

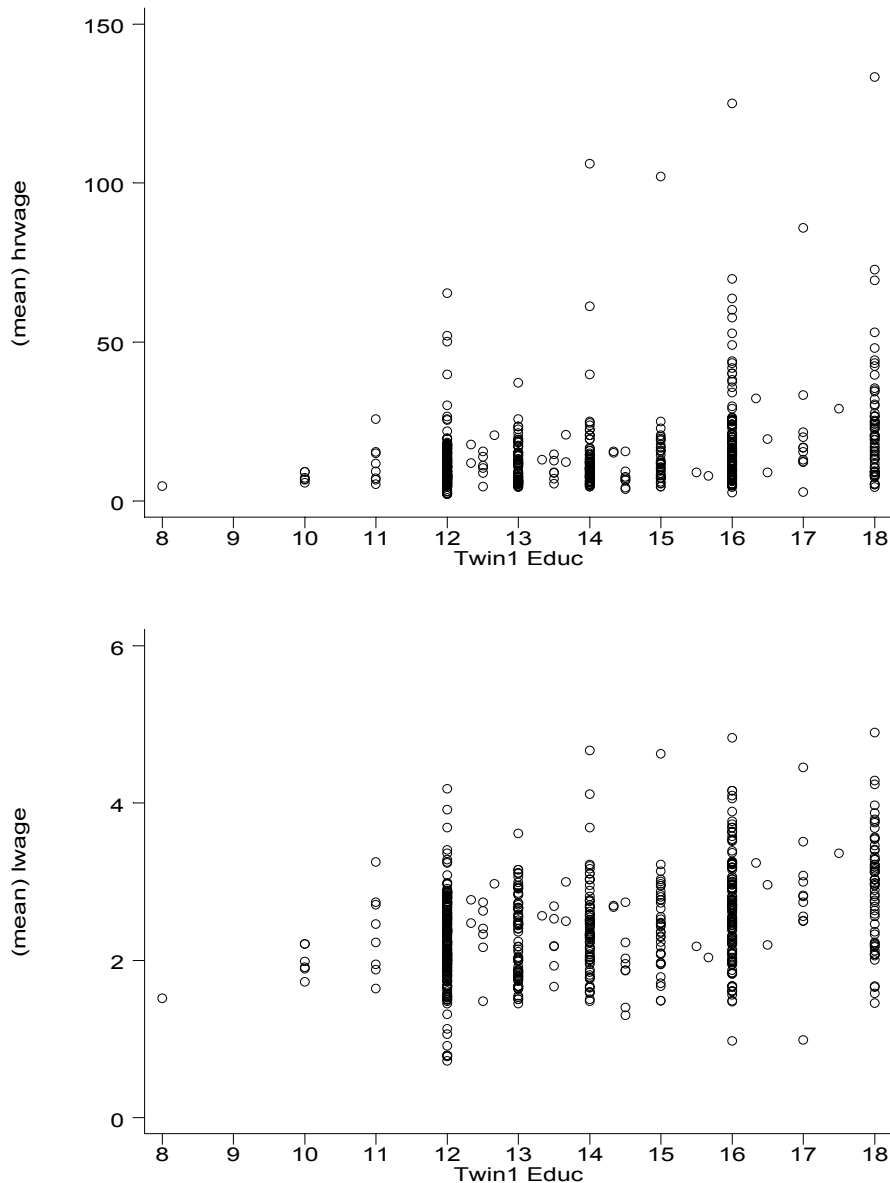
ECON 142

**SKETCH OF SOLUTIONS FOR APPLIED EXERCISE #2**

Question 1:

a.

Below are the scatter plots of hourly wages and log-wages on the y-axes and education on the x-axis, respectively. In both figures, there appears to be evidence of heteroskedasticity in the variance of the outcome with respect to education. Notice how the dispersion of points at a given level of education varies across the education levels. In addition, it appears that there may be more heteroskedasticity in the level of wages than in log-wages.



b.

If there is heteroskedasticity in the residuals of the log-wage regression, the OLS estimator of the returns to education is still unbiased and consistent but will no longer be the efficient estimator (not BLUE). In addition, the “conventional” estimator of the variance of the estimated return to education will be biased and inconsistent.

c.

Below are the STATA regression results uncorrected and corrected for heteroskedasticity, respectively.

```
. reg lwage educ age age2 female white
Source |           SS          df          MS              Number of obs =      680
-----+-----+-----+-----+-----+-----
Model   |  88.6616896         5   17.7323379          F( 5, 674) =      69.06
Residual| 173.065976        674   .256774445          Prob > F      = 0.0000
-----+-----+-----+-----+-----
Total   | 261.727666        679   .38546048          R-squared     = 0.3388
                                           Adj R-squared = 0.3339
                                           Root MSE     = .50673
```

```
-----+-----+-----+-----+-----+-----
lwage |           Coef.    Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----+-----+-----+-----+-----
educ  |   .1099923   .0095578    11.508  0.000   .0912256   .128759
age   |   .1039422   .0104989     9.900  0.000   .0833277   .1245567
age2  |  -.0010635   .0001261    -8.433  0.000  -.0013111  -.0008159
female|  -.317994   .0400314    -7.944  0.000  -.3965951  -.2393928
white |  -.1000952   .0722105    -1.386  0.166  -.2418798   .0416894
_cons |  -1.094912   .2612391    -4.191  0.000  -1.607853  -.581972
-----+-----+-----+-----+-----+-----
```

```
. reg lwage educ age age2 female white, robust
Regression with robust standard errors              Number of obs =      680
                                                    F( 5, 674) =      59.62
                                                    Prob > F      = 0.0000
                                                    R-squared     = 0.3388
                                                    Root MSE     = .50673
```

```
-----+-----+-----+-----+-----+-----
lwage |           Coef.    Robust Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----+-----+-----+-----+-----
educ  |   .1099923   .0104776    10.498  0.000   .0894197   .1305649
age   |   .1039422   .0119901     8.669  0.000   .0803997   .1274847
age2  |  -.0010635   .0001472    -7.225  0.000  -.0013525  -.0007745
female|  -.317994   .0399229    -7.965  0.000  -.3963823  -.2396056
white |  -.1000952   .0682216    -1.467  0.143  -.2340475   .0338571
_cons |  -1.094912   .2923846    -3.745  0.000  -1.669007  -.520818
-----+-----+-----+-----+-----+-----
```

When using the “robust” subcommand, the estimates of the standard errors are corrected for heteroskedasticity by allowing the variance of the residuals to vary across individuals in an unspecified way when the variance-covariance matrix of the slope coefficients is estimated. In effect, the squares of the estimated residuals for each individual are used in the calculation of the standard errors (see class notes for the Eicker-White formula). The “corrected” standard errors on education and age are about 10-15% greater than the “uncorrected” standard errors, while the standard errors on the gender and racial indicators are not very different. Consequently, based on this comparison, there is only slight evidence of heteroskedasticity.

d.

The reason why the R-squared and estimated coefficients from the regression of the estimated residuals on the regressors is virtually zero is that these “first-order” conditions (a.k.a., normal equations or moment/orthogonality conditions) were how the original least-squares estimates were derived in the first place. In other words, least squares estimation “calculates” the slope coefficients by imposing the

restriction implied by the model that the covariance between the residuals and regressors is zero (look at your notes and the last problem set for the equations).

e.

The following are the STATA regression results from the regressions of the squared residuals from the level of wages and log-wages regressions on the covariates, respectively:

```
. reg err2 educ age age2 female white
```

Source	SS	df	MS			
Model	17524249.8	5	3504849.96	Number of obs =	680	
Residual	335938741	674	498425.432	F( 5, 674) =	7.03	
Total	353462991	679	520564.051	Prob > F =	0.0000	
				R-squared =	0.0496	
				Adj R-squared =	0.0425	
				Root MSE =	705.99	

err2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	47.15739	13.31631	3.541	0.000	21.01095	73.30382
age	-8.443381	14.62742	-0.577	0.564	-37.16417	20.27741
age2	.1917732	.1757084	1.091	0.275	-.1532285	.5367749
female	-184.7258	55.77308	-3.312	0.001	-294.2357	-75.21595
white	89.81643	100.6062	0.893	0.372	-107.7228	287.3556
_cons	-484.8487	363.9673	-1.332	0.183	-1199.495	229.7974

```
. reg lerr2 educ age age2 female white
```

Source	SS	df	MS			
Model	4.03171842	5	.806343685	Number of obs =	680	
Residual	161.65299	674	.239841232	F( 5, 674) =	3.36	
Total	165.684709	679	.244012826	Prob > F =	0.0052	
				R-squared =	0.0243	
				Adj R-squared =	0.0171	
				Root MSE =	.48974	

lerr2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.0208984	.0092373	2.262	0.024	.002761	.0390358
age	-.0139254	.0101468	-1.372	0.170	-.0338486	.0059978
age2	.0001937	.0001219	1.589	0.112	-.0000456	.0004331
female	-.0954655	.0386889	-2.468	0.014	-.1714308	-.0195003
white	.0475033	.0697889	0.681	0.496	-.0895265	.1845331
_cons	.1979907	.2524784	0.784	0.433	-.297748	.6937294

The number of observations times the R-squareds from these regressions ( $N \cdot R^2$ ) is a test-statistic for heteroskedasticity that has an asymptotic chi-squared distribution under the null hypothesis of no heteroskedasticity, with the number of degrees of freedom equal to the number of regressors in the auxiliary squared residuals regression. In the above cases,  $680 \cdot R^2 \rightarrow \chi^2(5)$ . This test-statistic is 33.7 and 16.5 for the wage and log-wage regressions, respectively. The 5% critical value for a  $\chi^2(5)$  is 11.07. Consequently, we would reject the null hypothesis of homoskedasticity for both models at conventional levels of significance. It does appear, however, that the residuals from the wage regression exhibit more heteroskedasticity than the residuals from the log-wage regression.

When the squared residuals are regressed on the original regressors, their squares, and their interactions, then one is performing the White test for heteroskedasticity. The R-squareds from the regressions of the squared wage regression residuals and the squared log-wage regression residuals on all of these variables are 0.092 and 0.050, respectively. The resulting test-statistics are 62.6 and 34.0 and have asymptotic  $\chi^2(11)$  distributions under the null. Since the 5% critical value for a  $\chi^2(11)$  is 19.7, we again reject the null of no heteroskedasticity in both cases.

f.

Since identical twins are not just from the same family, but also from the same egg, there is a very good chance that there will be unobservable characteristics that are common among twins (e.g., personality, family background, schooling environment, etc.). Consequently, treating twins as being independent of each other is probably inappropriate, leading to a violation of the “pairwise” uncorrelated assumption on their residuals. Although the OLS estimator will be unbiased and consistent, it will be inefficient and the “conventional” estimates of the variances of the slope coefficients will be inconsistent. In many cases, this bias in the estimated standard errors can be quite substantial. Since twins are likely to be positively correlated with each other, the conventional standard error estimates will be biased down, overstating the precision of the coefficient estimates. Note that we may also worry that these twin “random effects” (unobservables) may also be correlated with both education and wages (e.g., differences across families in “ability” or motivation). In this case, there may be omitted variables bias in the OLS estimate of the returns to education. This is the subject of parts c and d of question 2.

g.

The STATA results are:

```
. reg lwage educ age age2 female white, cluster(id)
Regression with robust standard errors

Number of obs =      680
F( 5, 339) =    41.03
Prob > F      =    0.0000
R-squared     =    0.3388
Root MSE     =    .50673

Number of clusters (id) = 340
```

lwage	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.1099923	.0126008	8.729	0.000	.0852067	.1347779
age	.1039422	.0145249	7.156	0.000	.075372	.1325124
age2	-.0010635	.0001796	-5.922	0.000	-.0014167	-.0007103
female	-.317994	.0492271	-6.460	0.000	-.414823	-.2211649
white	-.1000952	.0686473	-1.458	0.146	-.2351234	.034933
_cons	-1.094912	.3385178	-3.234	0.001	-1.760772	-.4290524

So the standard error on the estimated return to education increases by over 30% when compared to the “conventional” standard error that is not corrected for clustering (the standard error on the age coefficient also increases substantially). When clustering is corrected for, we are now appropriately recognizing that the observations should not be treated as being independent of each other. When clustering is not corrected for, using the conventional OLS estimates of the standard errors inappropriately treats the data as being independent and overstates the amount of information contained in the data.

Question 2:

a.

If the self-reports of education are measured with error, and the measurement error is classical, then the estimated return to education from the multivariate log-wage regression will be biased down. This is also known as attenuation bias. The size of this bias depends on the relative amount of the variation in the self-report of schooling that is attributable to the measurement error. This is also known as the “noise-to-total variance” ratio. In the bivariate regression model with no additional regressors, the bias has the form:

$$\text{Bias}(\hat{\beta}_{OLS}) = -\frac{\beta\sigma_u^2}{\sigma_s^2} = -\frac{\beta\sigma_u^2}{\sigma_u^2 + \sigma_{S^*}^2} = -\lambda\beta,$$

where  $\sigma_u^2$  is the measurement error variance,  $\sigma_s^2$  is the total variance in the self-report of schooling,  $\sigma_{s^*}^2$  is the variance in true schooling (a.k.a, the signal), and  $\lambda$  is the noise-to-total variance ratio. If several regressors,  $X$ , are included in the regression as controls, then the bias has the form:

$$\text{Bias}(\hat{\beta}_{OLS}) = -\frac{\lambda\beta}{1-R_{S,X}^2}, \text{ where } R_{S,X}^2 \text{ is the R-squared from the regression of schooling on the } X\text{'s.}$$

As  $R_{S,X}^2$  goes up, the size of the attenuation bias increases for a fixed noise-signal ratio. Intuitively, if the measurement error is classical, and therefore independent of the additional regressors  $X$ , then including  $X$  in the regression will “soak up” some of the signal in schooling since it is related with true schooling. However, it will soak up none of the noise variance in self-reported schooling since it is independent of it. Consequently, while including additional regressors may alleviate the potential “omitted variables” problem, it could exacerbate attenuation bias arising from measurement error.

b.

The following are the results from using the other twin’s report of an individual’s education as an instrument for the self-reported education in two-stage least squares estimation:

```
. ivreg lwage age age2 female white (educ = educt_t)

Instrumental variables (2SLS) regression
```

Source	SS	df	MS			
Model	88.55968	5	17.711936	Number of obs =	680	
Residual	173.167986	674	.256925795	F( 5, 674) =	67.58	
Total	261.727666	679	.38546048	Prob > F =	0.0000	
				R-squared =	0.3384	
				Adj R-squared =	0.3335	
				Root MSE =	.50688	

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.1160166	.0103696	11.188	0.000	.095656	.1363771
age	.1040319	.0105022	9.906	0.000	.083411	.1246528
age2	-.0010624	.0001262	-8.421	0.000	-.0013101	-.0008147
female	-.3156432	.0400738	-7.877	0.000	-.3943277	-.2369586
white	-.0980293	.0722449	-1.357	0.175	-.2398815	.0438228
_cons	-1.187928	.2685695	-4.423	0.000	-1.715262	-.6605948

```
Instrumented:  educ
Instruments:  educt_t + age age2 female white
```

Here, the estimated return to education is about 6% higher than the estimate without instrumenting. Under the assumption that the self-reports of education are measured with classical measurement error and that the measurement error of the other twin’s report of the individual’s education is not correlated with this measurement error, then the instrumental variable will purge the attenuation bias in the “conventional” estimated return to education induced by the measurement error. This explains why the estimated return to education increases slightly.

c.

If there are unobservable, confounding variables that are correlated with both educational attainment and earnings, then there will be omitted variables bias in the least squares estimated return to education when these confounding variables are not controlled for. Intuitively, since education is not randomly assigned, if there are variables that drive both education and earnings (e.g., ability, motivation, etc.), we will inappropriately assign wage effects to education that are actually attributable to the other factors we have not controlled for. Suppose that the omitted variable is  $Z$  and that the “correct” regression model is:

$y_i = \alpha + \beta S_i + \theta Z_i + \varepsilon_i$ , where  $y_i$  is log-wages and  $S_i$  is schooling. Further suppose that  $Z_i$  is related to  $S_i$  in the following way:  $Z_i = \lambda S_i + u_i$ . Then the omitted variables bias from regressing  $y$  on  $S$  without controlling for  $Z$  is:  $\text{Bias}(\hat{\beta}) = \theta\lambda$ . For omitted variables bias, the omitted factors must be both correlated with schooling and correlated with earnings. If the omitted variables are positively correlated with both schooling and earnings, then the bias in the misspecified regression estimate is positive.

d.

If the omitted factors are identical between identical twins, then one can “first-difference” the data and compare differences in earnings and education across twin pairs to get an unbiased estimate of the return to education. The following are the STATA regression results from the first-differenced data:

```
. reg dlwage deduc if first==1, noconstant
```

Source	SS	df	MS	Number of obs = 340		
Model	2.80171106	1	2.80171106	F( 1, 339)	=	10.88
Residual	87.2987795	339	.257518524	Prob > F	=	0.0011
				R-squared	=	0.0311
				Adj R-squared	=	0.0282
				Root MSE	=	.50746

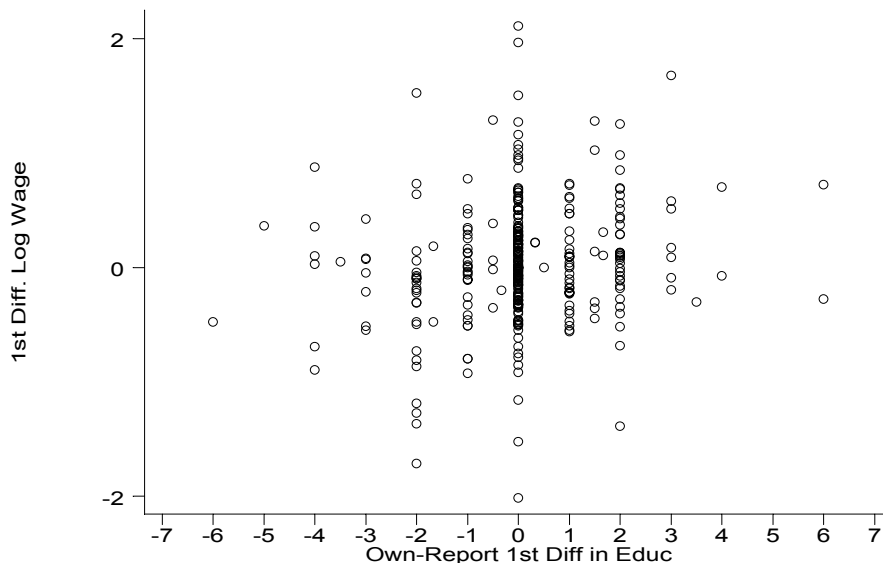
  

dlwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
deduc	.0617006	.018706	3.298	0.001	.0249061	.0984951

The estimated return to education is much lower (by over 40%) than the one based on the “conventional” log-wage regression. This may imply that the conventional cross-sectional estimate of the return to education is biased up since it does not control for “ability” differences across individuals with different levels of schooling (a.k.a., “ability” bias). We didn’t include controls for age, gender, and race since these are identical between identical twins and get “differenced out” when comparing differences between twins. This highlights that many factors may get differenced out when comparing twins.

e.

The following is the scatter plot of the difference in twins’ log-wages and the difference in twins’ self-reported education.



Most of the observations are clustered around zero, implying that most twins have identical education levels. This could imply that some of the observed differences in self-reported education between twins are attributable to measurement error. This could be another explanation of the result found in part d, since it could be that the estimated return to education based on first-differences is attenuated down due to the measurement error. In particular, if the measurement error is classical and the measurement errors in the twins' self-reports are independent of each other, then first-differencing the data will absorb much of the true signal in education while it absorbs none of the measurement error variance. This is because the education levels of twins are highly correlated (a correlation of about 0.7). This could exacerbate the attenuation bias attributable to measurement error quite a bit since the bias has the form:

$$\text{Bias}(\hat{\beta}_D) = -\frac{\lambda\beta}{(1-r_S)}, \text{ where } r_S \text{ is the correlation between two twins' education levels. Notice the}$$

similarity of this bias to the one shown above when additional regressors are included as controls. One way to think about this is that differencing the data is like including dummy variables for every twin pair (family) in the regression. Consequently, a lot of the signal may be absorbed relative to the amount of noise variance that is absorbed, and some of the observed differences in education between twins may be attributable to either or both twins misreporting.

f.

The following are the results from instrumenting the difference between the twins' self-reports with the difference between the twins' reports of the other twin's education using two-stage least squares:

```
. ivreg dlwage (deduc = deduct) if first==1, noconstant
Instrumental variables (2SLS) regression
```

Source	SS	df	MS			
Model	1.25589635	1	1.25589635	Number of obs =	340	
Residual	88.8445942	339	.262078449	F( 1, 339) =	.	
Total	90.1004906	340	.265001443	Prob > F =	.	
				R-squared =	.	
				Adj R-squared =	.	
				Root MSE =	.51194	

dlwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
deduc	.1075312	.0295439	3.640	0.000	.0494187	.1656438

```
Instrumented:   deduc
Instruments:   deduct
```

The estimated return to education has increased because under the assumption that the instrument is valid, the potentially large attenuation bias due to measurement error has now been purged. This conclusion depends crucially on the assumption that all of the potential reporting errors (associated with both the self-reports and the other twins' reports of each twin's education) are independent of each other.

If the measurement errors between an individual's self-report and his twin's report of the individual's education are correlated, then this instrumental variables estimate will be biased. In class, we discussed that under one scenario for correlated measurement errors, one twin's report of the difference in education is a valid instrument for the other twin's report of the education difference. Although this was not required to receive full credit, the below two-stage least squares regression uses twin 2's report of the education difference (dceduct) as an instrument for twin 1's report of the education difference (dceduc).

```
. ivreg dlwage (dceduc = dceduct) if first==1, noconstant
Instrumental variables (2SLS) regression
```

Source	SS	df	MS			
Model	3.29999374	1	3.29999374	Number of obs =	340	
Residual	86.8004968	339	.256048663	F( 1, 339) =	.	
Total	90.1004906	340	.265001443	Prob > F =	.	
				R-squared =	.	
				Adj R-squared =	.	
				Root MSE =	.50601	

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dceduc	.0877677	.0249433	3.519	0.000	.0387047	.1368308

Instrumented: dceduc  
Instruments: dceduct

The resulting instrumental variables (IV) estimate of the return to education is lower than the IV estimate based on the assumption that the measurement errors are independent. Thus, there appears to be some evidence of correlated measurement errors. See Ashenfelter and Krueger (1994) for more details on potential solutions to the problem of correlated reporting errors. Also, look at your class notes.

g. If the reporting errors are uncorrelated with each other, then the estimate in the first part of f. is purged of attenuation bias and is also more likely to be purged of ability bias than the “conventional” estimate of the return to education based on the regression of log-wages on educ, age, age2, female, and white. Since this estimate is only slightly lower than the conventional estimate, we might conclude that the size of the omitted variables bias in the conventional estimate is not very big. However, this is one of the more controversial debates ongoing in labor economics.

Consider the following model for log-wages ( $y_{ij}$ ):

$$y_{ij} = \gamma s_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} = \alpha_i + u_{ij},$$

where  $i$  indexes the family,  $j$  indexes the twin (1 or 2), and  $\alpha_i$  represents the unobservable factors shared by twins in the same family (i.e., the family effect). Suppose we compare twin differences in log-wages and education in order to difference out  $\alpha_i$ :

$$(y_{i1} - y_{i2}) = \gamma(s_{i1} - s_{i2}) + (u_{i1} - u_{i2}).$$

Comparing twin differences will reduce the omitted variables problem if the following holds:

$$\frac{\text{Cov}(s_{i1} - s_{i2}, \varepsilon_{i1} - \varepsilon_{i2})}{\text{Var}(s_{i1} - s_{i2})} < \frac{\text{Cov}(s_{ij}, \varepsilon_{ij})}{\text{Var}(s_{ij})}$$

Although it is believable that the numerator on the left-hand-side is less than the numerator on the right-hand-side, we also know that the denominator on the left-hand-side is much lower. That is, the variation in twin-differences in education across families is much lower than the overall variation in education across individuals. Consequently, the variation in the “treatment” is greatly reduced by focusing on twin differences, and even if the correlation between this treatment and the unobservables is smaller, the omitted variables bias may not fall. On the other hand, it is informative that the estimated return to education based on twin-differences is similar to the conventional OLS estimate.