

PART I ANSWERS

. * PROBLEM 1;

Variable	Obs	Mean	Std. Dev.	Min	Max
x	200	10	14.07089	-15	34
e2	200	.07804	1.959689	-5.903	5.965
e20	200	.089235	20.19952	-53.439	58.12
e50	200	5.106895	48.92489	-103.545	114.328

Of e2, e20, and e50, the mean of e2 is the closest to zero. The summary statistics do make sense. We were told that the e variables are normally distributed with means of 0 and standard deviations of 2, 20, and 50. Thus, e2 is less spread out than e20 or e50. In fact, it is somewhat surprising that the mean of e20 is so close to 0. Notice that the mean of e50 is fairly far away from 0, but this is expected as its standard deviation is 50. Finally, notice that each e variable has a sample standard deviation that is very close to its true standard deviation and that the maximum and minimum of each e variable is between 2 and 3 standard deviations away from its mean.

. * PROBLEM 2;

Variable	Obs	Mean	Std. Dev.	Min	Max
y2	200	92.07804	112.494	-107.249	285.987
y20	200	92.08923	114.7911	-130.262	314.116
y50	200	97.10689	119.8302	-173.059	342.328

These summary statistics make sense. As the mean of x is 10 exactly, the regression equation adds, on average, $12+8(10)=92$ to each error term. Thus, the mean of y2 is 92 plus the mean of e2, the mean of y20 is 92 plus the mean of e20, and the mean of y50 is 92 plus the mean of e50. Also notice that the standard deviations of y2, y20, and y50 are greater than the standard deviations of e2, e20, and e50 as the y variables must also reflect the variation inherent in the x variable.

. * PROBLEM 3;

Source	SS	df	MS	Number of obs = 200		
Model	2517564.67	1	2517564.67	F(1, 198)	=	.
Residual	762.620062	198	3.85161647	Prob > F	=	0.0000
-----				R-squared	=	0.9997
Total	2518327.29	199	12654.911	Adj R-squared	=	0.9997
-----				Root MSE	=	1.9626
y2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	7.993596	.0098872	808.48	0.000	7.974098	8.013094
_cons	12.14208	.170393	71.26	0.000	11.80606	12.4781

The estimated intercept is 12.14208, which is close to 12. The estimated slope is 7.993596, which is close to 8. The r-squared is 0.9997, which means that the estimated regression line explains almost all of the variation in the y variable. The standard error of the regression is 1.9626, which is fairly close to the true standard deviation of the error terms (e2).

. * PROBLEM 4;

Source	SS	df	MS	Number of obs = 200		
Model	2541063.07	1	2541063.07	F(1, 198)	=	6199.34
Residual	81158.657	198	409.892207	Prob > F	=	0.0000
-----				R-squared	=	0.9690
Total	2622221.72	199	13176.9936	Adj R-squared	=	0.9689
-----				Root MSE	=	20.246
y20	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	8.030815	.1019968	78.74	0.000	7.829675	8.231954
_cons	11.78109	1.757782	6.70	0.000	8.314711	15.24746

The estimated intercept is 11.78109, which is close to 12 but not as close as in problem 3. The estimated slope is 8.030815, which is close to 8 but not as close as in problem 3. The r-squared is 0.9690, which means that the estimated regression line explains over 96 percent of the variation in the y variable, which is quite a bit but not as much as in problem 3. The standard error of the regression is 20.246, which is fairly close to the true standard deviation of the error terms (e20).

. * PROBLEM 5;

Source	SS	df	MS	Number of obs = 200		
Model	2383114.29	1	2383114.29	F(1, 198)	=	994.68
Residual	474379.946	198	2395.85831	Prob > F	=	0.0000
-----				R-squared	=	0.8340
Total	2857494.24	199	14359.2675	Adj R-squared	=	0.8331
-----				Root MSE	=	48.948
y50	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	7.777219	.246594	31.54	0.000	7.290931	8.263507
_cons	19.33471	4.249724	4.55	0.000	10.95418	27.71523

The estimated intercept is 19.33471, which is not all that close to 12. The estimated slope is 7.777219, which is fairly close to 8 but not as close as in problem 4. The r-squared is 0.8331, which means that the estimated regression line explains over 83 percent of the variation in the y variable, which is quite a bit but not as much as in problem 4. The standard error of the regression is 48.948, which is fairly close to the true standard deviation of the error terms (e50).

. * PROBLEM 6;

When comparing the answers between problems 3 - 5, the estimated coefficients should all converge to the true values (12 & 8) as the number of observations gets big. Clearly this is happening, but it happens faster for the model in which the standard deviation of the error terms is smaller. The other statistics, such as the standard errors of the estimates, the r-squared, and the mean squared error, differ across regressions because the distribution of the error term differs across regressions.

. * PROBLEM 7;

(Graph omitted. See problem 10.)

The graph suggests that the mean error is zero, and errors are distributed roughly normal between -5 and 5.

Variable	Obs	Mean	Std. Dev.	Min	Max
y2	200	92.07804	112.494	-107.249	285.987
y2hat	200	92.07804	112.477	-107.7619	283.9243
error2	200	-1.96e-09	1.957616	-6.089904	5.842134

Notice that the average predicted value of y is exactly the average value of y. This must always be the case. Moreover, the average error term equals zero, which must also always be the case.

. * PROBLEM 8;

(Graph omitted. See question 10.)

The graph suggests that the mean error is zero, and errors are distributed roughly normal between -50 and 50.

Variable	Obs	Mean	Std. Dev.	Min	Max
y20	200	92.08923	114.7911	-130.262	314.116
y20hat	200	92.08923	113.0007	-108.6811	284.8288
error20	200	8.05e-09	20.19486	-54.11371	57.66098

As necessary, the average predicted value of y equals the average value of y , and the average error term equals zero.

. * PROBLEM 9;

(Graph omitted. See question 10.)

The graph suggests that the mean error is zero, and errors are distributed roughly normal between -100 and 100.

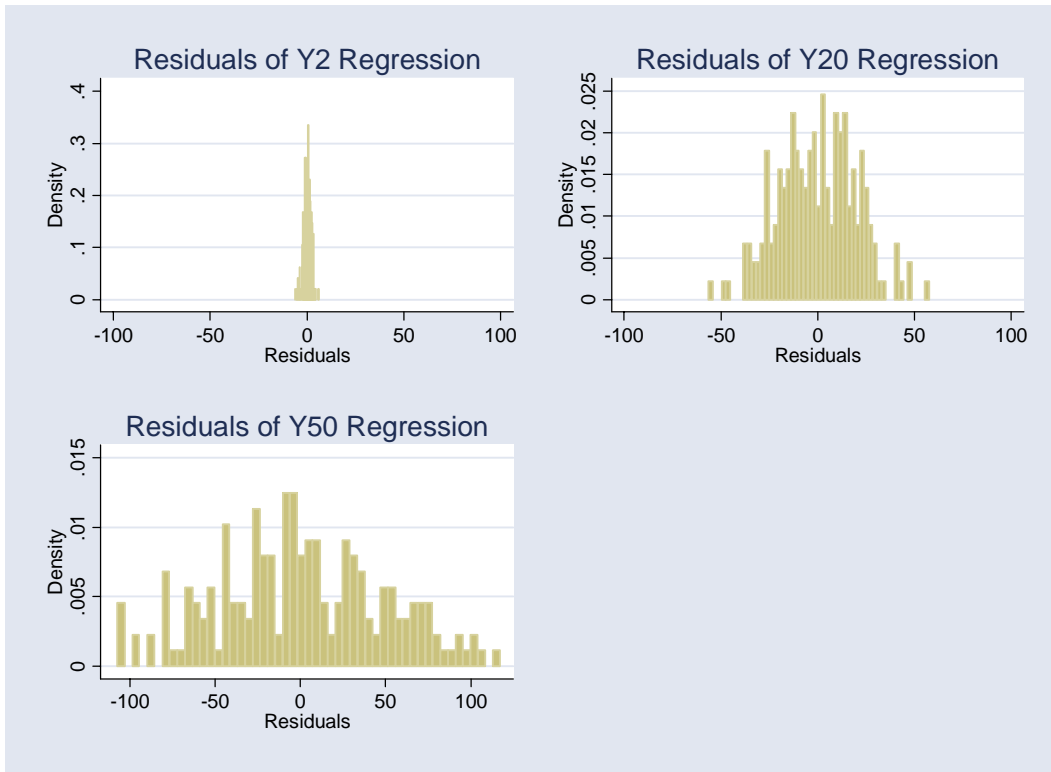
Variable	Obs	Mean	Std. Dev.	Min	Max
y50	200	97.10689	119.8302	-173.059	342.328
y50hat	200	97.1069	109.4324	-97.32358	283.7602
error50	200	-8.12e-08	48.82437	-107.3152	113.0084

As necessary, the average predicted value of y equals the average value of y , and the average error term equals zero.

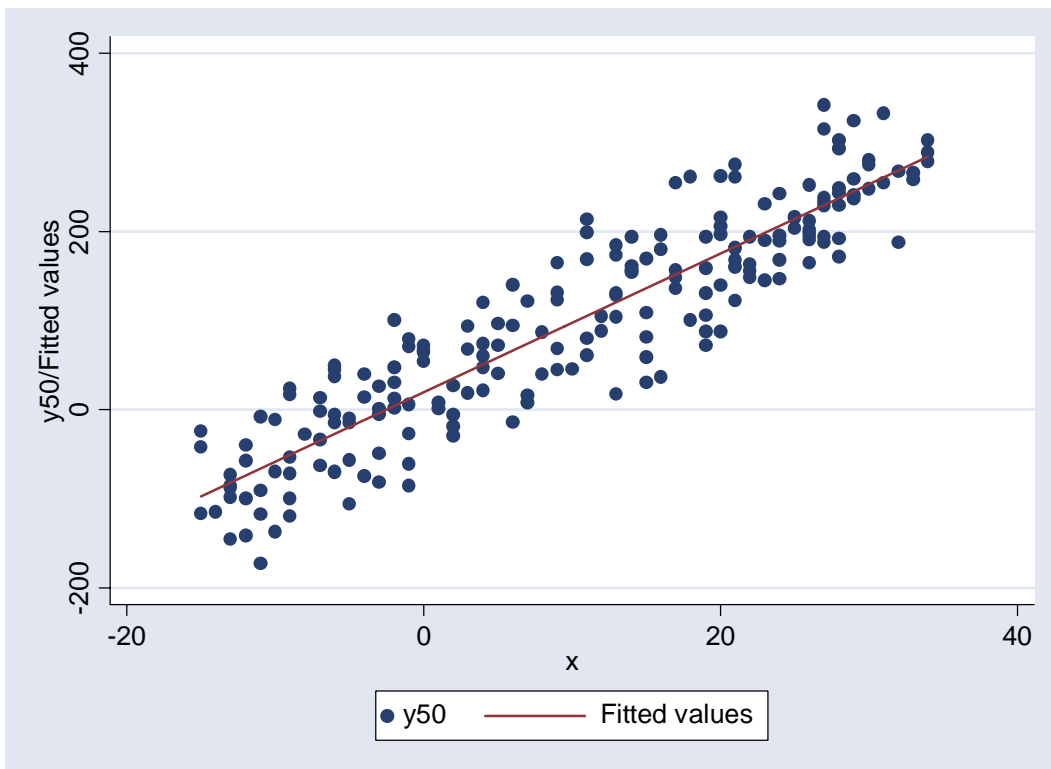
. * PROBLEM 10;

Regardless of the regression, the mean value of the predicted values must equal the average value of the dependent variable, and the mean value of the errors must equal zero (always). What is different is the distribution of the error terms. Because the standard deviation of the error terms is so much smaller for y_2 , the distribution of the error terms will be tighter and more bell-shaped. The distribution of the errors for y_{20} are more spread out, but still roughly normal. The errors for y_{50} , when the error terms are the most erratic, are greatly spread out, and normality is slower to kick in. (The graph is on the next page.)

Graph for problem 10.



. * PROBLEM 11;



PART II ANSWERS

* PROBLEM 1;

tab size;

size	Freq.	Percent	Cum.
Small	64	19.39	19.39
Middle	200	60.61	80.00
Large	66	20.00	100.00
Total	330	100.00	

Variable	Obs	Mean	Std. Dev.	Min	Max
-> size = Small					
enroll	64	437.5156	110.5547	204	617
dropout	64	.3489083	.325154	0	1.183432
lunch	64	28.70574	13.60803	5.857741	61.52381
act	64	21.62344	1.166428	18.9	25.2
-> size = Middle					
enroll	200	1359.955	625.7009	628	2849
dropout	200	.3544069	.2946964	0	1.772526
lunch	200	20.05279	11.4283	2.07305	69.51331
act	200	21.9595	.9177807	17.7	24.6
-> size = Large					
enroll	66	7636.515	12463.14	2853	99814
dropout	66	.5270103	.4997677	0	2.867333
lunch	66	17.81276	12.22709	.1482466	71.06017
act	66	22.55606	.9118246	19.1	24.6

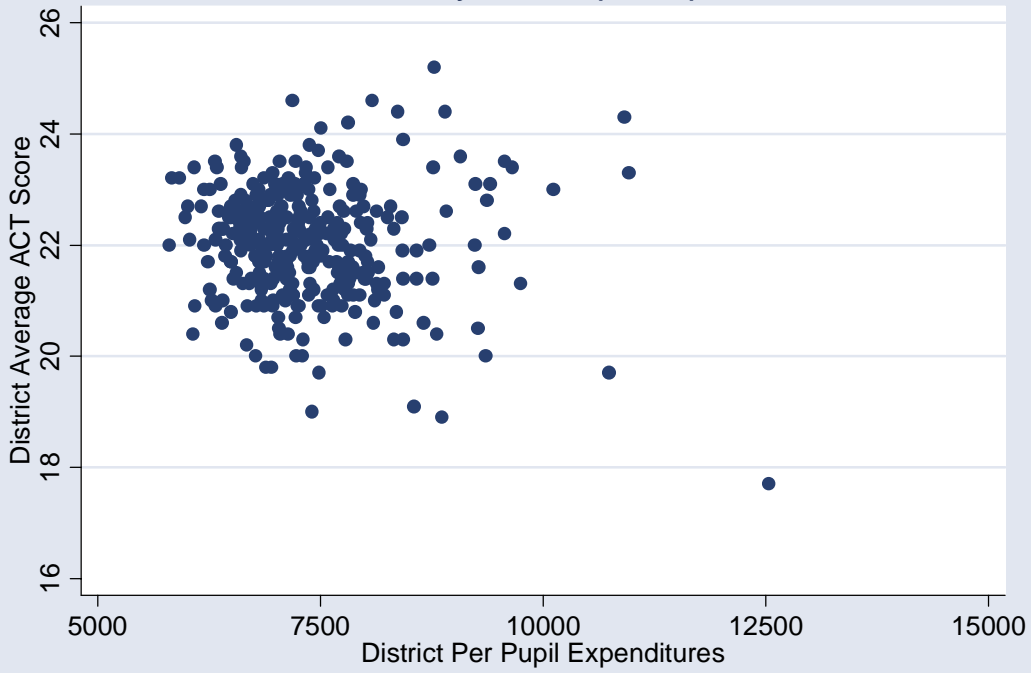
* PROBLEM 2;

Variable	Obs	Mean	Std. Dev.	Min	Max
bamin	330	25719.98	1301.497	22293	30815
bamax	330	36851.16	3648.96	27572	48476
mamin	330	28724.04	1746.136	24393	36032
mamax	330	44219.98	4048.267	32975	57209
ppexp	330	7377.191	875.6122	5807.12	12535.09

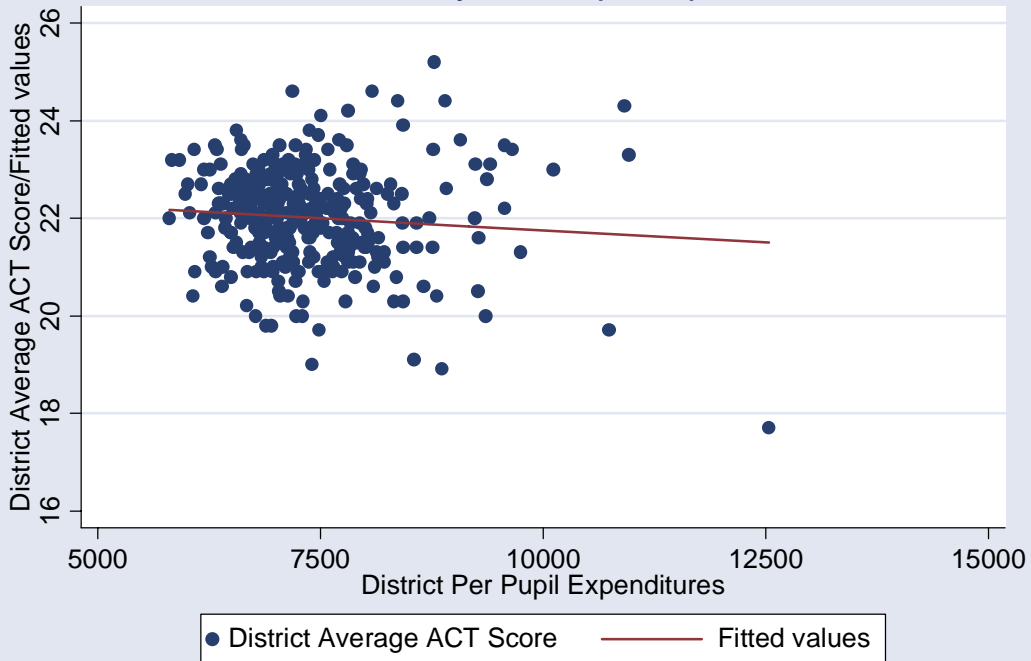
* PROBLEM 3;

The graphs follow below. In the first graph, there does not appear to be a clear relationship between expenditures and ACT scores. In the second graph, we see that the linear relationship is actually negative as the estimated regression line is sloped downward.

ACT Scores by Per Pupil Expenditures

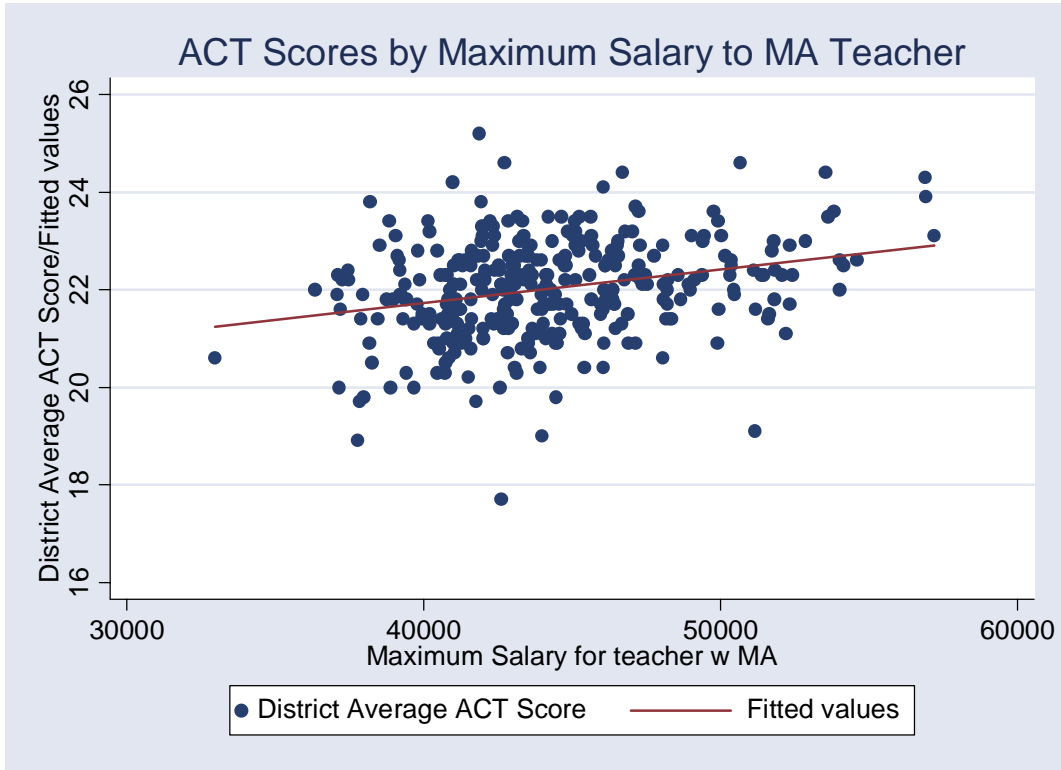


ACT Scores by Per Pupil Expenditures



* PROBLEM 4;

The estimated regression line implies that there is a positive relationship between the maximum salary paid to a teacher with a Masters Degree and ACT scores.



* PROBLEM 5 - MODEL 1;

Source	SS	df	MS	Number of obs =	330
Model	4.00811347	1	4.00811347	F(1, 328) =	3.95
Residual	333.120543	328	1.01561141	Prob > F =	0.0478
Total	337.128656	329	1.02470716	R-squared =	0.0119
				Adj R-squared =	0.0089
				Root MSE =	1.0078

act	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
bamin	.0000848	.0000427	1.99	0.048	8.26e-07 .0001688
_cons	19.83242	1.099378	18.04	0.000	17.6697 21.99514

The estimated coefficient on minimum BA salary is 0.0000848. This means that for every \$1 increase in the minimum salary paid to any teacher with a Bachelors Degree, a district's average ACT score is expected to increase by 0.0000848 points. The standard error of the coefficient is 0.0000427; the t-stat is 1.99, and the p-value is 0.048. The R-squared of the regression is 0.0119.

* PROBLEM 6 - MODEL 2;

Source	SS	df	MS	Number of obs = 330		
Model	2.53655055	1	2.53655055	F(1, 328)	=	2.49
Residual	334.592106	328	1.02009788	Prob > F	=	0.1158
				R-squared	=	0.0075
				Adj R-squared	=	0.0045
Total	337.128656	329	1.02470716	Root MSE	=	1.01

act	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ppexp	-.0001003	.0000636	-1.58	0.116	-.0002254	.0000248
_cons	22.75342	.4724228	48.16	0.000	21.82406	23.68278

The estimated coefficient on per pupil expenditures is -0.0001003. This means that for every \$1 increase in per pupil expenditures, a district's average ACT score is expected to **decrease** by 0.0001003 points. The standard error of the coefficient is 0.0000636; the *t*-stat is -1.58, and the *p*-value is 0.116. The *R*-squared of the regression is 0.0075.

* PROBLEM 7;

Variable	Obs	Mean	Std. Dev.	Min	Max
bamin	330	25.71998	1.301497	22.293	30.815
bamax	330	36.85116	3.64896	27.572	48.476
mamin	330	28.72404	1.746136	24.393	36.032
mamax	330	44.21998	4.048267	32.975	57.209
ppexp	330	7.377191	.8756122	5.80712	12.53509

* PROBLEM 8 - MODEL 3;

Source	SS	df	MS	Number of obs = 330		
Model	4.00811463	1	4.00811463	F(1, 328)	=	3.95
Residual	333.120542	328	1.01561141	Prob > F	=	0.0478
				R-squared	=	0.0119
				Adj R-squared	=	0.0089
Total	337.128656	329	1.02470716	Root MSE	=	1.0078

act	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
bamin	.0848065	.0426897	1.99	0.048	.0008264	.1687865
_cons	19.83242	1.099378	18.04	0.000	17.66969	21.99514

The estimated coefficient on minimum BA salary is 0.0848065. This means that for every \$1,000 increase in the minimum salary paid to any teacher with a Bachelors Degree, a district's average ACT score is expected to increase by 0.0848 points. The standard error of the coefficient is 0.0426897; the *t*-stat is 1.99, and the *p*-value is 0.048. The *R*-squared of the regression is 0.0119.

The t -stat reports the number of standard deviations our estimate of the coefficient on minimum BA salary is away from the true value of zero *if the true value is zero*. As the two-sided critical value at the 95% level is 1.96 so that the t -stat is greater than this (in absolute value), we would reject the claim that minimum BA salary is unrelated to ACT scores.

The p -value reports the lowest significance level at which one is willing to reject the claim that minimum BA salary is unrelated to ACT scores. In this case, as long as the significance level is 4.8% or greater, one would reject the claim. (Notice, this matches our previous result of rejecting the claim with 5% significance using the t -stat.)

Notice that parts of the regression are identical to MODEL 1. In particular, the statistics that describe how well the independent variable explains or is related to the dependent variable are all the same, i.e., the t -stat, p -value, and R -squared are all unchanged. What has changed is the estimated coefficient on minimum BA salary and its standard error, but notice how they have changed. They have changed in such a way as to keep the interpretation of the coefficient (and the statistical significance of the coefficient) unchanged. That is, there is no difference, in a linear model, between the following two interpretations:

MODEL 1: for every \$1 increase in the minimum salary paid to any teacher with a Bachelors Degree, a district's average ACT score is expected to increase by 0.0000848 points.

MODEL 3: for every \$1,000 increase in the minimum salary paid to any teacher with a Bachelors Degree, a district's average ACT score is expected to increase by 0.0848 points.

The conjecture, which is correct, is that rescaling variables cannot change the interpretation or statistical significance of estimated coefficients. Put differently, the computer doesn't understand units, and so the interpretation of any estimated relationship better be impervious to the unit of measurement. (Finally, notice that the estimated coefficient and standard error in MODEL 3 both increased 1000 times, thus keeping the t -stat unchanged, as the independent variable was divided by 1000.

* PROBLEM 9 - MODEL 4;

Source	SS	df	MS	Number of obs = 330		
Model	6.62562371	1	6.62562371	F(1, 328) =	3.95	
Residual	550.665728	328	1.67885893	Prob > F =	0.0478	
				R-squared =	0.0119	
				Adj R-squared =	0.0089	
Total	557.291352	329	1.69389469	Root MSE =	1.2957	

bamin	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
act	.1401895	.0705682	1.99	0.048	.0013661	.279013
_cons	22.6339	1.5551	14.55	0.000	19.57467	25.69313

The estimate on ACT is positive, but other than being the same sign as in MODEL 4, there is no obvious relationship between the two. Notice, however, that the

statistics from which something about the statistical relationship between the two variables can be inferred have all remained the same. That is, the *t*-stat, *p*-value, and *R*-squared of the regression are identical in MODELS 3 and 4. Again, this is indicative that the computer doesn't "think" about the model or causation. To the computer, all that matters is the strength of the linear relationship between the variables. The lesson, of course, is that we need to be cautious before concluding that one variable is "causing" or "affecting" another. Regression results report the "relationship" between two variables, and that is all.

* PROBLEM 10 - MODEL 5;

Source	SS	df	MS			
Model	2.53655007	1	2.53655007	Number of obs =	330	
Residual	334.592106	328	1.02009788	F(1, 328) =	2.49	
Total	337.128656	329	1.02470716	Prob > F =	0.1158	
				R-squared =	0.0075	
				Adj R-squared =	0.0045	
				Root MSE =	1.01	

act	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ppexp	-.1002795	.0635933	-1.58	0.116	-.2253816	.0248227
_cons	22.75342	.4724228	48.16	0.000	21.82406	23.68278

The estimated coefficient on per pupil expenditures is -0.1002795. This means that for every \$1,000 increase in per pupil expenditures, a district's average ACT score is expected to **decrease** by 0.1003 points. The standard error of the coefficient is 0.0636, the *t*-stat is -1.58, and the *p*-value is 0.116. The *R*-squared of the regression is 0.0075. One would reject the claim that ACT scores are unrelated to expenditures with 11.6% significance or more.

* PROBLEM 11;

Anti-tax organization might interpret the results in MODEL 5 by claiming that increased expenditures on public education are not associated with higher test scores. Thus, localities should spend less (and tax less). Moreover, although there is a positive relationship between salaries and test scores, the relationship is quite small economically. That is, if a district lowered salaries by \$1,000 across the board, the district's average ACT score is expected to fall by less than 0.10 of a point. (Specifically, it is expected to fall by 0.0848 points.) This is not a great cost when a town can save thousands of dollars.

The teachers' union or pro-education advocates could respond in numerous ways. (I will only list a few.) First, the negative relationship between per pupil expenditures and test scores is equally economically small, so let's not focus on the negative result. Second, the regressions consider two unimportant measures of spending - minimum salary for a teacher with a BA and total per pupil expenditures. It would be more appropriate to look at average salary rather than a starting salary. And it would be more appropriate to look at per pupil expenditure on academic instruction, instead of considering a measure of expenditure that includes construction costs, paying janitors, retiring district debt, funding athletics, funding fine arts, etc. Third, and probably most importantly, maximizing average ACT scores should not be the objective of any school district. The main objective is to get all students to function in society (i.e., able to read, write, and do math).

Thus, when the analysis focuses on the wrong objective function, the results are meaningless. Fourth, even if one thinks ACT scores are a valid outcome measure, they are flawed by the fact that not everyone takes the ACT. ACT tests are usually taken only by the college-bound. Thus, it is not all that surprising to find that the college-bound population across school districts is, more or less, performing the same. The results would probably be more striking (and positive) if all students took the ACT.

Finally, as an econometrician, there are many possible violations of the Classical Regression Model. Again, I will only list a few. First, it is a valid argument that the dependent variable (or outcome variable) is a faulty/meaningless one. Second, the right hand side model may be mis-specified. The linearity implied by the model can always be questioned, but this is usually a cheap shot. The better place to question the model is in the assumption that only salary or only per pupil expenditures affects ACT scores. One can well imagine that other factors also affect test scores, such as race and income. The analysis needs to somehow take these other factors into account. Third, one might be concerned about heteroskedasticity. In fact, this is a classic example of possible heteroskedasticity. Each district is associated with an average ACT score. Large districts, however, will have many more students take the ACT. Thus, their average score will be much more precisely estimated than, say, a district with only a handful of students taking the test. The difference in the number of test-takers will result in a different variance in the error term - a much smaller variance for large districts and a much greater variance for smaller districts.

```

#delimit;
set more 1;

log using project3.log, replace;

* PART I;

* Cut and paste excel data into Stata.
  Save the data with the command: save residuals, replace.
  Clear the memory with the command: clear.
  Now you are ready to run the program.;

use residuals.dta;

* PROBLEM 1;
sum x e2 e20 e50;

* PROBLEM 2;
gen y2=12+8*x+e2;
gen y20=12+8*x+e20;
gen y50=12+8*x+e50;
sum y2 y20 y50;

* PROBLEM 3;
reg y2 x;

* PROBLEM 4;
reg y20 x;

* PROBLEM 5;
reg y50 x;

* PROBLEM 7;
reg y2 x;
predict y2hat;
predict error2, resid;
hist error2, bin(50);
sum y2 y2hat error2;

* PROBLEM 8;
reg y20 x;
predict y20hat;
predict error20, resid;
hist error20, bin(50);
sum y20 y20hat error20;

* PROBLEM 9;
reg y50 x;
predict y50hat;
predict error50, resid;
hist error50, bin(50);
sum y50 y50hat error50;

```

```

* PROBLEM 10;
hist error2, bin(50) xlabel(-100 -50 0 50 100) title("Residuals of Y2 Regression")
saving(graph_e2, replace);
hist error20, bin(50) xlabel(-100 -50 0 50 100) title("Residuals of Y20 Regression")
saving(graph_e20, replace);
hist error50, bin(50) xlabel(-100 -50 0 50 100) title("Residuals of Y50 Regression")
saving(graph_e50, replace);
graph combine graph_e2.gph graph_e20.gph graph_e50.gph, saving(graph_all, replace);

* PROBLEM 11;
scatter y50 x || lfit y50 x;

clear;

* PART II;

use coefficients.dta;

describe;
sum;

* PROBLEM 1;
_pctile enroll, p(20 80);
gen size=1*(enroll<r(r1))
      +2*(enroll>=r(r1)&enroll<=r(r2))
      +3*(enroll>r(r2));
label define sizes 1 Small 2 Middle 3 Large;
label values size sizes;
tab size;
sort size;
by size: sum enroll dropout lunch act;

* PROBLEM 2;
gen ppexp=totalexpend/enroll;
label variable ppexp "District Per Pupil Expenditures";
sum bamin bamax mamin mamax ppexp;

* PROBLEM 3;
scatter act ppexp, xlabel(5000 7500 10000 12500 15000)
      ylabel(16 18 20 22 24 26)
      title("ACT Scores by Per Pupil Expenditures")
      saving(graph_q3a, replace);
scatter act ppexp || lfit act ppexp,
      xlabel(5000 7500 10000 12500 15000)
      ylabel(16 18 20 22 24 26)
      title("ACT Scores by Per Pupil Expenditures")
      saving(graph_q3b, replace);

* PROBLEM 4;
scatter act mamax || lfit act mamax,
      xlabel(30000 40000 50000 60000)
      ylabel(16 18 20 22 24 26)
      title("ACT Scores by Maximum Salary to MA Teacher")
      saving(graph_q5, replace);

```

```
* PROBLEM 5;
reg act bamin;

* PROBLEM 6;
reg act ppexp;

* PROBLEM 7;
replace bamin=bamin/1000;
replace bamax=bamax/1000;
replace mamin=mamin/1000;
replace mamax=mamax/1000;
replace ppexp=ppexp/1000;
sum bamin bamax mamin mamax ppexp;

* PROBLEM 8;
reg act bamin;

* PROBLEM 9;
reg bamin act;

* PROBLEM 10;
reg act ppexp;

save coefficients2, replace;

clear;
log close;
```