Economics 475: Econometrics

Homework #4: Answers

This homework is Monday, February 13th.

Your Midterm Exam will occur on Wednesday, February 15th.

1. A large number of regressions investigating why some counties experience higher murder rates. These regressions typically estimate equations similar to:

(1) $M_i = \beta_0 + \beta_1 P_i + \beta_2 U_i + e_{1i}$

where M is the number of murders per 100,000 residents, P is the number of policemen per 100,000 residents, U is the unemployment rate, i indexes counties, and e_{1i} is mean zero, variance σ_1^2.

a. What signs do you expect β_1 and β_2 to take?

I would expect counties with more police to have lower crime rates (B_1<0) and with higher unemployment rates to have greater crime rates (B_2 >0).

b. Many have argued that crime is not an exogenous variable. Indeed, one might think of murders being determined simultaneously with police presence. Consider the simultaneous system of equations:

(2) $M_i = \beta_0 + \beta_1 P_i + \beta_2 U_i + e_{1i}$

(3) $P_i = \alpha_0 + \alpha_1 M_i + \alpha_2 Inc_i + e_{2i}$

where Inc_i is the county's level of per capita income.

What are the reduced form equations for M and P?

$$M_i = \frac{\beta_0 + \beta_1\alpha_0}{1-\beta_1\alpha_1} + \frac{\beta_1\alpha_2}{1-\beta_1\alpha_1}Inc_i + \frac{\beta_2}{1-\beta_1\alpha_1}U_i + \frac{\beta_1}{1-\beta_1\alpha_1}e_2 + \frac{1}{1-\beta_1\alpha_1}e_1$$

$$P_i = \frac{\alpha_0 + \alpha_1\beta_0}{1-\beta_1\alpha_1} + \frac{\alpha_2}{1-\beta_1\alpha_1}Inc_i + \frac{\alpha_1\beta_2}{1-\beta_1\alpha_1}U_i + \frac{\alpha_1}{1-\beta_1\alpha_1}e_1 + \frac{1}{1-\beta_1\alpha_1}e_2$$

c. If equations (2) and (3) describe the murder rate, what is the covariance between e_1 and P? What is the covariance between e_1 and U? Given these covariances, what will happen to an OLS estimate of (2)? Specifically, what will $\hat{\beta}_1$ and $\hat{\beta}_2$ be relative to their true values?

A high M (caused by a high e) would lead to counties hiring more police; thus a positive correlation occurs between P and e. Specifically, the covariance is $E[e_1(P-\bar{P})] = \frac{\alpha_1}{1-\beta_1\alpha_1}\sigma_{e1}^2$.

The covariance between e_1 and U is zero.

Estimating the regression in (1) would thus lead to biased coefficients (the estimate of B_1 would be biased in a positive manner. The estimate of B_2 is biased in a direction that depends upon U's correlation with P and M).

d. Are structural equations (1) and (2) over, exactly, or underidentified?

In this case, there are two exogenous variables, U and Inc. In equation (1) there are two slope variables. Since there are as many slope variables as exogenous variables, equation (1) is exactly identified. Likewise, equation (2) is exactly identified.

e. When I solve for the reduced form equations for M and P, I get:

(3) $M_i = \Pi_0 + \Pi_1 Inc_i + \Pi_2 U_i + w_i$

(4) $P_i = \Pi_3 + \Pi_4 Inc_i + \Pi_5 U_i + v_i$

where the Π's are functions of the α's and β's and the w's and v's are functions of the random error terms and the α's and β's. After using OLS to estimate equations (3) and (4), I find: $\hat{\Pi}_0 = .01$, $\hat{\Pi}_1 = -5$, $\hat{\Pi}_2 = 12$, $\hat{\Pi}_3 = 8$, $\hat{\Pi}_4 = 7$, $\hat{\Pi}_5 = 1$

What are your ILS estimates of β_0, β_1, β_2, α_0, α_1, α_2?

Using these six estimates and the six equations given in part c, I can isolate each α and β. I find:

(3) $M_i = 5.72429 - .714286P_i + 12.7143\ U_i + e_{1i}$

(4) $P_i = 7.9917 + .083333M_i + 7.41667Inc_i + v_i$

2. Perhaps the most frequently estimated regression is known as a Mincer Earnings Equation which expresses the natural log of wages as a function of individual observables including things like gender, age, experience and education. Economists have used the Mincer Earnings Equation to estimate the returns to education; that is the percent increase in wages given another year of education. However, this estimation is commonly criticized as having omitted variable bias; namely individuals going to school longer likely have characteristics that simultaneously make them better students and lead to higher pay. Thus, the coefficient on education is probably biased.

a. If one estimates the regression:

$$\ln(Wage_i) = \beta_0 + \beta_1 Educ_i + \varepsilon_i$$

but one omits variables such as ability and motivation, in what direction will OLS' estimate of β_1 be biased? What assumptions are you making in order to identify the direction of this bias?

If ability/motivation are positively related to both education and wages, than omitting ability/motivation will cause OLS to overestimate β_1.

b. Economists have long sought an instrumental variable that could be used to eliminate the bias from the regression in part a. What characteristics does such an instrument require? Some possible instruments suggested for this problem have been: 1) the number of siblings an individual has; 2) the distance from the nearest college an individual lives; 3) the education of an individual's parents. Comment on if these are appropriate or not.

Any instrument must be correlated with the independent variable but not the error term of our structural equation. In this case, we want an instrument that is correlated with education but not correlated with the part of wages that is unexplained by the regression.

1. Number of siblings is correlated with education (more siblings, the harder it is for parents to provide an education for any individual child) but it is also probably correlated with the error term (siblings may provide social skills and an environment in which individuals learn job skills).
2. The distance the nearest college is to an individual is probably correlated with education (closer colleges are less expensive to attend) but might be correlated with the error term, especially if parents choose to live near colleges for their amenities (which would show the parents care about things that likely influence wages of their children).
3. The education of a child's parents is also likely correlated with their education and, again, probably correlated with the error term. More educated parents convey skills/opportunities to their children differently than less educated ones.

c. One famous idea for an instrument was proposed by Joshua Angrist and Alan Krueger in a 1991 paper published by the Quarterly Journal of Economics. Before introducing this instrument, open the data set entitled "NEW7080.dta." This is the original data used by Angrist and Krueger and contains 247,199 observations of men born between 1920 and 1929 from the 1970 U.S. Census. Using this data estimate the equation:

$$LWKLYWGE = \beta_0 + \beta_1 EDUC_i + \beta_2 BLACK_i + \beta_3 MARRIED_i + \beta_4 SMSA_i + \beta_5 NEWENG_i + \beta_6 MIDATL_i + \beta_7 ENOCENT_i + \beta_8 WNOCENT_i + \beta_9 SOATL_i + \beta_{10} ESOCENT_i + \beta_{11} WSOCENT_i + \beta_{12} MT_i + \beta_{13} YR20_i + \beta_{14} YR21_i + \beta_{15} YR22_i + \beta_{16} YR23_i + \beta_{17} YR24_i + \beta_{18} YR25_i + \beta_{19} YR26_i + \beta_{20} YR27_i + \beta_{21} YR28_i + \beta_{23} AGE_i + \beta_{24} AGEQSQ_i$$

In this case, the dependent variable is the natural log of weekly wages, EDUC is the years of education, BLACK and MARRIED are dummy variables, SMSA is a dummy variable indicating if an individual lives in a city, the next 8 variables are location dummy variables (e.g., NEWENG = new England); AGE and AGESQ are age and age squared, and the dummy variables starting with YR indicate the year the individual was born.

What is your estimate of β_1? How do you interpret this number?

I find:

```
. reg LWKLYWGE EDUC BLACK MARRIED SMSA NEWENG MIDATL ENOCENT WNOCENT SOATL ESOCENT WS
> OCENT MT YR20 YR21 YR22 YR23 YR24 YR25 YR26 YR27 YR28 AGE AGEQSQ
```

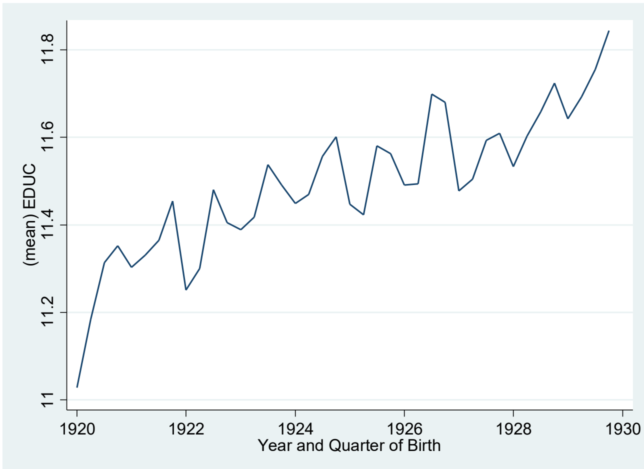
Source	SS	df	MS			
				Number of obs	=	247,199
				F(23, 247175)	=	3203.34
Model	24077.2575	23	1046.83728	Prob > F	=	0.0000
Residual	80775.7623	247,175	.326795842	R-squared	=	0.2296
				Adj R-squared	=	0.2296
Total	104853.02	247,198	.424166133	Root MSE	=	.57166

LWKLYWGE	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
EDUC	.0701244	.0003547	197.68	0.000	.0694291	.0708196
BLACK	-.2979589	.0043445	-68.58	0.000	-.3064741	-.2894437
MARRIED	.2928037	.0037449	78.19	0.000	.2854638	.3001437
SMSA	-.1343198	.0025648	-52.37	0.000	-.1393467	-.1292928
NEWENG	-.0327318	.0059551	-5.50	0.000	-.0444037	-.0210599
MIDATL	-.0131083	.0041124	-3.19	0.001	-.0211684	-.0050481
ENOCENT	.0197556	.0040477	4.88	0.000	.0118222	.027689
WNOCENT	-.1414295	.0054027	-26.18	0.000	-.1520186	-.1308404
SOATL	-.103773	.0044283	-23.43	0.000	-.1124524	-.0950936
ESOCENT	-.2077559	.0058936	-35.25	0.000	-.2193071	-.1962046
WSOCENT	-.1513897	.0050703	-29.86	0.000	-.1613274	-.141452
MT	-.1268585	.006706	-18.92	0.000	-.1400021	-.113715
YR20	-.0184507	.0384707	-0.48	0.632	-.0938523	.0569508
YR21	-.0106333	.0337788	-0.31	0.753	-.0768387	.0555722
YR22	-.0089803	.0292744	-0.31	0.759	-.0663575	.0483968
YR23	-.0026575	.0249214	-0.11	0.915	-.0515028	.0461877
YR24	.0015686	.0206755	0.08	0.940	-.0389548	.042092
YR25	.012714	.0166376	0.76	0.445	-.0198953	.0453233
YR26	.0147386	.0127953	1.15	0.249	-.0103399	.039817
YR27	.0167465	.0092345	1.81	0.070	-.0013529	.0348459
YR28	.0161007	.0064108	2.51	0.012	.0035356	.0286657
AGE	-.0021751	.0042163	-0.52	0.606	-.0104389	.0060887
AGEQSQ	.0000618	.0000729	0.85	0.397	-.0000811	.0002047
_cons	4.176986	.107386	38.90	0.000	3.966512	4.387459

The coefficient of .07 indicates that for each additional year of education, an individual can expect a 7% increase in their weekly wages.

d. Angrist and Krueger argue that the quarter-of-birth of an individual might be correlated with their education. Their argument has to do with the fact that individuals are required to attend school until the age of 16 (in many states). Someone born at the beginning of the year (quarter 1) will reach the age of 16 at an earlier point in their grade than someone born later in the year (say quarter 4). Thus, among two students dropping out of school at age 16, one will have more school than the other because they were born earlier in the year.

As evidence, they present this graph:



In this graph, the lowest points within a year are the first quarter of the year and the highest are the fourth. I made this graph using your data set and the following commands:

```
gen y = YOB + 0*QTR1 + .25*QTR2 + .5*QTR3 + .75*QTR4
collapse EDUC, by(y)
label variable y "Year and Quarter of Birth"
line EDUC y
```

Comment on the quarter of birth as an instrument.

In hindsight (and many, many research papers that have investigated this) we know a lot about quarter of birth as an instrument. From the graph above, it does appear that quarter of birth is connected to education and it is hard to imagine that the quarter you are born in influences your wages directly.

However, it turns out that quarter of birth is what is known as a "weak" instrument in that it doesn't explain much of education. Looking at the graph, it appears that at most, quarter of birth accounts for around .1 years of education (within year of birth—in other words, someone born in early 1924 on average has about .1 years of education less than someone born later in the year. Remember what an instrument does, it finds the exogenous variation in our X variable (education in this case) and uses that variation to explain wages. In this case, we are hoping to explain wage differences using a difference in education of about .1 years—or about one month. Trying to see what happens to someone's wages if they earn an additional month of education is going to be difficult.

e. From the graph in part d, it is clear that education is a function of the quarter of birth and the year of birth (there is more education for people born later in the decade). Angrist and Krueger propose as the instruments all possible dummy variables that represent year and quarter of birth (i.e., one dummy variable for 1920 quarter 1, another for 1920 quarter 2, etc.). Fortunately, these variables were included in your data set entitled QTR120, QTR121, QTR122, etc.

Using these instruments, estimate your first stage regression (don't forget the other exogenous variables from part c). What do you find? Evaluate if these are good instruments or not.

My results are:

r(198);

```
. reg EDUC BLACK MARRIED SMSA NEWENG MIDATL ENOCENT WNOCENT SOATL ESOCENT WSOCENT MT YR20 YR21 YR2
> 2 YR23 YR24 YR25 YR26 YR27 YR28 AGE AGEQSQ QTR120- QTR329
note: YR20 omitted because of collinearity
note: QTR220 omitted because of collinearity
```

Source	SS	df	MS
Model	195087.803	50	3901.75607
Residual	2596780.99	247,148	10.5069877
Total	2791868.8	247,198	11.294059

Number of obs	=	247,199
F(50, 247148)	=	371.35
Prob > F	=	0.0000
R-squared	=	0.0699
Adj R-squared	=	0.0697
Root MSE	=	3.2414

EDUC	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
BLACK	-2.317499	.0241911	-95.80	0.000	-2.364913	-2.270085
MARRIED	.40606	.0212205	19.14	0.000	.3644684	.4476517
SMSA	-.5882687	.0144953	-40.58	0.000	-.6166791	-.5598582
NEWENG	-.4105518	.0337588	-12.16	0.000	-.4767181	-.3443855
MIDATL	-.455914	.0233019	-19.57	0.000	-.5015851	-.4102429
ENOCENT	-.7393628	.0229049	-32.28	0.000	-.7842558	-.6944699
WNOCENT	-.5818435	.0306148	-19.01	0.000	-.6418476	-.5218393
SOATL	-1.104921	.0250126	-44.17	0.000	-1.153946	-1.055897
ESOCENT	-1.652889	.0332536	-49.71	0.000	-1.718066	-1.587713
WSOCENT	-1.139735	.0286601	-39.77	0.000	-1.195908	-1.083562
MT	-.155019	.0380265	-4.08	0.000	-.2295499	-.080488
YR20	0	(omitted)				
YR21	.0246532	.0626896	0.39	0.694	-.0982168	.1475231
YR22	-.0753639	.0648424	-1.16	0.245	-.2024533	.0517254
YR23	-.073623	.0653613	-1.13	0.260	-.2017294	.0544834
YR24	-.04658	.0640866	-0.73	0.467	-.1721881	.079028
YR25	-.1545625	.0631237	-2.45	0.014	-.2782833	-.0308417
YR26	-.0809895	.0611047	-1.33	0.185	-.2007532	.0387741
YR27	-.1566795	.0578458	-2.71	0.007	-.2700557	-.0433032
YR28	-.1114989	.0576259	-1.93	0.053	-.2244442	.0014465
AGE	.2413608	.1144003	2.11	0.035	.0171392	.4655823
AGEQSQ	-.0033822	.0012372	-2.73	0.006	-.0058072	-.0009573
QTR120	-.2772703	.0755927	-3.67	0.000	-.42543	-.1291106
QTR121	-.1429502	.0632063	-2.26	0.024	-.2668329	-.0190674
QTR122	-.1605565	.0649139	-2.47	0.013	-.2877861	-.0333269
QTR123	-.1172529	.0655002	-1.79	0.073	-.2456315	.0111257
QTR124	-.163552	.0651411	-2.51	0.012	-.2912269	-.0358771
QTR125	-.0903028	.0666899	-1.35	0.176	-.2210132	.0404077
QTR126	-.1896083	.0682698	-2.78	0.005	-.3234153	-.0558013
QTR127	-.1805322	.0675823	-2.67	0.008	-.3129916	-.0480727
QTR128	-.2186307	.0697671	-3.13	0.002	-.3553723	-.0818891
QTR129	-.2469622	.0678914	-3.64	0.000	-.3800275	-.1138969
QTR220	0	(omitted)				
QTR221	.0328102	.0836292	0.39	0.695	-.1311009	.1967213
QTR222	.0290673	.0838653	0.35	0.729	-.1353065	.193441
QTR223	.0655263	.0825516	0.79	0.427	-.0962726	.2273252
QTR224	.0127437	.0807353	0.16	0.875	-.1454954	.1709828
QTR225	.0599339	.0808523	0.74	0.459	-.0985344	.2184023
QTR226	-.0244348	.0802388	-0.30	0.761	-.1817006	.1328311
QTR227	.0213937	.0784145	0.27	0.785	-.1322966	.175084
QTR228	.0126851	.078441	0.16	0.872	-.1410571	.1664273
QTR229	-.0240162	.0769025	-0.31	0.755	-.1747431	.1267108
QTR320	.0662101	.0516619	1.28	0.200	-.0350458	.1674661
QTR321	-.0266348	.0654477	-0.41	0.684	-.1549105	.101641
QTR322	.1279635	.0654735	1.95	0.051	-.0003629	.2562899
QTR323	.1109581	.0655973	1.69	0.091	-.0176108	.239527
QTR324	.0313575	.0637611	0.49	0.623	-.0936126	.1563276
QTR325	.1133112	.0643369	1.76	0.078	-.0127874	.2394099
QTR326	.0911089	.0644221	1.41	0.157	-.0351567	.2173746
QTR327	.0224528	.0627369	0.36	0.720	-.1005099	.1454154
QTR328	.0019646	.064042	0.03	0.976	-.1235561	.1274853
QTR329	-.0378384	.0627387	-0.60	0.546	-.1608047	.0851279
_cons	8.384242	2.5794	3.25	0.001	3.328686	13.4398

Notice, the 1st quarter births (starting with QTR1) are all negative and smaller than the 2nd and 3rd quarter ones, and smaller than the fourth quarter (which are the omitted dummy variables).

To determine if these are good instruments, we need to determine if the quarter variables are statistically correlated with education AND if they are uncorrelated with the error terms. Since we don't observe the error terms, we cannot accomplish the second of these tasks. However, we can test if all the QTR variables are statistically different than zero through an F-test. I do this in Stata using the test command (or you can do it by estimating a restricted version of this regression and constructing the F-test yourself).

```
. test QTR120 QTR121 QTR122 QTR123 QTR124 QTR125 QTR126 QTR127 QTR128 QTR129 QTR220 QTR221 QTR222 QTR223 QTR224 QTR22
> 5 QTR227 QTR228 QTR229 QTR320 QTR321 QTR322 QTR323 QTR324 QTR325 QTR326 QTR327 QTR328 QTR329

 ( 1)  QTR120 = 0
 ( 2)  QTR121 = 0
 ( 3)  QTR122 = 0
 ( 4)  QTR123 = 0
 ( 5)  QTR124 = 0
 ( 6)  QTR125 = 0
 ( 7)  QTR126 = 0
 ( 8)  QTR127 = 0
 ( 9)  QTR128 = 0
 (10)  QTR129 = 0
 (11)  o.QTR220 = 0
 (12)  QTR221 = 0
 (13)  QTR222 = 0
 (14)  QTR223 = 0
 (15)  QTR224 = 0
 (16)  QTR225 = 0
 (17)  QTR227 = 0
 (18)  QTR228 = 0
 (19)  QTR229 = 0
 (20)  QTR320 = 0
 (21)  QTR321 = 0
 (22)  QTR322 = 0
 (23)  QTR323 = 0
 (24)  QTR324 = 0
 (25)  QTR325 = 0
 (26)  QTR326 = 0
 (27)  QTR327 = 0
 (28)  QTR328 = 0
 (29)  QTR329 = 0
       Constraint 11 dropped

       F( 28,247148) =    2.87
            Prob > F =    0.0000
```

Here, we find that the QTR variables are statistically different than zero—but not by much. A F-statistics of 2.87 is not large though it is statistically significant. Thus, it appears that quarter of birth does explain years of education but it doesn't provide large differences in years of education between people with different birth quarters.

f. Estimate equation c using the instruments developed from the first stage in part e. What do you find? Do your results change relative to those found in part c?

I create the variable instrument in the first line of the command, below. It follows immediately after the commands in part e. I find:

```
. predict instr, xb

. reg LWKLYWGE instr BLACK MARRIED SMSA NEWENG MIDATL ENOCENT WNOCENT SOATL ESOCENT WSOCENT MT YR20 YR21 YR22 YR23 YR
> 24 YR25 YR26 YR27 YR28 AGE AGEQSQ
```

Source	SS	df	MS			
				Number of obs	=	247,199
				F(23, 247175)	=	1299.29
Model	11309.4716	23	491.716155	Prob > F	=	0.0000
Residual	93543.5482	247,175	.378450686	R-squared	=	0.1079
				Adj R-squared	=	0.1078
Total	104853.02	247,198	.424166133	Root MSE	=	.61518

LWKLYWGE	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
instr	.1034924	.035891	2.88	0.004	.033147	.1738378
BLACK	-.2206243	.0833085	-2.65	0.008	-.3839067	-.0573419
MARRIED	.2792428	.015132	18.45	0.000	.2495844	.3089011
SMSA	-.1146864	.0212963	-5.39	0.000	-.1564266	-.0729462
NEWENG	-.0190226	.0160774	-1.18	0.237	-.0505339	.0124888
MIDATL	.0020947	.0169398	0.12	0.902	-.0311069	.0352962
ENOCENT	.0444112	.0268737	1.65	0.098	-.0082605	.097083
WNOCENT	-.122022	.0216683	-5.63	0.000	-.1644913	-.0795527
SOATL	-.0669082	.0399353	-1.68	0.094	-.1451804	.011364
ESOCENT	-.1525944	.0596669	-2.56	0.011	-.26954	-.0356489
WSOCENT	-.1133815	.0412422	-2.75	0.006	-.1942152	-.0325478
MT	-.1216926	.0091077	-13.36	0.000	-.1395434	-.1038417
YR20	-.0743603	.0730067	-1.02	0.308	-.2174515	.0687308
YR21	-.0620168	.0661486	-0.94	0.348	-.1916663	.0676328
YR22	-.0521483	.0561082	-0.93	0.353	-.162119	.0578224
YR23	-.0403016	.0485647	-0.83	0.407	-.1354871	.0548839
YR24	-.0293715	.0400306	-0.73	0.463	-.1078303	.0490873
YR25	-.0102282	.0304868	-0.34	0.737	-.0699816	.0495251
YR26	-.0027782	.0233356	-0.12	0.905	-.0485154	.042959
YR27	.0080884	.0136189	0.59	0.553	-.0186042	.034781
YR28	.012718	.0077995	1.63	0.103	-.0025688	.0280047
AGE	-.0028613	.0045969	-0.62	0.534	-.0118711	.0061486
AGEQSQ	.0001609	.0001324	1.22	0.224	-.0000985	.0004204
_cons	3.625952	.6038256	6.00	0.000	2.44247	4.809434

The coefficient on the instrument is .10 suggesting that an additional year of education raises wages by 10 percent.

Notice, this is the opposite direction of what we would have expected. We believed that OLS overstated the returns to education in our original, uncorrected model. However, after correcting it, we find that the returns actually rose. This should suggest that our instrument is questionable—now that we know it is a "weak" instrument we probably shouldn't trust these results.

3. Suppose you want to test whether girls who attend a girls' high school do better in math than girls who attend coed schools. You have a random sample of senior high school girls and measure the variable *score*, an outcome of a mathematics standardized test. Let *girlhs* be a dummy variable indicating whether a student attends a girls' high school. Consider the regression $Score_i = B_0 + B_1 Girlhs_i + \varepsilon_i$.

a. Suppose that parental support and motivation are unmeasured factors in ε. How does this fact impact estimates of B_1?

In this case, parental support is positively correlated with Girlshs and with score so the OLS coefficient B_1 will be positively biased.

b. Consider the variable *Numgirl* where *Numgirl* is the number of girls' high schools within a 20 mile radius of the observation's home. Under what conditions could *Numgirl* be used as a valid IV for *Girlhs*.

Numgirl must be correlated with girlhs but not with any part of Score that isn't explained by girlhs.

4. Describe the data you will use in your final project. If possible, show me a regression from this data.