Economics 475: Econometrics

Homework #2

This homework is due on Monday, January 30th.

1. In class we demonstrated that the OLS estimates of $\hat{B}_1$ is an unbiased estimate of β_1. Show that $\hat{B}_0$ is an unbiased estimate of β_0. (Hint: Remember $\hat{B}_0 = \overline{Y} - \hat{B}_1 \overline{X}$). What assumptions are necessary for $\hat{B}_0$ to be an unbiased estimate of β_0?

To do this, we must show $E\left[\hat{B}_0\right] = \beta_0$. I do this in a few simple steps:

$$E\left[\hat{B}_0\right] = E\left[\overline{Y} - \hat{B}_1 \overline{X}\right] = E\left[\frac{1}{n}\sum\left(\beta_0 + \beta_1 X_i + \varepsilon_i\right) - \hat{B}_1 \overline{X}\right] = E\left[\beta_0 + \beta_1 \overline{X} + \frac{\sum \varepsilon_i}{n} - \hat{B}_1 \overline{X}\right]$$

$$= \beta_0 + \beta_1 \overline{X} + E\left[\frac{\sum \varepsilon_i}{n}\right] - \beta_1 \overline{X} = \beta_0$$

I used two assumptions for this proof: $E[\varepsilon] = 0$ and $E\left[\hat{B}_1\right] = \beta_1$. The first of these is identical to the assumptions used in the Gauss Markov proof we completed in class. The second of these is true, by the Gauss Markov proof, only if $E[\varepsilon] = 0$ and $E[\varepsilon X] = 0$. Thus, $\hat{B}_0$ is an unbiased estimate of β_0 under exactly the same conditions that $\hat{B}_1$ is an unbiased estimate of β_1.

2. Open the data set, "Whatcom County Homesales" posted on my website. This data consists of observations from all home sales in the year 2000 in Whatcom County.

The data are defined as:
Area: A code for the home's location within Whatcom County
Number: The numerical portion of the home's address
Address: The street portion of the home's address
New: Binary equal to 1 if home is new
Month: The month of home sale (1 = January, 2 = February)
Price: The home's sale price
Sqft: Square footage of house
Style: A categorical variable indicating style
Yr_Built: Year the house was built
Bedrooms: # of home's bedrooms
Age: 2000 – Yr_Built
lnprice: Natural log of Price

Consider the regression:

$$\text{lnprice}_i = \beta_0 + \beta_1 \text{sqft}_i + \beta_2 \text{sqft}_i^2 + \beta_3 \text{bedrooms}_i + \beta_4 \text{age}_i$$

a. Estimate the regression above and interpret the coefficients. Carefully describe the relationship between the home price and square footage.
To do this, I conduct the following:

```
. gen sqft2 = sqft^2

. reg lnprice sqft sqft2 bedrooms age
```

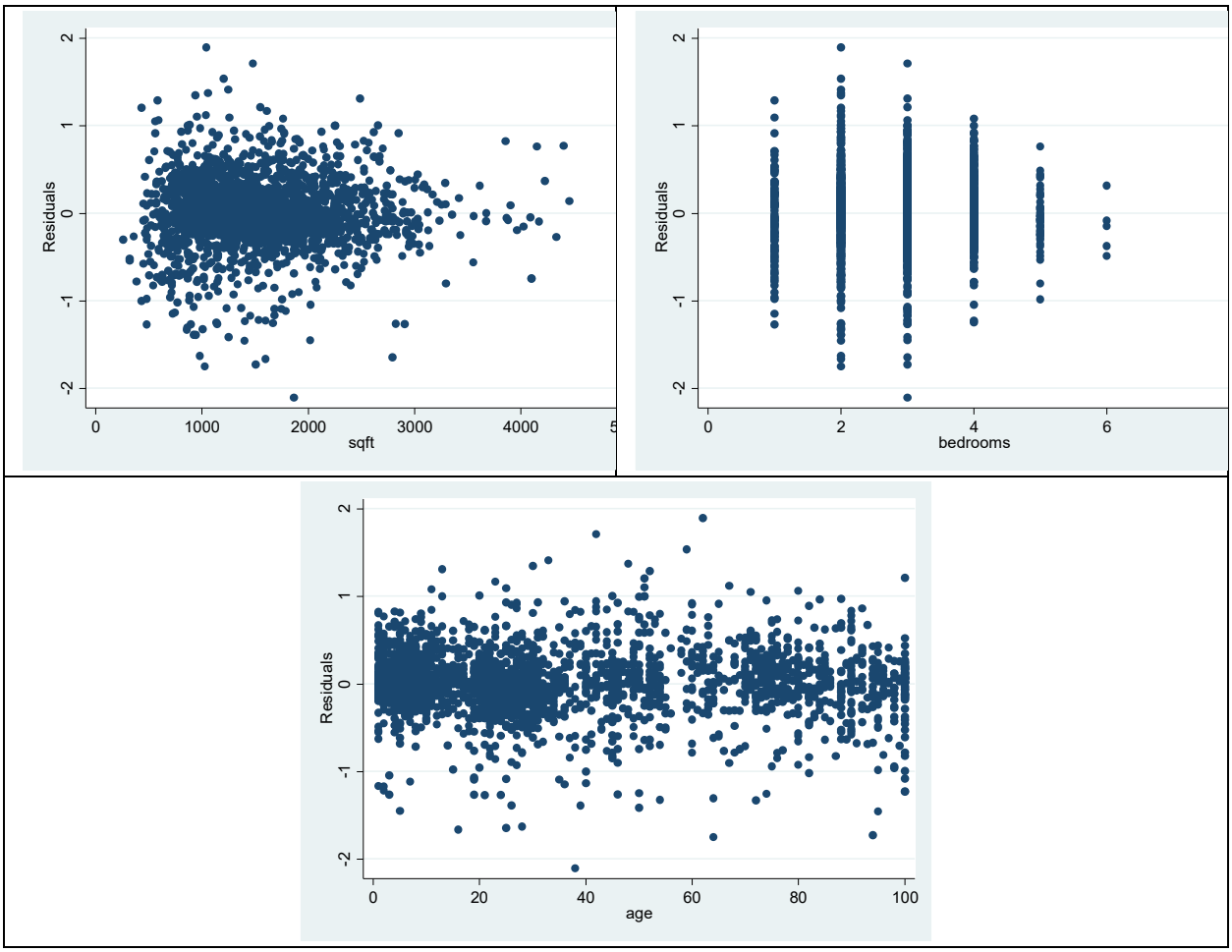
Source	SS	df	MS	Number of obs	=	2,476
				F(4, 2471)	=	578.08
Model	307.460319	4	76.8650798	Prob > F	=	0.0000
Residual	328.56202	2,471	.132967228	R-squared	=	0.4834
				Adj R-squared	=	0.4826
Total	636.02234	2,475	.256978723	Root MSE	=	.36465

lnprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sqft	.0008863	.0000508	17.46	0.000	.0007868	.0009859
sqft2	-8.90e-08	1.24e-08	-7.17	0.000	-1.13e-07	-6.46e-08
bedrooms	.0074663	.0112865	0.66	0.508	-.0146657	.0295983
age	-.0010433	.0002597	-4.02	0.000	-.0015525	-.0005341
_cons	10.78723	.0451769	238.78	0.000	10.69864	10.87582

Interpreting the impact of sqft on lnprice is not straightforward since sqft shows up as a quadratic. The appropriate interpretation is to take the derivative of lnprice with respect to sqft or $\frac{dlnprice}{dsqft} = .00088 - .000000178 \times Sqft$. So, a one unit increase in sqft changes the price of a home by a .00088 - .0000178 percent times the amount of sqft of the house

b. The coefficient on bedrooms turns out to be not statistically different than zero. However, it seems that people like homes with more bedrooms. What explains this odd result?
Remember, OLS controls for other factors. Usually, when we think about a home with more bedrooms, we also think about a larger home. However, this OLS model controls for the square foot of the house. So, the appropriate way to think of the coefficients on bedrooms is: given two houses both with the same square foot (and age), one with an additional bedroom will sell for the same price as one without the additional bedroom.

c. Use the residuals from the regression in part a and create a plot of the residuals and an independent variable (your choice) to search for heteroskedasticity. What do you find? (Question: is it appropriate to search for heterskedasticity by plotting residuals against one of the two sqft variables?)



It is dangerous to identify heteroskedasticity by looking at plots, since often the quantity of data is hidden by points on top of each other. However, in this case it seems as if two and three bedroom homes have a much wider distribution of prices than one and four bedroom homes. This suggests heteroskedasticity. There also may be wider variation of home price around smaller square foot homes, though this is harder to see.

d. Perform a Park Test on Age. Does this test indicate a heteroskedasticity problem?

I find:

```
. gen lnresid2 = log(resid2)

. reg lnresid2 age
```

Source	SS	df	MS	Number of obs	=	2,476
				F(1, 2474)	=	29.23
Model	178.573648	1	178.573648	Prob > F	=	0.0000
Residual	15112.1032	2,474	6.1083683	R-squared	=	0.0117
				Adj R-squared	=	0.0113
Total	15290.6768	2,475	6.17805125	Root MSE	=	2.4715

lnresid2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0090166	.0016676	5.41	0.000	.0057466	.0122867
_cons	-4.159215	.077157	-53.91	0.000	-4.310514	-4.007916

It appears that older homes have larger squared residuals. To the extent that the residuals are unbiased estimates of the error terms, this suggests that home price is heteroskedastic in relation to home age.

You can also do this test without the natural log of the residuals.

e. Perform a White test on the regression in part a. Do you find heteroskedasticity? Describe the pros and cons of the White test versus the Park test.

The White test is a more comprehensive test for heteroskedasticity in that it checks many sources of heteroskedasticity rather than simply one variable. It is a little harder to construct, since one has to create many independent variables to account for different functional forms of heteroskedasticity. In this case, I construct squared independent variables (though, since sqft is already squared, I don't square that again) and I construct all possible interaction terms among the independent variables:

```
. gen age2 = age^2

. gen bedrooms2 = bedrooms^2

. gen sqftage = sqft*age

. gen sqftbedrooms = sqft*bedrooms

. gen agebedrooms = age*bedrooms

. reg resid2 sqft sqft2 age age2 bedrooms bedrooms2 sqftage sqftbedrooms agebedrooms
```

Source	SS	df	MS	Number of obs	=	2,476
				F(9, 2466)	=	9.30
Model	7.77271535	9	.863635039	Prob > F	=	0.0000
Residual	229.113719	2,466	.092909051	R-squared	=	0.0328
				Adj R-squared	=	0.0293
Total	236.886434	2,475	.095711691	Root MSE	=	.30481

resid2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sqft	-.0000696	.0000565	-1.23	0.218	-.0001804	.0000413
sqft2	1.53e-08	1.39e-08	1.10	0.273	-1.20e-08	4.27e-08
age	.0016104	.0011832	1.36	0.174	-.0007098	.0039307
age2	-.0000292	8.49e-06	-3.44	0.001	-.0000458	-.0000125
bedrooms	-.1068397	.0389948	-2.74	0.006	-.1833056	-.0303737
bedrooms2	.0066386	.007565	0.88	0.380	-.0081957	.021473
sqftage	1.09e-06	4.92e-07	2.21	0.027	1.23e-07	2.05e-06
sqftbedrooms	8.14e-06	.0000183	0.44	0.657	-.0000278	.0000441
agebedrooms	.000181	.0003001	0.60	0.546	-.0004073	.0007694
_cons	.3407687	.0723818	4.71	0.000	.1988333	.4827042

In this case, my F-statistic is 9.3 and an F critical value with 9 and 2466 degrees of freedom at the 95% level is 1.883 so I reject the null hypothesis and conclude that there is heteroskedasticity.

Remember, there is no "correct" structural form for the White test. You can add as many independent variables (and their products, polynomials, etc.) but if you don't add something that does belong there, then you will get biased results (think of omitted variable bias).

f. Regardless of your answers to parts c through e, imagine that heteroskedasticity existed in the regression of part a. Specifically, assume that the $\text{Var}(\varepsilon) = \text{Age}_i \times \sigma^2$. Use the weighted least squares technique to correct for this type of heteroskedasticity and make comparisons to your original regression in part a.

If the variance of the error term is proportional to Age (and that is a big IF), then the appropriate weighting system is to multiply my regression by the inverse of the square root of age:

```
. gen weight = 1/(age^.5)

. gen wlnprice = weight*lnprice

. gen wsqft = weight*sqft

. gen wsqft2 = weight*sqft2

. gen wage = weight*age

. gen wbedrooms = weight*bedrooms

. reg wlnprice weight wsqft wsqft2 wage wbedrooms, noconstant
```

Source	SS	df	MS	Number of obs	=	2,476
				F(5, 2471)	>	99999.00
Model	37091.3181	5	7418.26362	Prob > F	=	0.0000
Residual	25.0617416	2,471	.010142348	R-squared	=	0.9993
				Adj R-squared	=	0.9993
Total	37116.3799	2,476	14.9904604	Root MSE	=	.10071

wlnprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	10.93674	.0451695	242.13	0.000	10.84817	11.02532
wsqft	.0008125	.0000457	17.79	0.000	.000723	.0009021
wsqft2	-6.69e-08	1.06e-08	-6.31	0.000	-8.77e-08	-4.61e-08
wage	-.0019259	.0004117	-4.68	0.000	-.0027333	-.0011185
wbedrooms	-.0155367	.0106036	-1.47	0.143	-.0363296	.0052562

Notice, in this case we do not include a constant—instead the coefficient on weight represents the constant.

g. Using the weighted least squares technique based upon Age in part f, has the heteroskedasticity problem been eliminated?

I can check this by performing another White test, but this time on the residuals from the model in part f and using the weighted coefficients from part f on the right hand side (I can also construct squared and interacted terms of the weighted coefficients and put on the right hand side of the White test):

```
. predict resid, resid

. gen resid2 = resid^2

. gen wage2 = wage^2

. gen wbedrooms2 = wbedrooms^2

. gen wsqftwage = wsqft*wage

. gen wsqftwbedrooms = wsqft*wbedrooms

. gen wagewbedrooms = wage*wbedrooms

. reg resid2 wsqft wsqft2 wage wbedrooms wage2 wbedrooms2 wsqftwage wsqftwbedrooms wage
> wbedrooms
```

Source	SS	df	MS		
Model	.636847362	9	.070760818	Number of obs	= 2,476
Residual	5.33672091	2,466	.00216412	F(9, 2466)	= 32.70
				Prob > F	= 0.0000
Total	5.97356827	2,475	.002413563	R-squared	= 0.1066
				Adj R-squared	= 0.1034
				Root MSE	= .04652

resid2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
wsqft	.0000337	.0000304	1.11	0.268	-.0000259	.0000932
wsqft2	-1.19e-08	4.25e-09	-2.80	0.005	-2.02e-08	-3.57e-09
wage	-.005594	.0050914	-1.10	0.272	-.015578	.0043899
wbedrooms	-.0171996	.0166255	-1.03	0.301	-.049801	.0154017
wage2	.0003519	.0003265	1.08	0.281	-.0002883	.0009921
wbedrooms2	.0028855	.0036083	0.80	0.424	-.0041902	.0099611
wsqftwage	9.09e-07	5.25e-06	0.17	0.863	-9.40e-06	.0000112
wsqftwbedrooms	.0000125	7.77e-06	1.61	0.107	-2.71e-06	.0000278
wagewbedrooms	-.0018106	.0030419	-0.60	0.552	-.0077755	.0041542
_cons	.0298929	.0196	1.53	0.127	-.0085413	.0683271

It turns out that I still have heteroskedasticity, even after weighting with the inverse of the square root of Age. Clearly, the appropriate weight is not this. Of course, it could be because heteroskedasticity is a function of something other than Age (perhaps Age^2, or $Age^{.631}$, or something else) but more likely it is a function of other things (one tip-off is that the coefficient on the $sqft^2$ is statistically different from zero in the above test).

h. Rather than knowing the form of the heteroskedastiticy as given in part f, it is unlikely (often impossible) to know the true form of the heteroskedasticity. Using the original regression in part a, re-estimate this model using Feasible GLS. Compare this estimator to that presented in part a.

GLS requires you to choose a potential function form for heteroskedasticity. One candidate that encompasses many, but not all, possibilities is $Var(\varepsilon) = Exp[\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots]$. One can add as many squared or interacted terms to the right hand side of this as one wants. Remember, omitting a necessary

variable will bias your estimates of the variance of ε so it is important to include as many things as you can think of:

```
. predict resid, resid

. gen resid2 = resid^2

. gen lnresid2 = log(resid2)

. reg lnresid2 sqft sqft2 bedrooms bedrooms2 age age2 sqftage sqftbedrooms agebedrooms
```

Source	SS	df	MS	Number of obs	=	2,476
				F(9, 2466)	=	12.80
Model	682.287323	9	75.8097025	Prob > F	=	0.0000
Residual	14608.3895	2,466	5.92392113	R-squared	=	0.0446
				Adj R-squared	=	0.0411
Total	15290.6768	2,475	6.17805125	Root MSE	=	2.4339

lnresid2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sqft	.000041	.0004514	0.09	0.928	-.0008442	.0009262
sqft2	4.38e-08	1.11e-07	0.39	0.694	-1.75e-07	2.62e-07
bedrooms	-1.611883	.3113741	-5.18	0.000	-2.222465	-1.001302
bedrooms2	.1698266	.0604063	2.81	0.005	.0513743	.2882789
age	.0252233	.0094482	2.67	0.008	.006696	.0437506
age2	-.000285	.0000678	-4.20	0.000	-.000418	-.0001521
sqftage	4.02e-06	3.93e-06	1.02	0.306	-3.69e-06	.0000117
sqftbedrooms	.0000881	.0001465	0.60	0.548	-.0001992	.0003754
agebedrooms	.0015952	.002396	0.67	0.506	-.0031031	.0062936
_cons	-1.958642	.5779698	-3.39	0.001	-3.091999	-.8252859

```
. predict h, xb

. gen g = exp(h)
```

```
. gen weight = 1/(g^.5)

. drop wlnprice- wagewbedrooms

. gen wlnprice = weight*lnprice

. gen wsqft = weight*sqft

. gen wsqft2 = weight*sqft2

. gen wage = weight*age

. gen wbedrooms = weight*bedrooms

. reg wlnprice weight wsqft wsqft2 wage wbedrooms, noconstant
```

Source	SS	df	MS	Number of obs	=	2,476
				F(5, 2471)	>	99999.00
Model	18473985.5	5	3694797.1	Prob > F	=	0.0000
Residual	14210.3081	2,471	5.75083289	R-squared	=	0.9992
				Adj R-squared	=	0.9992
Total	18488195.8	2,476	7466.96114	Root MSE	=	2.3981

wlnprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	10.834	.0523314	207.03	0.000	10.73139	10.93662
wsqft	.0007969	.0000603	13.21	0.000	.0006786	.0009151
wsqft2	-7.08e-08	1.63e-08	-4.35	0.000	-1.03e-07	-3.88e-08
wage	-.0012512	.0002479	-5.05	0.000	-.0017373	-.0007651
wbedrooms	.0246997	.0114089	2.16	0.030	.0023276	.0470718

At this point, I would want to conduct another White test to ensure that I eliminated the heteroskedasticity. If I did not, then I would want to add variables to the lnresid2 regression.

I can simplify this by simply using the White Corrected Standard errors:

```
. reg lnprice sqft sqft2 age bedrooms, robust

Linear regression                               Number of obs     =      2,476
                                                F(4, 2471)        =     486.77
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4834
                                                Root MSE          =     .36465
```

lnprice	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
sqft	.0008863	.0000605	14.65	0.000	.0007677	.001005
sqft2	-8.90e-08	1.50e-08	-5.92	0.000	-1.18e-07	-5.95e-08
age	-.0010433	.0002707	-3.85	0.000	-.0015741	-.0005125
bedrooms	.0074663	.0115988	0.64	0.520	-.0152782	.0302107
_cons	10.78723	.0546976	197.22	0.000	10.67997	10.89449

When I do this, I see that the "correct" standard error are very similar to those of my FGLS regression. This suggests I did a good job in that weighted regression. It also begs the question of why we just didn't use the White Correction to begin with (which, of course, researchers do—we rarely use FGLS anymore because computers can so easily compute the White Correction today). The White's correction also

avoids the problem of trying to estimate the functional form of the heteroscedasticity—clearly a problem for a researcher that doesn't know this functional form.

3. Using your final project data, answer the following questions.

a. Describe each variable to me. What is your dependent variable? Independent variable(s)? What do they measure? Where do they come from?

b. Estimate a regression using your variables. Show me your results. Describe what you are looking for in this regression.

c. Does your regression have heteroskedasticity?