.00662)	(0.0243)	(0.0236)
expersq	-0.00229	-0.00233	-0.00233
	(0.000317)	(0.000343)	(0.000333)
black	-0.199	-0.147	-0.147
	(0.0182)	(0.0554)	(0.0538)
smsa	0.136	0.112	0.112
	(0.0201)	(0.0326)	(0.0316)
south	-0.148	-0.145	-0.145
	(0.0260)	(0.0281)	(0.0272)
smsa66	0.0262	0.0185	0.0185
	(0.0194)	(0.0222)	(0.0216)
reg664	0.0551	0.0499	0.0499
	(0.0417)	(0.0450)	(0.0436)
reg665	0.128	0.146	0.146
	(0.0418)	(0.0484)	(0.0469)
reg666	0.141	0.163	0.163
	(0.0452)	(0.0534)	(0.0518)
reg667	0.118	0.135	0.135
	(0.0448)	(0.0508)	(0.0493)
reg668	-0.0564	-0.0831	-0.0831
	(0.0513)	(0.0610)	(0.0592)
reg669	0.119	0.108	0.108
	(0.0388)	(0.0430)	(0.0417)
educhat		0.132	
		(0.0565)	
_cons	4.621	3.666	3.666
	(0.0742)	(0.951)	(0.922)

Standard errors in parentheses