Economics 475: Econometrics
Homework #2
This homework is due on Monday, January 30th.

1. In class we demonstrated that the OLS estimates of B1 is an unbiased estimate of β 1. Show that $B0$ is an unbiased estimate of β_0. (Hint: Remember B0YB1 X). What assumptions are necessary for B0 to be an unbiased estimate of β_0?

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$$
$$\bar{Y} - \hat{\beta}_1 X_1 = \hat{\beta}_0$$
$$E(\beta_0) = E(Y) - E(\beta_1), \text{ where } E(\hat{\beta}_0\ E(Y) = \mu = \bar{Y}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 X$$

$\hat{\beta}_0$ is an unbiased estimate of $\hat{\beta}_0$ only if the OLS assumptions hold true. These assumptions include: *linearity* (the relationship between X and Y is linear), *independence* (the observations are independent), *homoskedasticity* ($var(\varepsilon_i)$ is constant ∀ i), and *normality* (errors are normally distributed).

2. Open the data set, "Whatcom County Homesales" posted on my website. This data consists ofobservations 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
CountyNumber: 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
styleYr_Built: Year the house was built
Bedrooms: # of home's bedrooms
Age: 2000 – Yr_Built
lnprice: Natural log of
Price

Consider the regression: $lnprice_i = \beta 0 + \beta 1 sqft_i + \beta 2 sqft^2 + \beta 3 bedrooms_i + \beta 4 age_i$

a. Estimate the regression above and interpret the coefficients. Carefully describe the relationshipbetween the home price and square footage.

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	Coefficient	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

Figure 1. Regression of lnprice

$$\beta 0 \approx 10.78723, \beta 1 \approx 0.0008863, \ \beta 2 \approx -8.90e^{-8}, \ \beta 3 \approx -0.0010433$$

- $\beta 0$ indicates that the lnprice-intercept of the regression line is roughly 10.78723.
- $\beta 1$ indicates that as sqft increases by 1 unit, the percent change of price increases on average by 0.0008863 percent.
- $\beta 2$ indicates that as sqft^2 increases by 1 unit, the percent change diminishes on average by $-8.9e^{8}$ percent.
- β4 indicates that as age increases by 1 unit, the percent change of price diminishes on average by -0.0010433 percent.

b. The coefficient on bedrooms turns out to be not statistically different than zero. However, it seemsthat people like homes with more bedrooms. What explains this odd result?

With the coefficient of bedrooms being statistically similar to zero, we can interpret that the increase of bedrooms by 1 unit, should not significantly raise the sell price of houses, which seems counterintuitive. One hypothesis for this phenomenon could be that the bedrooms variable could be dependent on another variable such as sqft and sqft^2. This means that bedrooms could have significant covariance or multicollinearity with another variable, leading to a skewed coefficient of bedrooms.

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 tosearch for heteroskedasticity by plotting residuals against one of the two sqft variables?)

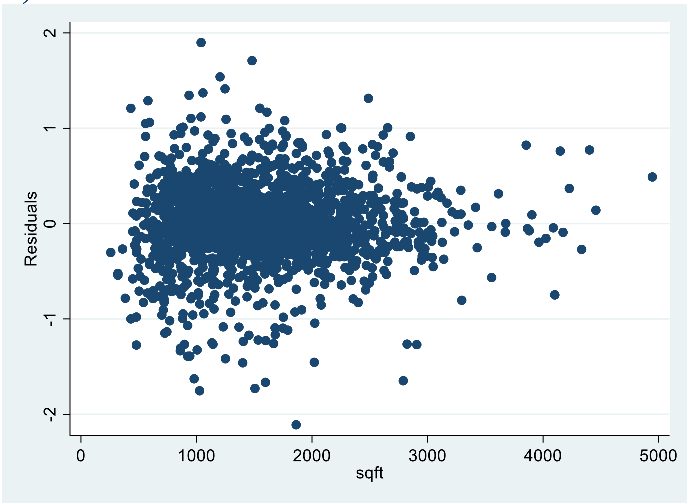

Figure 2. Scatter plot of Residuals against sqft

There seems to be a decreasing trend when we create a scatterplot using residuals and sqft, implying some level of heteroskedasticity. However, it is not appropriate to search for heteroskedasticity by plotting residuals against either of the sqft variables as they are two correlated variables.

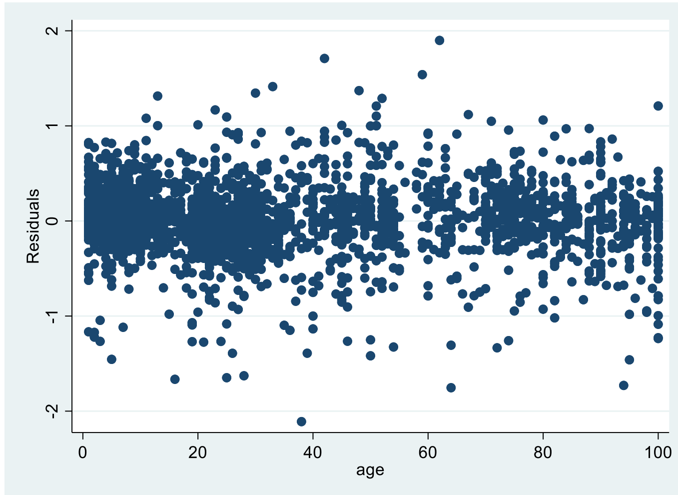

Figure 3. Scatter plot of Residuals against age

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

Source	SS	df	MS		
				Number of obs =	2,476
				F(1, 2475) =	370.20
Model	36.4948435	1	36.4948435	Prob > F =	0.0000
Residual	243.991349	2,475	.098582363	R-squared =	0.1301
				Adj R-squared =	0.1298
Total	280.486192	2,476	.113281984	Root MSE =	.31398

res2	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	.002624	.0001364	19.24	0.000	.0023566	.0028914

Figure 4. Parks Test regression on age

We can look at the coefficient and the t-score to analyze if this problem is heteroskedastic. Since the coefficient of age is not zero and t-score is 19.24, which is larger than the t-critical value of 1.96 with a 95% confidence interval. So, the Parks Test indicates heteroskedasticity.

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.

Source	SS	df	MS		
				Number of obs =	2,476
				F(13, 2462) =	7.19
Model	8.66102124	13	.666232403	Prob > F =	0.0000
Residual	228.225413	2,462	.092699193	R-squared =	0.0366
				Adj R-squared =	0.0315
Total	236.886434	2,475	.095711691	Root MSE =	.30447

res2	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
sqft	-.0008949	.0002919	-3.07	0.002	-.0014673	-.0003224
sqft2	6.89e-07	2.80e-07	2.46	0.014	1.40e-07	1.24e-06
bedrooms	-.0257112	.0931173	-0.28	0.782	-.2083075	.1568851
age	.0015238	.001186	1.28	0.199	-.0008018	.0038495
sqft2	0	(omitted)				
sqftsqft2	-2.03e-10	1.19e-10	-1.71	0.087	-4.36e-10	2.95e-11
sqft2bedrooms	1.63e-08	4.68e-08	0.35	0.728	-7.55e-08	1.08e-07
sqftage	1.36e-06	5.02e-07	2.70	0.007	3.73e-07	2.34e-06
sqftbedrooms	-.0000617	.0001324	-0.47	0.641	-.0003213	.0001979
bedroomsage	.000064	.0003037	0.21	0.833	-.0005315	.0006595
sqft22	2.05e-14	1.64e-14	1.25	0.213	-1.17e-14	5.27e-14
bedrooms2	.0061001	.0078344	0.78	0.436	-.0092626	.0214628
age2	-.0000282	8.50e-06	-3.32	0.001	-.0000449	-.0000116
sqft22bedrooms	-3.02e-16	1.36e-15	-0.22	0.824	-2.96e-15	2.36e-15
_cons	.6064764	.1427992	4.25	0.000	.3264575	.8864954

Figure 4. White Test on the regression in part a.

By looking at the F-statistic and the p-value, we can see a significantly large F-statistic of 7.19 and a p-value of 0, indicating that the residuals have a significant amount of variation that is not explained by the predictor variables. This suggests that the error variance is non-constant across all of predictor variables thus indicating heteroskedasticity.

Parks Test		**White Test**	
Pros	**Cons**	**Pros**	**Cons**
Relatively simple test	Requires all assumptions of OLS to be true	More "powerful" version of Parks Test	Can be lengthy and impractical
With first-order approximations, given all assumptions of OLS are satisfied, Parks Test is all you need	Can only reliably be used on first-order approximations	Can be used for first and second-order approximations	Could give significant values from other underlying issues other than heteroskedasticity

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

Source	SS	df	MS
Model	17414948.1	5	3482989.62
Residual	216349.235	2,471	87.555336
Total	17631297.3	2,476	7120.87938

Number of obs	=	2,476
F(5, 2471)	=	39780.44
Prob > F	=	0.0000
R-squared	=	0.9877
Adj R-squared	=	0.9877
Root MSE	=	9.3571

wlnprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
w	30.65357	1.327886	23.08	0.000	28.04969	33.25745
wsqft	.0098307	.0001563	62.91	0.000	.0095243	.0101371
wsqft2	-2.58e-06	4.27e-08	-60.37	0.000	-2.66e-06	-2.50e-06
wbedrooms	.4783409	.0391459	12.22	0.000	.4015788	.5551029
wage	.0302422	.0011363	26.61	0.000	.0280139	.0324705

Figure 5. WLS correction on regression on lnprice from part a.

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

When considering the F-statistic and P-value, we can see a significantly large F-statistic of 39780.44 and P-value of 0, indicating that the residuals have a significant amount of variation that is not explained by our correction. This suggests that the error variance is non-constant across all of predictor variables thus still indicating heteroskedasticity.

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.

Source	SS	df	MS			
				Number of obs	=	2,476
				F(5, 2471)	>	99999.00
Model	17616850.3	5	3523370.05	Prob > F	=	0.0000
Residual	14447.0672	2,471	5.84664798	R-squared	=	0.9992
				Adj R-squared	=	0.9992
Total	17631297.3	2,476	7120.87938	Root MSE	=	2.418

wlnprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
w	10.7741	.0522539	206.19	0.000	10.67164	10.87657
wsqft	.0008943	.0000628	14.25	0.000	.0007712	.0010174
wsqft2	-9.62e-08	1.73e-08	-5.57	0.000	-1.30e-07	-6.23e-08
wbedrooms	.0148552	.0101573	1.46	0.144	-.0050626	.0347729
wage	-.0011159	.0002713	-4.11	0.000	-.001648	-.0005839

Figure 7. FGLS re-estimate of part a.

Part a: $\beta 0 \approx 10.78723$ Part h: $\beta 0 \approx 10.7741$

The estimators in the original model compared to our FGLS model are fairly similar, implying that there is minimal heteroskedasticity in our original model. However, our f-statistic is even higher than it was it part a. This suggests that I most likely calculated the weight incorrectly or even used a wrong variable. It could also mean that there is some form of heteroskedasticity that we have not addressed yet.

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?

Data:	**Sourced from:**	
Exit Survey of Undergraduate Students Completing Degrees in the Spring of 2010	This data is sourced from a survey conducted by the Office of Survey Research at Western Washington University in 2010. The data is “a mixture of open-ended, multiple-choice, and numerical response questions” (WWU OSR 2011)	
Dependent Variable:	**Purpose:**	**Sourced by/from:**
Job Offer Success Rate of Undergraduates at WWU	to analyze the success rate of obtaining job offers dependent on specific variables discussed below	• Section B.15.b. on page 16 • “how many job offers have you received” • will be a percentage of total observations
Independent Variables:		
Field of Study (FOS)	to categorize students to their general field of study	• Sourced from part A. • lays out which college the respondent graduated from • data will most likely be in the form of 8 dummy variables.
Mean GPA (GPA)	to assess the students’ perceived success in school versus its reliability to obtain job offers	• A.2 makes each college’s mean GPA.

Support Satisfaction (SUP)	to factor students' experience with WWU professors and willingness to ask for help	• C.2 measures relationships with other faculty on a scale from 1-7 • Or C.6 that asks how often did you talk with an advisor or faculty member about career plans and attend learning events on campus? • Organized by college • Outputting means of responses
Graduation Delays (GRAD)	to analyze the effect of graduating late	• C.4.a asks the students what reasons they had for graduating late • Will utilize C.4. to assess the percentage of students that took longer to graduate than expected from each school
Further Education (EDU)	to account for the students who plan on delaying job searching because of further education	• C.12.b asks the students where they are with their application process • will focus on the totals of each college
Debt (DEBT)	to account for the possible urgency to look for a job	• C.11. provides the number of students in each college who borrowed money to fund their education. • C.11.a. provides the mean debt from each college.

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?

```
*Open Dataset*
cd "C:\Users\leej207\Desktop\475 HW\HW2"
use "Whatcom County Homesales.dta"

*2a) generate regression*
gen lnprice = ln(price)
gen sqft2 = sqft^2
reg lnprice sqft sqft2 bedrooms age

*2c) Residuals from regression to create a plot of residuals and an independent variable to search for heteroskedasticity*
predict res, resid
scatter res sqft

 *2d) Park's test on Age*
 gen res2 =  res^2
 reg res2 age, noconstant

 *2e) White Test on part a*
drop res2
gen sqft22 = sqft2^2
gen sqftsqft2 =  sqft*sqft2
gen sqft2bedrooms = sqft2*bedrooms
gen bedroomsage = bedrooms*age
gen res2 = res^2
gen sqft22 = sqft2^2
gen bedrooms2 = bedrooms^2
gen sqft22bedrooms = sqft22*bedrooms
gen sqftage = sqft*age
gen bedrooms2 = bedrooms^2
gen age2 = age^2
gen sqftbedrooms = sqft*bedrooms
reg res2 sqft sqft2 bedrooms age sqft2 sqftsqft2 sqft2bedrooms sqftage sqft22 bedrooms2 age2 sqft22bedrooms bedroomsage sqft22 bedrooms2 age2 sqftbedrooms

*2f) WLS technique
drop w
gen w = 1/(age^0.5)
gen wlnprice = w*lnprice
gen wsqft = w*sqft
gen wsqft2 = w*sqft2
gen wbedrooms = w*bedrooms
gen wage = w*age
reg wlnprice w wsqft wsqft2 wbedrooms wage, noconstant

*2h) FGLS
drop res2
drop res
reg lnprice sqft sqft2 bedrooms age
predict res, resid
gen res2 = res^2
gen lnres2 = ln(res2)
reg lnres2 sqft sqft2 bedrooms age
predict res2ageres, xb
gen e_res2ageres = exp(res2ageres)
gen w= 1/(e_res2ageres^0.5)
gen wlnprice = w*lnprice
gen wsqft = w*sqft
gen wsqft2 = w*sqft2
gen wbedrooms = w*bedrooms
gen wage = w*age
reg wlnprice w wsqft wsqft2 wbedrooms wage, noconstant
```