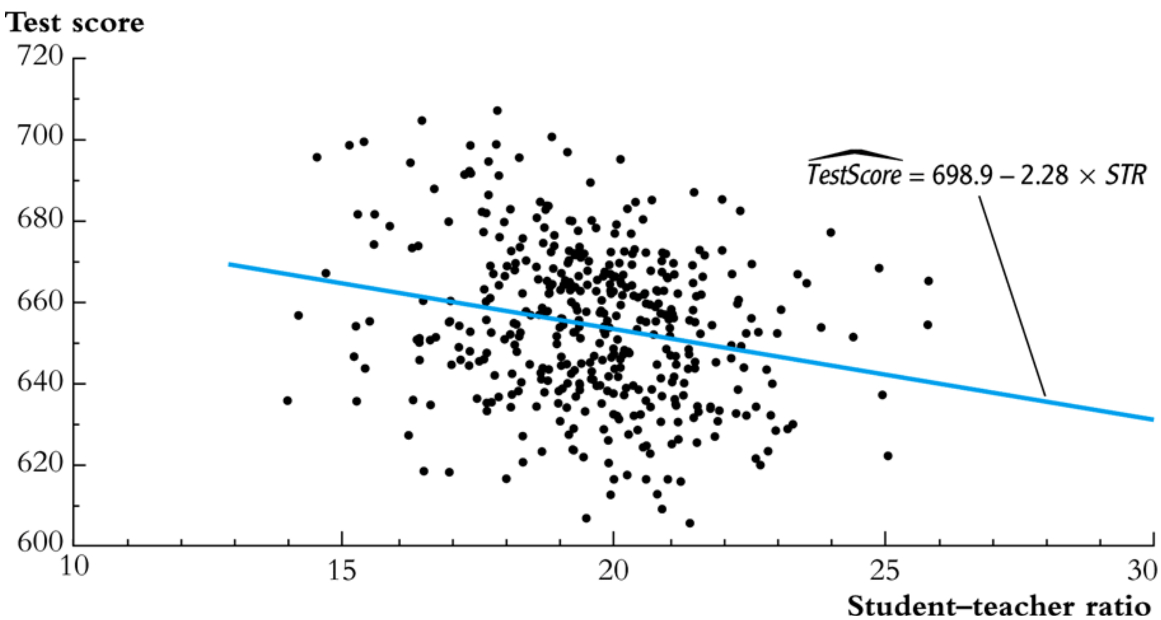
David Gabel Econometrics Fall 2017

Happy Holidays!



On October 19, 2017 I put the following graph on the board and posed the question, does the error term appear to be *Heteroskedastic or homoskedastic?* The following two tables provide the Stata output for this data set using Homoskedasticity-only standard errors and heteroskedasticity – robust standard errors. Based on your reading of the graph, as well as the two tables, do you think the error term is best characterized as homoscedastic or heteroskedastic? Does it make a substantial difference in this problem (note: as always, the coefficient estimates are identical).

```
. * OLS

. reg AverageTestScore StudentTeacherRatio

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11004         1  7794.11004   Prob > F        =    0.0000
    Residual |  144315.484       418  345.252353   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030056   Root MSE        =    18.581

------------------------------------------------------------------------------
AverageTes~e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
StudentTea~o |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------
```

```
. * robust

. reg AverageTestScore StudentTeacherRatio, robust

Linear regression                               Number of obs     =        420
                                                F(1, 418)         =      19.26
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0512
                                                Root MSE          =     18.581

--------------------------------------------------------------------------------
------
                    |               Robust
   AverageTestScore |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------------+-----------------------------------------------------------
------
StudentTeacherRatio |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.259
              _cons |    698.933   10.36436    67.44   0.000     678.5602    719.31
```

---

project. Run regression with sat as explanatory variable and get only 690 observations. If exclude sat, over 2, 000 observations. Point out large change in coefficients. Is this all due to omitted variable bias?
reg mr_kq5_pq1 flagship black_share_fall_2000 hisp_share_fall_2000
asian_or_pacific_share_fall_2000 tier if sat_avg_2013 <2000& sat_avg_2001 <2000, robust
what if ran regression for 690 observations, but exclude sat scores. Is this an equally large change in coefficients?
reg mr_kq5_pq1 flagship black_share_fall_2000 hisp_share_fall_2000
asian_or_pacific_share_fall_2000 tier sat_avg_2013 sat_avg_2001 , robust
vs full data set and no sat

reg scorecard *median*earnings_2011 iclevel region type public
drop iclevel
big drop in r2. Does this prove omitted variable bias? No. correlation. This done by bouzi. Try some other variable other than type since this is somewhat redundant with public. Could use this as example of multicollinearity.

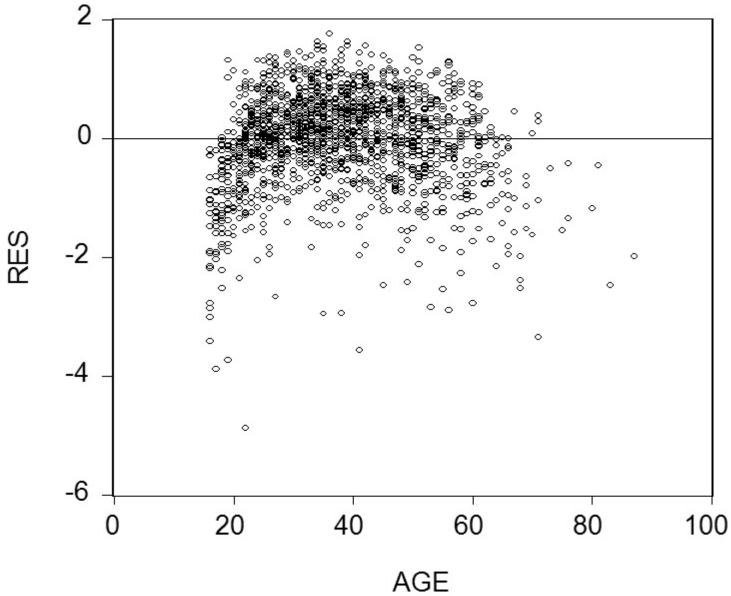
Could do earnings function reg scorecard_median_earnings_2011 iclevel public temp tier
Where temp =public*iclevel; transform tier to different dummy variables

Does earnings belong as an explanatory variable for mobility? Is this what should be replicated by other schools?

I would have liked to see someone explore why P(parents in q1) varies across states. For example, pell grants over 100% of the tuition in NY. Does something equivalent exist in NC?

You have learned that earnings functions are one of the most investigated relationships in economics. These typically relate the logarithm of earnings to a series of explanatory variables such as education, work experience, gender, race, etc.
(a) Why do you think that researchers have preferred a log-linear specification over a linear specification? In

addition to the interpretation of the slope coefficients, also think about the distribution of the error term.
(b) To establish age-earnings profiles, you regress ln(*Earn*) on Age, where *Earn* is weekly earnings in dollars, and *Age* is in years. Plotting the residuals of the regression against age for 1,744 individuals looks as shown in the figure:



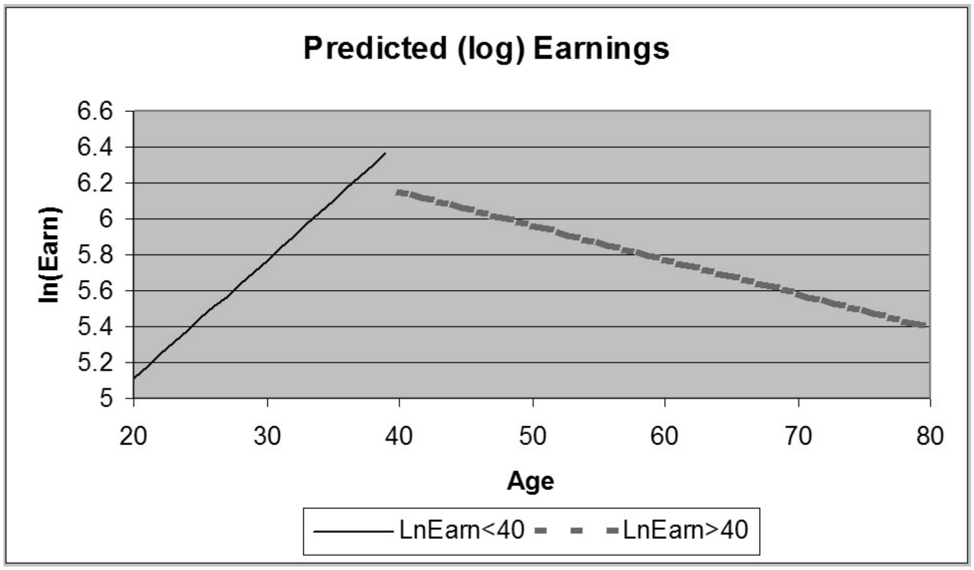
Do you sense a problem?
(c) You decide, given your knowledge of age-earning profiles, to allow the regression line to differ for the below and above 40 years age category. Accordingly you create a binary variable, *Dage*, that takes the value one for age 39 and below, and is zero otherwise. Estimating the earnings equation results in the following output (using heteroskedasticity-robust standard errors):

$$\widehat{LnEarn} = \underset{(38.33)}{6.92} - \underset{(0.22)}{3.13} \times Dage - \underset{(0.004)}{0.019} \times Age + \underset{(0.005)}{0.085} \times (Dage \times Age),\ R^2=0.20,\ SER=0.721.$$

Sketch both regression lines: one for the age category 39 years and under, and one for 40 and above. Does it make sense to have a negative sign on the Age coefficient? Predict the ln(*earnings*) for a 30 year old and a 50 year old. What is the percentage difference between these two?
(d) The F-statistic for the hypothesis that both slopes and intercepts are the same is 124.43. Can you reject the null hypothesis?

ANSWERS
(a) The error variance and the variance of the dependent variable are related. Given that the dependent variable (earnings) is not normally distributed, it is difficult to postulate that the error variance is normally distributed. Using logarithms results in a distribution that is closer to a normal. In addition, there seems to be a better fit for the log-linear specification, and the coefficients can be interpreted as percentage changes.
(b) There seems to be a pattern in the residuals when sorted by age. This suggests a misspecified functional form.
(c) According to the specification, earnings increase with age until the individual is 39 years old. It is only from age 40 onwards that the regression predicts a negative relationship between earnings and age.

According to the estimates, a 30-year-old would have ln(*earnings*) of 5.77, while the predicted value for a 50-year-old would be 5.97. The difference between the two is approximately 20 percent.



(d) The critical value from the *F*-table is 4.61 at the 1% level. Hence the null hypothesis is rejected.

Earnings functions attempt to find the determinants of earnings, using both continuous and binary variables. One of the central questions analyzed in this relationship is the returns to education.
(a) Collecting data from 253 individuals, you estimate the following relationship

$$\widehat{\ln(Earn_i)} = \underset{(0.14)}{0.54} + \underset{(0.011)}{0.083} \times Educ,\ R^2 = 0.20,\ SER = 0.445$$

where *Earn* is average hourly earnings and *Educ* is years of education.

What is the effect of an additional year of schooling? If you had a strong belief that years of high school education were different from college education, how would you modify the equation? What if your theory suggested that there was a "diploma effect"?
(b) You read in the literature that there should also be returns to on-the-job training. To approximate on-the-job training, researchers often use the so called Mincer or potential experience variable, which is defined as $Exper = Age - Educ - 6$. Explain the reasoning behind this approximation. Is it likely to resemble years of employment for various sub-groups of the labor force?
(c) You incorporate the experience variable into your original regression

$$\widehat{\ln(Earn_i)} = \underset{(0.16)}{-0.01} + \underset{(0.012)}{0.101} \times Educ + \underset{(0.006)}{0.033} \times Exper - \underset{(0.0001)}{0.0005} \times Exper^2,$$

$R^2 = 0.34$, $SER = 0.405$

What is the effect of an additional year of experience for a person who is 40 years old and had 12 years of education? What about for a person who is 60 years old with the same education background?
(d) Test for the significance of each of the coefficients of the added variables. Why has the coefficient on education changed so little? Sketch the age-(log)earnings profile for workers with 8 years of education and 16 years of education.
(e) You want to find the effect of introducing two variables, gender and marital status. Accordingly you specify a binary variable that takes on the value of one for females and is zero otherwise (*Female*), and another binary variable that is one if the worker is married but is zero otherwise (*Married*). Adding these variables to the regressors results in:

$$\widehat{\ln(Earn_i)} = \underset{(0.16)}{0.21} + \underset{(0.012)}{0.093} \times Educ + \underset{(0.006)}{0.032} \times Exper - \underset{(0.0001)}{0.0005} \times Exper^2$$
$$- \underset{(0.049)}{0.289} \times Female + \underset{(0.056)}{0.062}\, Married,$$

$R^2 = 0.43$, $SER = 0.378$

Are the coefficients of the two added binary variables individually statistically significant? Are they economically important? In percentage terms, how much less do females earn per hour, controlling for education and experience? How much more do married people make? What is the percentage difference in earnings between a single male and a married female? What is the marriage differential between males and females?

(f) In your final specification, you allow for the binary variables to interact. The results are as follows:

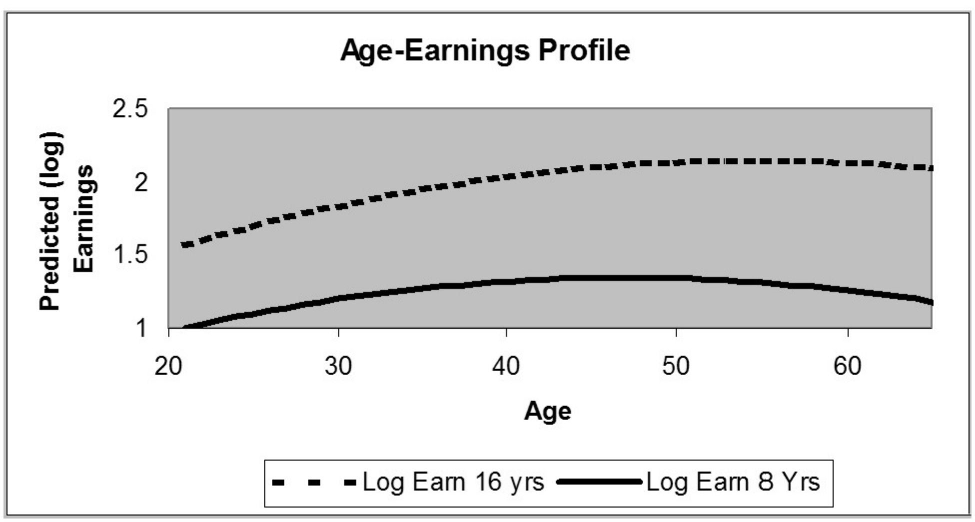
$$\widehat{\ln(Earn_i)} = \underset{(0.16)}{0.14} + \underset{(0.011)}{0.093} \times Educ + \underset{(0.006)}{0.032} \times Exper - \underset{(0.001)}{0.0005} \times Exper^2$$

$$- \underset{(0.075)}{0.158} \times Female + \underset{(0.080)}{0.173} \times Married - \underset{(0.097)}{0.218} \times (Female \times Married),$$

$R^2 = 0.44$, $SER = 0.375$

Repeat the exercise in (e) of calculating the various percentage differences between gender and marital status.

Answer:
(a) One additional year of education carries an 8.3 percent increase, or a return, on earnings. You would need additional data to see if this coefficient was different for high school versus college education. Including both variables in the regression would then allow you to test for equality of the coefficients. A "diploma effect" could be studied by creating a binary variable for a high school diploma, a junior college diploma, a B.A. or B.Sc. diploma, and so forth.
(b) The idea is that everybody works except in the first six years of life and during the time spent in school/university for education. This approximation will work better for people with a strong attachment to the labor force. It will not work well for females and those who are frequently unemployed or out of the workforce.
(c) For the first person, the *Exper* variable increases from 22 to 23, and results in a 1.1 percent earnings increase. For the 60 year old, there is an expected decrease of 1 percent.
(d) Both coefficients are highly significant using conventional levels of significance. The fact that the coefficient on the education variable hardly changed suggests that education and experience are not highly correlated.



(e) The coefficient for the female binary variable is statistically significant even at the 1% level. The coefficient for the married binary variable only has a *t*-statistic of 1.11 and is not statistically significant at the 10% level. Both coefficients indicate economic importance, since females make approximately 29 percent less than males and married people earn roughly 6 percent more. A married female earns roughly 23 percent less than a single male. Married females earn 29 percent less than married males, the same percentage that single females earn less than single males.
(f) The default is the single male. Single females earn 15.8 percent less. Married males earn 17.3 percent more. Married females earn 20.3 percent less. Comparing married females with married males now results in a percentage differential of 37.6 percent in favor of the males.

Source	SS	df	MS			
				Number of obs	=	2,084
				F(18, 2065)	=	89.50
Model	.156186344	18	.008677019	Prob > F	=	0.0000
Residual	.200196322	2,065	.000096947	R-squared	=	0.4383
				Adj R-squared	=	0.4334
Total	.356382666	2,083	.000171091	Root MSE	=	.00985

mr_kq5_pq1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tier1	-.0049063	.0060303	-0.81	0.416	-.0167324	.0069197
tier2	-.0001905	.0053377	-0.04	0.972	-.0106583	.0102773
tier3	.0083899	.0059687	1.41	0.160	-.0033154	.0200952
tier4	.0022309	.005225	0.43	0.669	-.0080159	.0124777
tier5	.0080979	.0056371	1.44	0.151	-.0029572	.0191529
tier6	.005209	.0049891	1.04	0.297	-.0045751	.0149931
tier7	.005785	.0057326	1.01	0.313	-.0054573	.0170273
tier8	.0056686	.0050777	1.12	0.264	-.0042894	.0156265
tier9	.0024706	.0056089	0.44	0.660	-.0085292	.0134703
tier10	.0013454	.005071	0.27	0.791	-.0085995	.0112902
tier11	.0024026	.0051589	0.47	0.641	-.0077147	.0125199
tier12	0	(omitted)				
public	.0002544	.002732	0.09	0.926	-.0051034	.0056121
sticker_price_2013	7.84e-08	4.89e-08	1.60	0.109	-1.75e-08	1.74e-07
scorecard_netpric~2013	-1.73e-07	5.94e-08	-2.92	0.004	-2.90e-07	-5.68e-08
asian_or_pacific_~2000	.0676247	.0044088	15.34	0.000	.0589786	.0762709
black_share_fall_2000	.0213408	.0013651	15.63	0.000	.0186636	.024018
hisp_share_fall_2000	.0536294	.0020194	26.56	0.000	.0496691	.0575897
alien_share_fall_2000	.0284841	.0072636	3.92	0.000	.0142393	.0427289
_cons	.0057157	.0050992	1.12	0.262	-.0042844	.0157158

test tier1=tier2=tier3=tier4=tier5=tier6=tier7=tier8=tier9=tier10=tier11
F(10, 2065) = 11.03
Write Ho, Ha. Do I reject

Based on the f-test at the five percent level of significance, should I drop the tier values?

I am interested in knowing the economic impact of some of the explanatory variables. I execute the ey/ex command in Stata and obtain the following table. The eleasticity estimates appear under the heading ey/ex:

	ey/ex	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
sticker_price_2013	**.0788468**	**.0490377**	**1.61**	**0.108**	**-.0173217**	**.1750154**
scorecard_netpric~2013	**-.1352265**	**.0471959**	**-2.87**	**0.004**	**-.2277829**	**-.04267**

Sticker_price is the list price of a college, and net price is defined as follows: Net Price is the amount that a student pays to attend an institution in a single academic year AFTER subtracting scholarships and grants the student receives. Scholarships and grants are forms of financial aid that a student does not have to pay back.

Which variable has a larger economic impact on the mobility rate? Explain.

Source |	SS	df	MS		
				Number of obs =	2,166
				F(13, 2152) =279.47	
Model |	97.91	13	7.53	Prob > F =	
Residual |	57.99	2,152	.027	R-squared =0.6280	
				Adj R-squared =	0.6258
Total |	155.91	2,165	.072	Root MSE =	.16417

lnearnings |	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tier2 |	-.193772	.0517062	-3.75	0.000	-.2951713	-.0923727
tier3 |	-.2915432	.0574108	-5.08	0.000	-.4041296	-.1789568
tier4 |	-.3278334	.0514185	-6.38	0.000	-.4286686	-.2269983
tier5 |	-.533204	.0485032	-10.99	0.000	-.628322	-.4380859
tier6 |	-.4783959	.048294	-9.91	0.000	-.5731036	-.3836882
tier7 |	-.7036004	.0520176	-13.53	0.000	-.8056103	-.6015905
tier8 |	-.654977	.0515035	-12.72	0.000	-.7559788	-.5539753
tier9 |	-.8132432	.0484315	-16.79	0.000	-.9082205	-.7182658
tier10 |	-.7188926	.0514106	-13.98	0.000	-.8197121	-.618073
tier11 |	-.8346025	.051589	-16.18	0.000	-.9357719	-.7334331
tier12 |	-1.036492	.0538835	-19.24	0.000	-1.142161	-.9308228
pct_stem|	.0035129	.0002637	13.32	0.000	.0029957	.0040301
region |	-.0197212	.0035467	-5.56	0.000	-.0266766	-.0127658
_cons |	11.111	.0486186	228.53	0.000	11.01565	11.20634

F TEST OVERALL FITNESS OF MODEL
REGION. PROVIDE DESCRIPTION. DO YOU AGREE WITH WAY VARIABLE HAS BEEN INCLUDED? WOULD YOU PROPOSE CHANGE. EXPLAIN.