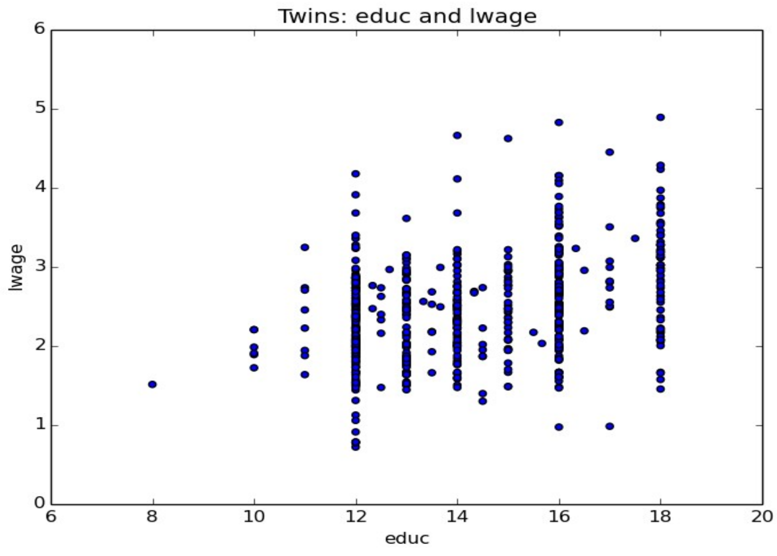
Rudy Gilman

Applied Econometrics, Problem Set 1, March 1, 2016

1. a) Run the bivariate regression of log-wages on a constant and education and show the scatter plot

```
                        OLS Regression Results
==============================================================================
Dep. Variable:                  lwage   R-squared:                       0.116
Model:                            OLS   Adj. R-squared:                  0.115
Method:                 Least Squares   F-statistic:                     89.32
Date:                Tue, 01 Mar 2016   Prob (F-statistic):           5.34e-20
Time:                        20:06:11   Log-Likelihood:                -598.17
No. Observations:                 680   AIC:                             1200.
Df Residuals:                     678   BIC:                             1209.
Df Model:                           1
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      1.0077      0.153      6.573      0.000         0.707     1.309
educ           0.1022      0.011      9.451      0.000         0.081     0.123
==============================================================================
Omnibus:                       22.938   Durbin-Watson:                   1.469
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               32.785
Skew:                           0.306   Prob(JB):                     7.60e-08
Kurtosis:                       3.884   Cond. No.                         97.5
==============================================================================
```



Now regress log-wages on a constant, education, age, age-squared, and the gender and racial indicators. Briefly interpret the "economic meaning" of each slope coefficient. What do the coefficients on age and age-squared imply about the life-cycle profile of earnings? Would including just a linear term for age lead to a more appropriate regression model? Explain.

OLS Regression Results

==

Dep. Variable:	lwage	R-squared:	0.339
Model:	OLS	Adj. R-squared:	0.334
Method:	Least Squares	F-statistic:	69.06
Date:	Tue, 01 Mar 2016	Prob (F-statistic):	2.70e-58
Time:	20:06:12	Log-Likelihood:	-499.62
No. Observations:	680	AIC:	1011.
Df Residuals:	674	BIC:	1038.
Df Model:	5		

==

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	-1.0949	0.261	-4.191	0.000	-1.608	-0.582
educ	0.1100	0.010	11.508	0.000	0.091	0.129
age	0.1039	0.010	9.900	0.000	0.083	0.125
age2	-0.0011	0.000	-8.433	0.000	-0.001	-0.001
female	-0.3180	0.040	-7.944	0.000	-0.397	-0.239
white	-0.1001	0.072	-1.386	0.166	-0.242	0.042

==

Omnibus:	45.596	Durbin-Watson:	1.559
Prob(Omnibus):	0.000	Jarque-Bera (JB):	103.915
Skew:	0.376	Prob(JB):	2.72e-23
Kurtosis:	4.762	Cond. No.	2.50e+04

==

Answer: Each year of education raises wages by 11%. Each year of life does approximately the same until one reaches a certain point, after which it decreases or flattens out (negative coef on age2). Looking at a quick scatterplot of age and lwage, this looks appropriate. It also seems intuitively correct; earning power seems to peak around age 50 or so. We should keep age2 in the model. Being female decreases wages by about 32%. Being white seems to decrease wages by 10%, though the std err is large enough to make us suspicious of this value.

Now add age3 and age4 to the regression. Does this substantially improve the fit of the regression model?

```
                OLS Regression Results
==============================================================================
Dep. Variable:                  lwage   R-squared:                       0.341
Model:                            OLS   Adj. R-squared:                  0.335
Method:                 Least Squares   F-statistic:                     49.76
Date:                Tue, 01 Mar 2016   Prob (F-statistic):           4.96e-57
Time:                        20:06:12   Log-Likelihood:                -498.25
No. Observations:                 680   AIC:                             1013.
Df Residuals:                     672   BIC:                             1049.
Df Model:                           7
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     -3.9369      1.832     -2.149      0.032        -7.533    -0.340
educ           0.1089      0.010     11.285      0.000         0.090     0.128
age            0.4145      0.194      2.132      0.033         0.033     0.796
age2          -0.0130      0.007     -1.770      0.077        -0.027     0.001
age3           0.0002      0.000      1.640      0.102     -3.81e-05     0.000
age4       -1.118e-06    6.8e-07     -1.643      0.101     -2.45e-06  2.18e-07
female        -0.3175      0.040     -7.916      0.000        -0.396    -0.239
white         -0.1103      0.073     -1.519      0.129        -0.253     0.032
==============================================================================
Omnibus:                       49.420   Durbin-Watson:                   1.561
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              114.791
Skew:                           0.403   Prob(JB):                     1.18e-25
Kurtosis:                       4.844   Cond. No.                     5.16e+08
==============================================================================
```

Answer: R squared is essentially the same as before adding in extra age terms. Fit not improved.

b. Compare the estimated return to education to the one from the bivariate regression model. Are they different? What might this imply about how education is distributed across the twins population?

Answer: Estimated returns are very similar. Education is probablly distrubuted evenly across the population.

Now compare the mean characteristics of individuals with a college degree (educ=16) to individuals with just a high school degree (educ=12). Can you think of variables that we have not controlled for that may be related to both educational attainment and earnings? What does this imply about how we should interpret the least squares estimate of the relation between log-wages and education?

Means:

White

('12: ', 0.921875)

('16: ', 0.87681162)

female

('12: ', 0.62946427)

('16: ', 0.5869565)

selfemp

('12: ', 0.11532738)

('16: ', 0.13043478)

twoplus

('12: ', 0.23214285714285715)

('16: ', 0.18115942028985507)

dlwage

('12: ', -0.059820525)

('16: ', 0.015984911)

duncov

('12: ', 0.0066964286)

('16: ', -0.02657005)

dmaried

('12: ', -0.0044642859)

('16: ', 0.036231883)

Answer: At the p > .05 level: Highschool group is slightly whiter. College group is slightly more self-employed. Highschool group is slightly more female. Ability is correlated with both education and wages. Omitting this variable would give our estimates of the effects of educ on wages a positive bias--ie giving educ more credit than is due. In light of this, we should be skeptical of this estimate of the return on educ.

c. Now create dummy variables for each of the eleven levels of schooling (8-18). Regress both wages and log-wages on just the dummy variables. Is the effect of education on wages linear in education? How about its effect on log-wages? Focusing on log-wages, describe where the “nonlinearities” are, if any.

OLS Regression Results

==

Dep. Variable:	hrwage	R-squared:	0.599
Model:	OLS	Adj. R-squared:	0.593
Method:	Least Squares	F-statistic:	100.1
Date:	Tue, 01 Mar 2016	Prob (F-statistic):	7.31e-126
Time:	20:06:12	Log-Likelihood:	-2671.6
No. Observations:	680	AIC:	5363.
Df Residuals:	670	BIC:	5408.
Df Model:	10		

==

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	-4.22e+13	8.6e+13	-0.490	0.624	-2.11e+14	1.27e+14
educ8	4.22e+13	8.6e+13	0.490	0.624	-1.27e+14	2.11e+14
educ9	1.295e+10	2.64e+10	0.490	0.624	-3.89e+10	6.48e+10
educ10	-1.295e+10	2.64e+10	-0.490	0.624	-6.48e+10	3.89e+10
educ11	4.5709	6.694	0.683	0.495	-8.574	17.715
educ12	-0.6498	4.457	-0.146	0.884	-9.401	8.102
educ13	-0.0585	1.571	-0.037	0.970	-3.143	3.026
educ14	1.7485	1.875	0.933	0.351	-1.933	5.430
educ15	1.2384	2.316	0.535	0.593	-3.310	5.787
educ16	3.7897	2.182	1.737	0.083	-0.494	8.073
educ17	5.1742	3.728	1.388	0.166	-2.146	12.494
educ18	0.8079	3.909	0.207	0.836	-6.868	8.484

==

Omnibus:	643.125	Durbin-Watson:	1.510
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23990.157
Skew:	4.250	Prob(JB):	0.00

```
                         OLS Regression Results
==============================================================================
Dep. Variable:                  lwage   R-squared:                       0.947
Model:                            OLS   Adj. R-squared:                  0.946
Method:                 Least Squares   F-statistic:                     1197.
Date:                Tue, 01 Mar 2016   Prob (F-statistic):               0.00
Time:                        20:06:12   Log-Likelihood:                -594.39
No. Observations:                 680   AIC:                             1209.
Df Residuals:                     670   BIC:                             1254.
Df Model:                          10
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept   -1.12e+12   4.06e+12     -0.276      0.782     -9.08e+12  6.84e+12
educ8        1.12e+12   4.06e+12      0.276      0.782     -6.84e+12  9.08e+12
educ9       3.438e+08   1.24e+09      0.276      0.782      -2.1e+09  2.79e+09
educ10     -3.438e+08   1.24e+09     -0.276      0.782     -2.79e+09   2.1e+09
educ11         0.3673      0.316      1.164      0.245        -0.252     0.987
educ12        -0.0809      0.210     -0.385      0.700        -0.493     0.332
educ13         0.0139      0.074      0.187      0.852        -0.132     0.159
educ14         0.0748      0.088      0.847      0.398        -0.099     0.248
educ15         0.0846      0.109      0.775      0.439        -0.130     0.299
educ16         0.1943      0.103      1.890      0.059        -0.008     0.396
educ17         0.2168      0.176      1.234      0.218        -0.128     0.562
educ18         0.0638      0.184      0.347      0.729        -0.298     0.426
==============================================================================
Omnibus:                       24.262   Durbin-Watson:                   1.478
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               38.988
Skew:                           0.281   Prob(JB):                     3.42e-09
Kurtosis:                       4.029   Cond. No.                          nan
==============================================================================
```

Answer: The effects of education are nonlinear in both the hrwage and lwage models. The 16th and 17th years of education (ie graduating from college) has a disproportionately positive effect on wages in both cases.

Now run the dummy variable regression for log-wages including age, age-squared, gender, and race as controls. Does allowing for nonlinearities in the return to education improve the fit of the regression model substantially?

OLS Regression Results

==

Dep. Variable:	lwage	R-squared:	0.960
Model:	OLS	Adj. R-squared:	0.959
Method:	Least Squares	F-statistic:	1149.
Date:	Tue, 01 Mar 2016	Prob (F-statistic):	0.00
Time:	20:06:12	Log-Likelihood:	-496.62
No. Observations:	680	AIC:	1021.
Df Residuals:	666	BIC:	1085.
Df Model:	14		

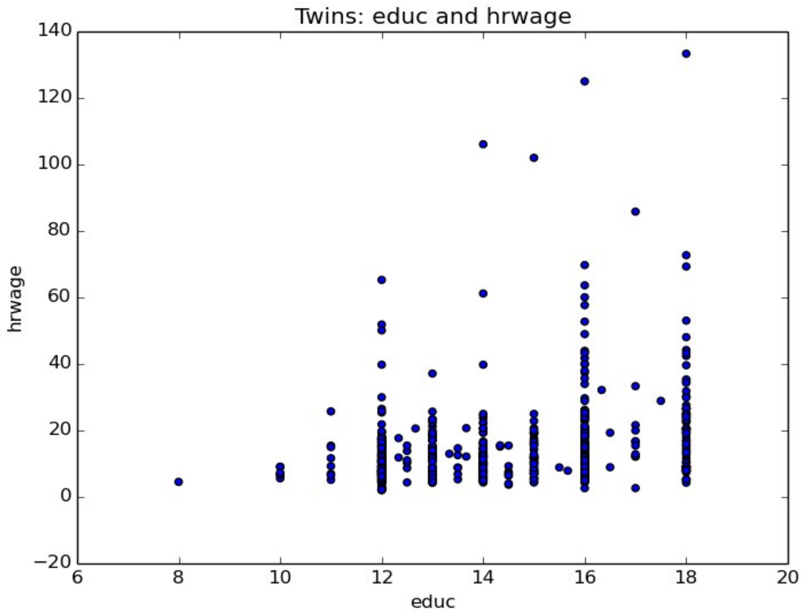
==

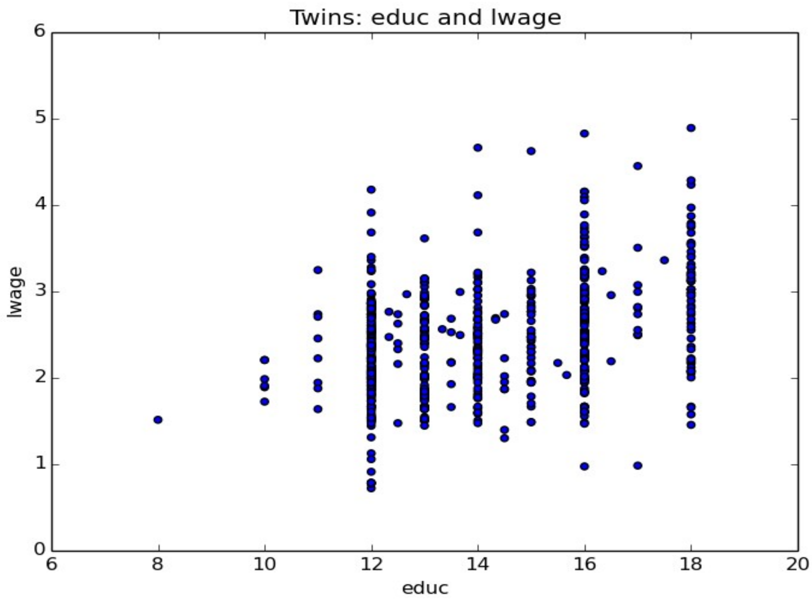
	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	-0.2471	0.280	-0.883	0.378	-0.797	0.303
educ8	-0.2471	0.280	-0.883	0.378	-0.797	0.303
educ9	0.2047	0.274	0.746	0.456	-0.334	0.743
educ10	0.2047	0.274	0.746	0.456	-0.334	0.743
educ11	0.4274	0.275	1.557	0.120	-0.112	0.966
educ12	-0.1360	0.183	-0.742	0.458	-0.496	0.224
educ13	0.1324	0.065	2.033	0.042	0.004	0.260
educ14	0.0801	0.077	1.042	0.298	-0.071	0.231
educ15	0.0727	0.095	0.765	0.444	-0.114	0.259
educ16	0.1818	0.089	2.034	0.042	0.006	0.357
educ17	0.2500	0.153	1.634	0.103	-0.050	0.550
educ18	-0.0832	0.161	-0.517	0.605	-0.399	0.232
age	0.1050	0.011	9.883	0.000	0.084	0.126
age2	-0.0011	0.000	-8.468	0.000	-0.001	-0.001
female	-0.3240	0.040	-8.020	0.000	-0.403	-0.245
white	-0.1023	0.073	-1.409	0.159	-0.245	0.040

==

Answer: The fit of the model has improved substantially (r-squared increased to .96)

d. Based on the scatter plot of hourly wages on the y-axis and education on the x-axis, is there any evidence on homoskedasticity/heteroskedacity in the wage regression model? What about with logwages on the y-axis?





Answer: There is evidence of possible heteroskedacity in both hrwage and lwage models with larger variation in wages as educ increases

e. Regress log-wages on education, age, age2, and the gender and racial indicators, using the “robust” option in STATA to calculate the Eicker-White consistent standard errors. Explain briefly how these estimates of the standard errors are corrected for heteroskedasticity. How do they compare to the “uncorrected” (conventional) least squares estimates of the standard errors. Is there any evidence of heteroskedasticity?

OLS Regression Results

==

Dep. Variable:	lwage	R-squared:	0.339
Model:	OLS	Adj. R-squared:	0.334
Method:	Least Squares	F-statistic:	69.06
Date:	Tue, 01 Mar 2016	Prob (F-statistic):	2.70e-58
Time:	20:06:12	Log-Likelihood:	-499.62
No. Observations:	680	AIC:	1011.
Df Residuals:	674	BIC:	1038.
Df Model:	5		

==

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	-1.0949	0.261	-4.191	0.000	-1.608	-0.582
educ	0.1100	0.010	11.508	0.000	0.091	0.129
age	0.1039	0.010	9.900	0.000	0.083	0.125
age2	-0.0011	0.000	-8.433	0.000	-0.001	-0.001
female	-0.3180	0.040	-7.944	0.000	-0.397	-0.239
white	-0.1001	0.072	-1.386	0.166	-0.242	0.042

==

Omnibus:	45.596	Durbin-Watson:	1.559
Prob(Omnibus):	0.000	Jarque-Bera (JB):	103.915
Skew:	0.376	Prob(JB):	2.72e-23
Kurtosis:	4.762	Cond. No.	2.50e+04

==

Warnings:

[1] The condition number is large, 2.5e+04. This might indicate that there are

strong multicollinearity or other numerical problems.

normal standard errors

Intercept 0.261239

educ 0.009558

age 0.010499

age2 0.000126

female 0.040031

white 0.072211

dtype: float64

White standard errors

Intercept 0.291092

educ 0.010431

age 0.011937

age2 0.000147

female 0.039746

white 0.067920

dtype: float64

p-value of the f-statistic of the hypothesis that the error variance does not depend on x:

0.000433064862979

Answer: White standard errors are slightly larger for intercept, educ, age, and age2. They are slightly smaller for female and white. Estimates of the standard errors are corrected for heteroskedasticity by allowing them to vary with x values. The White test shows evidence of heteroskedasticity. The p-value of the f-statistic of the hypothesis that the error variance does not depend on x is .0004.

f. Using the “predict” STATA command [predict (var. name), residual], save the residuals from both the wage and log-wage regressions. Now regress the squared values of the residuals from the two sets of regressions on education, age, age2, female, and white. From the R-squareds of these regressions, test for heteroskedasticity in the two sets of residuals. Does one set of residuals appear to be more heteroskedastic than the other?

OLS Regression Results

==

Dep. Variable:	res_lwage2	R-squared:	0.024
Model:	OLS	Adj. R-squared:	0.017
Method:	Least Squares	F-statistic:	3.362
Date:	Tue, 01 Mar 2016	Prob (F-statistic):	0.00523
Time:	20:06:12	Log-Likelihood:	-476.42
No. Observations:	680	AIC:	964.8
Df Residuals:	674	BIC:	992.0
Df Model:	5		

==

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	0.1980	0.252	0.784	0.433	-0.298	0.694
educ	0.0209	0.009	2.262	0.024	0.003	0.039
age	-0.0139	0.010	-1.372	0.170	-0.034	0.006
age2	0.0002	0.000	1.589	0.112	-4.56e-05	0.000
female	-0.0955	0.039	-2.468	0.014	-0.171	-0.020
white	0.0475	0.070	0.681	0.496	-0.090	0.185

==

Omnibus:	723.585	Durbin-Watson:	1.768
Prob(Omnibus):	0.000	Jarque-Bera (JB):	42665.706
Skew:	4.962	Prob(JB):	0.00
Kurtosis:	40.515	Cond. No.	2.50e+04

==

Warnings:

[1] The condition number is large, 2.5e+04. This might indicate that there are

strong multicollinearity or other numerical problems.

OLS Regression Results

==

Dep. Variable:	res_hrwage2	R-squared:	0.050
Model:	OLS	Adj. R-squared:	0.043
Method:	Least Squares	F-statistic:	7.032
Date:	Tue, 01 Mar 2016	Prob (F-statistic):	2.03e-06
Time:	20:06:12	Log-Likelihood:	-5422.4
No. Observations:	680	AIC:	1.086e+04
Df Residuals:	674	BIC:	1.088e+04
Df Model:	5		

==

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	-484.8487	363.967	-1.332	0.183	-1199.495	229.797
educ	47.1574	13.316	3.541	0.000	21.011	73.304
age	-8.4434	14.627	-0.577	0.564	-37.164	20.277
age2	0.1918	0.176	1.091	0.275	-0.153	0.537
female	-184.7258	55.773	-3.312	0.001	-294.236	-75.216
white	89.8164	100.606	0.893	0.372	-107.723	287.356

==

Omnibus:	1155.379	Durbin-Watson:	1.848
Prob(Omnibus):	0.000	Jarque-Bera (JB):	511482.287
Skew:	10.721	Prob(JB):	0.00
Kurtosis:	135.637	Cond. No.	2.50e+04

==

Warnings:

[1] The condition number is large, 2.5e+04. This might indicate that there are

strong multicollinearity or other numerical problems.

Answer: Residuals from hrwage model appear more heteroskedastic (r-squared of .05 vs r-squared of .025 for lwage model) ie the regressors from the hrwage model are better predictors of that model's residuals than the lwage model's regressors are of that model's residuals.

Now regress the squared residuals on education, education2, age, age2, female, white, and the interactions education*age, female*age, female*education, white*age, and white*education. Again, test for heteroskedasticty based on the R-squareds of the regressions.

```
                         OLS Regression Results
==============================================================================
Dep. Variable:           res_lwage2   R-squared:                       0.048
Model:                          OLS   Adj. R-squared:                  0.034
Method:               Least Squares   F-statistic:                     3.356
Date:              Tue, 01 Mar 2016   Prob (F-statistic):           0.000279
Time:                      20:06:13   Log-Likelihood:                -468.16
No. Observations:               680   AIC:                             958.3
Df Residuals:                   669   BIC:                             1008.
Df Model:                        10
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      0.1488      0.853      0.174      0.862        -1.526     1.824
educ           0.0017      0.051      0.033      0.974        -0.098     0.101
age           -0.0243      0.017     -1.398      0.163        -0.059     0.010
age2           0.0002      0.000      1.572      0.117     -4.87e-05     0.000
female         0.6279      0.314      2.001      0.046         0.012     1.244
white          0.4246      0.660      0.643      0.520        -0.872     1.721
educAge        0.0013      0.001      1.635      0.102        -0.000     0.003
femAge         0.0021      0.003      0.611      0.542        -0.005     0.009
femEduc       -0.0567      0.019     -3.018      0.003        -0.094    -0.020
whiteAge      -0.0107      0.006     -1.748      0.081        -0.023     0.001
whiteEduc      0.0016      0.037      0.045      0.965        -0.071     0.074
==============================================================================
Omnibus:                      725.359   Durbin-Watson:                   1.785
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            44463.168
Skew:                           4.965   Prob(JB):                         0.00
Kurtosis:                      41.350   Cond. No.                     1.05e+05
==============================================================================
```

OLS Regression Results

==

Dep. Variable:	res_hrwage2	R-squared:	0.092
Model:	OLS	Adj. R-squared:	0.078
Method:	Least Squares	F-statistic:	6.760
Date:	Tue, 01 Mar 2016	Prob (F-statistic):	4.47e-10
Time:	20:06:13	Log-Likelihood:	-5407.0
No. Observations:	680	AIC:	1.084e+04
Df Residuals:	669	BIC:	1.089e+04
Df Model:	10		

==

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	1928.5595	1216.899	1.585	0.113	-460.841	4317.960
educ	-119.6241	72.506	-1.650	0.099	-261.990	22.742
age	-70.5448	24.831	-2.841	0.005	-119.300	-21.789
age2	0.1864	0.177	1.052	0.293	-0.161	0.534
female	1458.7901	447.572	3.259	0.001	579.975	2337.605
white	-851.1648	942.010	-0.904	0.367	-2700.817	998.487
educAge	4.4441	1.165	3.815	0.000	2.157	6.731
femAge	-9.9143	4.847	-2.045	0.041	-19.432	-0.396
femEduc	-88.6309	26.799	-3.307	0.001	-141.250	-36.011
whiteAge	5.8957	8.710	0.677	0.499	-11.207	22.999
whiteEduc	45.0022	52.880	0.851	0.395	-58.828	148.832

==

Omnibus:	1122.513	Durbin-Watson:	1.887
Prob(Omnibus):	0.000	Jarque-Bera (JB):	440023.287
Skew:	10.141	Prob(JB):	0.00
Kurtosis:	125.959	Cond. No.	1.05e+05

==

Answer: Including these interaction variables helps explain more of the variation in the residuals of both models, ie the residuals appear more heteroskedastic when tested with these variables. Adjusted r-squares nearly doubled for both models.

g. Explain how the assumption that the residuals from the log-wage regression are “pairwise” uncorrelated may be violated when using the twins data. Use the following STATA commands to create a variable that separately identifies each twin pair in the data set (Note: the data must be in its original order for this to work):

Answer: Twins are probably correlated with one another across a number of characteristics, e.g. ability, family upbringing, etc. If one twin has a positive residual, we could guess that the other might, as well.

Run the regression of log-wages on education, age, age2, female, and white using the “cluster” STATA option to correct the estimated standard errors for correlation in the residuals between twins. Explain why the standard errors on the estimated return to education are higher (and t-ratio lower) than when clustering is not corrected for.

OLS Regression Results

==

Dep. Variable:	lwage	R-squared:	0.339
Model:	OLS	Adj. R-squared:	0.334
Method:	Least Squares	F-statistic:	69.06
Date:	Tue, 01 Mar 2016	Prob (F-statistic):	2.70e-58
Time:	20:06:13	Log-Likelihood:	-499.62
No. Observations:	680	AIC:	1011.
Df Residuals:	674	BIC:	1038.
Df Model:	5		

==

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	-1.0949	0.261	-4.191	0.000	-1.608	-0.582
educ	0.1100	0.010	11.508	0.000	0.091	0.129
age	0.1039	0.010	9.900	0.000	0.083	0.125
age2	-0.0011	0.000	-8.433	0.000	-0.001	-0.001
female	-0.3180	0.040	-7.944	0.000	-0.397	-0.239
white	-0.1001	0.072	-1.386	0.166	-0.242	0.042

==

Omnibus:	45.596	Durbin-Watson:	1.559
Prob(Omnibus):	0.000	Jarque-Bera (JB):	103.915
Skew:	0.376	Prob(JB):	2.72e-23
Kurtosis:	4.762	Cond. No.	2.50e+04

Answer: SE is significantly bigger on educ because the intracluster correlations are positive. Without accounting for clustering, we effectively overestimate our sample size. Note: cluster option in my stats package isn't working. These appear to be normal errors, not clustered, but I know what the values should look like!

h. Now run the regression of the average of the log-wages of each twin pair on each twin pair's average education (i.e., you now have 340 twin pair observations based on twin averages). Does this correct the "clustering" problem in the residuals? Explain briefly.

```
  OLS Regression Results
==============================================================================
Dep. Variable:                  lwage   R-squared:                       0.137
Model:                            OLS   Adj. R-squared:                  0.135
Method:                 Least Squares   F-statistic:                     53.81
Date:                Tue, 01 Mar 2016   Prob (F-statistic):           1.65e-12
Time:                        20:06:13   Log-Likelihood:                -262.90
No. Observations:                 340   AIC:                             529.8
Df Residuals:                     338   BIC:                             537.5
Df Model:                           1
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      0.9258      0.209      4.440      0.000         0.516     1.336
educ           0.1080      0.015      7.335      0.000         0.079     0.137
==============================================================================
Omnibus:                        8.381   Durbin-Watson:                   2.119
Prob(Omnibus):                  0.015   Jarque-Bera (JB):                9.868
Skew:                           0.253   Prob(JB):                      0.00720
Kurtosis:                       3.664   Cond. No.                         104.
==============================================================================
```

Answer: Yes, this effectively corrects for the clustering problem. Collapsing to group means is an extreme version of using clustered standard errors--we've reduced our number of observations down to the number of groups.

i. Suppose that a twin's self-report of education is an imperfect measure of the twin's actual educational attainment due to misreporting. In addition, suppose that this measurement error is "classical" in the sense that it is independently and identically distributed. What is the formula for the bias in the estimated return to education from the regression of log-wages on educ, age, age2, white, female in terms of the "noise-to-total variance ratio"?

The expected value of our estimator will be biased downwards (1-(d/(1-r))) where d is the noise to total variance ratio (var of measurement error / var of observed lwages) and r is the r-squared value from our regression of schooling on the X's.

j. Now run the following STATA command [ivreg lwage age age2 female white (educ = educt_t), cluster(id)]. This performs two-stage least squares estimation of the return to education using the other twin's report of the individual's education level as an instrument for the individual's self-reported education (for each individual, both the individual and his twin were asked about the individual's education level). Explain why the estimated return to education from this procedure is greater than the estimated return from standard OLS. Calculate the reliability ratio of the education data under the assumption that the measurement errors are classically distributed.

```
                    OLS Regression Results
==============================================================================
Dep. Variable:                  lwage   R-squared:                       0.332
Model:                            OLS   Adj. R-squared:                  0.327
Method:                 Least Squares   F-statistic:                     66.91
Date:                Tue, 01 Mar 2016   Prob (F-statistic):           9.32e-57
Time:                        20:06:13   Log-Likelihood:                -503.22
No. Observations:                 680   AIC:                             1018.
Df Residuals:                     674   BIC:                             1046.
Df Model:                           5
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     -1.2070      0.271     -4.455      0.000        -1.739    -0.675
educ_pred      0.1156      0.010     11.132      0.000         0.095     0.136
age            0.1059      0.011     10.031      0.000         0.085     0.127
age2          -0.0011      0.000     -8.591      0.000        -0.001    -0.001
female        -0.3247      0.040     -8.078      0.000        -0.404    -0.246
white         -0.0964      0.073     -1.327      0.185        -0.239     0.046
==============================================================================
Omnibus:                       42.613   Durbin-Watson:                   1.539
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               88.277
Skew:                           0.381   Prob(JB):                     6.77e-20
Kurtosis:                       4.592   Cond. No.                     2.58e+04
==============================================================================
```

Return on education is slightly higher because by instrumenting with the twin's estimate we've mitigated the attenuation bias. This assumes a twin's and an individual's measurement errors aren't correlated.

OLS Regression Results

==

Dep. Variable:	lwage	R-squared:	0.332
Model:	OLS	Adj. R-squared:	0.327
Method:	Least Squares	F-statistic:	66.91
Date:	Tue, 01 Mar 2016	Prob (F-statistic):	9.32e-57
Time:	20:06:13	Log-Likelihood:	-503.22
No. Observations:	680	AIC:	1018.
Df Residuals:	674	BIC:	1046.
Df Model:	5		

==

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	-1.2070	0.271	-4.455	0.000	-1.739	-0.675
educ_pred	0.1156	0.010	11.132	0.000	0.095	0.136
age	0.1059	0.011	10.031	0.000	0.085	0.127
age2	-0.0011	0.000	-8.591	0.000	-0.001	-0.001
female	-0.3247	0.040	-8.078	0.000	-0.404	-0.246
white	-0.0964	0.073	-1.327	0.185	-0.239	0.046

==

Omnibus:	42.613	Durbin-Watson:	1.539
Prob(Omnibus):	0.000	Jarque-Bera (JB):	88.277
Skew:	0.381	Prob(JB):	6.77e-20
Kurtosis:	4.592	Cond. No.	2.58e+04

==

k. Suppose there is an unmeasured factor that is associated with both an individual's educational attainment and an individual's earnings (e.g., innate ability, family background, school quality). Explain how this could lead to "omitted variables" bias in the least squares estimate of the return to education. Suppose that the omitted variable is Ai. Write out the "omitted variables bias" in terms of the linear relationships between education and A and log-wages and A.

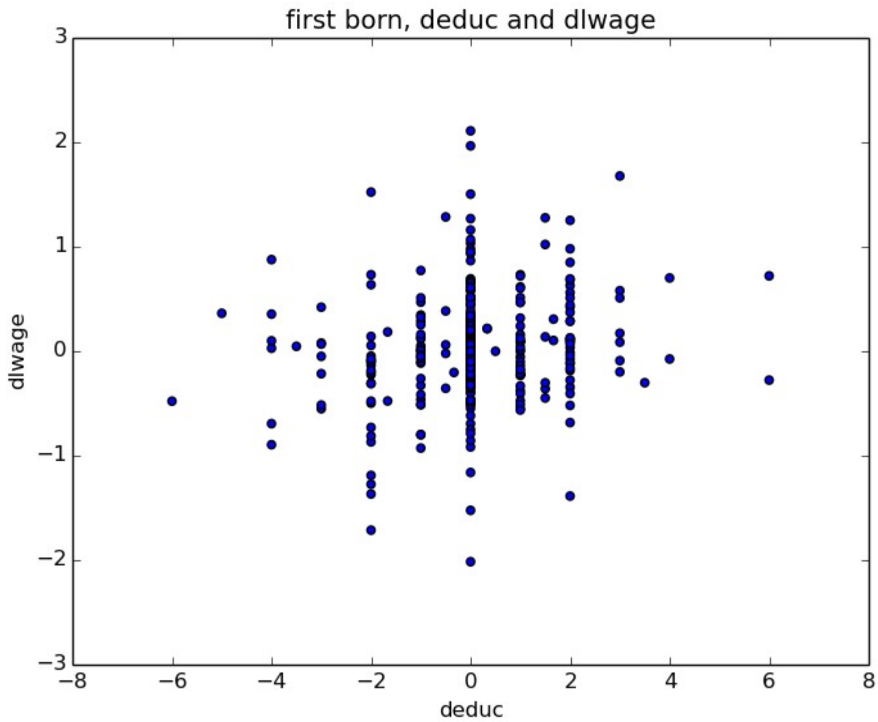
Answer: If the omitted variable is positively correlated with educ and with wages, as we expect ability would be, then omitting it will bias our estimate of the returns to education upwards, ie educ will steal ability's thunder. The expected value of our biased estimator will be equal to the unbiased estimator + t(cov(educ,Ai) / var(educ)) where 't' is the effect of Ai on lwage, controlling for educ. If t and cov(educ, Ai) are greater than zero, our estimator will be biased upwards, ie it will overstate the actual effects of educ on lwage.

l. Now suppose that all omitted factors are held constant when comparing identical twins. Run the regression of the difference in log-wages between twins on the difference in educational attainment using the STATA command [reg dlwage deduc if first==1, noconstant robust]. How does this estimate of the return to education compare to the one based on the regression of lwage on educ, age, age2, female, white? Explain what this might imply about the omitted variables bias.

```
                    OLS Regression Results
==============================================================================
Dep. Variable:              dlwage   R-squared:                       0.031
Model:                         OLS   Adj. R-squared:                  0.028
Method:              Least Squares   F-statistic:                     10.64
Date:             Tue, 01 Mar 2016   Prob (F-statistic):            0.00122
Time:                     20:06:13   Log-Likelihood:                -250.73
No. Observations:              340   AIC:                             505.5
Df Residuals:                  338   BIC:                             513.1
Df Model:                        1
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      0.0296      0.028      1.074      0.284        -0.025     0.084
deduc          0.0610      0.019      3.262      0.001         0.024     0.098
==============================================================================
Omnibus:                    28.586   Durbin-Watson:                   2.269
Prob(Omnibus):               0.000   Jarque-Bera (JB):               96.191
Skew:                        0.254   Prob(JB):                     1.30e-21
Kurtosis:                    5.556   Cond. No.                         1.47
==============================================================================
```

Answer: The fact that our estimated returns to educ halved when holding (ostensibly) all omitted variables constant shows that our previous results were probably biased upwards as predicted.

m. Graph a scatter plot with the difference in twins' log-wages (dlwage) on the y-axis and the difference in twins' self-reported education (deduc) on the x-axis only using the first-born twin's observations (first=1). Where are most of the observations clustered with respect to the x-axis of deduc? What could this imply about the importance of measurement error in this variable? How might this be another explanation for the result you found in part (l) that the estimated return to education is lower when running the "first-differences" regression? Explain how first-differencing the data may exacerbate the measurement error problem.



Answer: Most observations are clustered close to zero This implies that our model may be very sensitive to measurement error. By first-differencing the data, we're giving up signal but keeping all the noise, so our attenuation bias will be even worse than before. This could exacerbate the measurement error problem and explain why our estimated returns to educ are lower when running the first-differenced regression.

n. Now run the STATA command [ivreg dlwage (deduc = deduct) if first==1, noconstant robust]. This two-stage least squares regression uses deduct as an instrument for deduc. Explain why the estimated return to education is now larger than the one in part (l). How does the unbiasedness of this estimate depend on the classical measurement error assumption? Will it be unbiased if the measurement errors between an individual's self-report of education and his twin's report of the individual's education are correlated? Describe a solution to this problem.

OLS Regression Results

==

Dep. Variable:	dlwage	R-squared:	0.039
Model:	OLS	Adj. R-squared:	0.037
Method:	Least Squares	F-statistic:	27.17
Date:	Tue, 01 Mar 2016	Prob (F-statistic):	2.47e-07
Time:	20:06:13	Log-Likelihood:	-499.99
No. Observations:	680	AIC:	1004.
Df Residuals:	678	BIC:	1013.
Df Model:	1		

==

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	4.337e-19	0.019	2.24e-17	1.000	-0.038	0.038
deduc_pred	0.1075	0.021	5.213	0.000	0.067	0.148

n. Much as it did above, using an instrument helps us get rid of the attenuation bias we had from measurement error in the self-reported data, increasing our estimate substantially. As above, this assumes a twin's and an individual's measurement errors aren't correlated--if they are, this estimate is biased.

o. Suppose that the classical measurement error assumption holds, what might one conclude about the size of the omitted variables bias in the "conventional" OLS estimate of the returns to education (i.e., the estimate from regressing lwage on educ, age, age2, female, white)? Do you think that comparing twin pair differences across families reduces the omitted variables problem? Explain.

Answer: Our estimate of the return on education after dealing with attenuation bias and OVB with regards to ability isn't much smaller than our original, plain vanilla estimate. Looks like the OVB wasn't that serious to begin with. Alternatively, comparing twin-pair differences might just not being getting rid of OVB. There may be significant variation in twins among things like ability, motivation, etc. Don't they say the first twin is usually more successful or something?

2. Estimating the effect of computers using the German data

a. Regress log-wages on a constant, education, experience, experience-squared, the gender and marital status indicators, and the computer indicator, while adjusting for heteroskedasticity. Briefly interpret the “economic meaning” of each slope coefficient.

OLS Regression Results

==

Dep. Variable:	lnw	R-squared:	0.333
Model:	OLS	Adj. R-squared:	0.333
Method:	Least Squares	F-statistic:	1670.
Date:	Tue, 01 Mar 2016	Prob (F-statistic):	0.00
Time:	20:06:13	Log-Likelihood:	-8857.7
No. Observations:	20042	AIC:	1.773e+04
Df Residuals:	20035	BIC:	1.778e+04
Df Model:	6		

==

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	1.7014	0.019	90.298	0.000	1.664	1.738
ed	0.0703	0.001	59.508	0.000	0.068	0.073
exp	0.0298	0.001	28.463	0.000	0.028	0.032
exp2	-0.0457	0.002	-22.179	0.000	-0.050	-0.042
female	-0.2153	0.006	-38.425	0.000	-0.226	-0.204
mar	0.0353	0.006	5.592	0.000	0.023	0.048
computer	0.1722	0.006	29.020	0.000	0.161	0.184

==

Omnibus:	3664.899	Durbin-Watson:	1.705
Prob(Omnibus):	0.000	Jarque-Bera (JB):	21550.860
Skew:	-0.752	Prob(JB):	0.00
Kurtosis:	7.852	Cond. No.	211.

==

Answer: Each year of education increases wages by 7%. Each year of experience increases wages by 3% until a certain pt, after which wages flatten or fall. Being female reduces wages by 22%. Being married increases wages by 4%. Using a computer at work increases wages by 17%.

b. Now add the indicators for pencil, telephone, calculator, and hammer use to the regression you ran in part (a). Compare the estimated return to education and computer use to the ones from the part (a) regression model. Are they different?

OLS Regression Results

==

Dep. Variable:	lnw	R-squared:	0.343
Model:	OLS	Adj. R-squared:	0.343
Method:	Least Squares	F-statistic:	1047.
Date:	Tue, 01 Mar 2016	Prob (F-statistic):	0.00
Time:	20:06:13	Log-Likelihood:	-8706.8
No. Observations:	20042	AIC:	1.744e+04
Df Residuals:	20031	BIC:	1.752e+04
Df Model:	10		

==

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	1.7513	0.020	88.344	0.000	1.712	1.790
ed	0.0647	0.001	52.917	0.000	0.062	0.067
exp	0.0288	0.001	27.650	0.000	0.027	0.031
exp2	-0.0438	0.002	-21.401	0.000	-0.048	-0.040
female	-0.2294	0.006	-39.077	0.000	-0.241	-0.218
mar	0.0334	0.006	5.329	0.000	0.021	0.046
computer	0.1197	0.007	18.001	0.000	0.107	0.133
hammer	-0.0354	0.006	-5.505	0.000	-0.048	-0.023
telefon	0.0418	0.008	5.172	0.000	0.026	0.058
calc	0.0461	0.007	6.603	0.000	0.032	0.060
pencil	0.0311	0.008	3.861	0.000	0.015	0.047

==

Omnibus:	3809.746	Durbin-Watson:	1.709
Prob(Omnibus):	0.000	Jarque-Bera (JB):	22828.704
Skew:	-0.781	Prob(JB):	0.00
Kurtosis:	7.990	Cond. No.	225.

==

Answer: Returns to ed reduced to 6%, returns to computer reduced to 12%. Results significantly different.

c. Now run a regression that also controls for the individual's occupation category as "fixed effects" – e.g., areg y x, absorb(occ) robust (data must be sorted by occ). Interpret the implications of your findings for the role of potential omitted variables bias in the OLS estimate of the effect of computer use on log-wages (see DiNardo and Pischke for their interpretation).

ed	0.0404	0.002	23.629	0.000	0.037	0.044
exp	0.0261	0.001	25.240	0.000	0.024	0.028
exp2	-0.0384	0.002	-18.916	0.000	-0.042	-0.034
female	-0.1604	0.008	-21.141	0.000	-0.175	-0.146
mar	0.0341	0.006	5.574	0.000	0.022	0.046
computer	0.0682	0.007	9.354	0.000	0.054	0.082
hammer	-0.0206	0.008	-2.649	0.008	-0.036	-0.005
telefon	0.0484	0.008	5.793	0.000	0.032	0.065
calc	0.0223	0.007	3.147	0.002	0.008	0.036
pencil	0.0074	0.008	0.914	0.361	-0.009	0.023

Answer: The effect of using a computer at work has gone down to 7%. As DiNardo and Pischke conclude, these results seem to suggest that computer users have unobserved skills which might have little to do with computers, but which are rewarded on the job market, or that computers were introduced first in higher-paying jobs. Note, these results make it clear that this is not an appropriate way to measure the returns to a given technology—are we to assume that using a hammer yields negative returns?