

**Problem Set 1**  
ANSWERS

*Part I. Multiple Choice Problems*

1. If  $\mathbf{X}$  and  $\mathbf{Z}$  are two random variables, then  $E[\mathbf{X}-\mathbf{Z}]$  is

**d.  $E[\mathbf{X}] - E[\mathbf{Z}]$**

This is just a simple application of one of the properties of expected value.

2. Any linear combination of normally distributed random variables:

**a. is also normally distributed**

This is also just a simple application of probability theory.

(These first 2 were just checking that you reviewed the Econ 10 material!)

3. What can be said about the estimated slope coefficient for a regression of  $Y$  on  $X$ , versus the estimated slope coefficient for a regression of  $X$  on  $Y$ ?

**b. the slopes are not reciprocals**

The estimate of  $\beta_1$  from the population model  $Y=\beta_0+\beta_1X+u$  can be written as  $Cov(X,Y)/Var(X)$ , thus the estimate of  $\beta_1$  from the population model  $X=\beta_0+\beta_1Y+u$  can be written as  $Cov(X,Y)/Var(Y)$ . Clearly, these will have the same sign, will not be identical, and will not be reciprocals.

4. Suppose the following is the true population model of the effect of smoking on birth weight in ounces:  $birth\ weight = \beta_0 + \beta_1 daily\ cigarettes + u$ , and that the average birth weight is 125 oz and the average mother smoked 2 cigarettes per day. If your estimate of  $\beta_1$  is -2.5, what must your estimate of  $\beta_0$  be?

**c. 130**

We know that an OLS regression fits exactly at the mean of the sample. Thus, based on the information given, we have:  $125 = \text{estimated } \beta_0 - 2.5 * 2$ , so our estimate of  $\beta_0$  must be  $125+5=130$ .

5. Suppose a new sample is chosen and you estimate the following:  $predicted\ birth\ weight\ in\ oz = 136 - 2.1\ daily\ cigarettes$ . What would be your estimate of  $\beta_0$  be if birth weight were recorded in pounds?

**b. 8.5**

There are 16 ounces in a pound, so we need to divide our coefficients by 16 to obtain the same relationship.

6. Using the same sample as in question 5, you now estimate the following:  $birth\ weight\ in\ oz = \beta_0 + \beta_1 weekly\ cigarettes + u$ . Your estimate of  $\beta_1$  would be:

**c. - 0.3**

From question 5 we know that each daily cigarette reduces birth weight by 2.1 ounces. Since a daily cigarette implies 7 weekly cigarettes, 1 weekly cigarette has an effect 1/7 the size, reducing weight by 0.3 ounces.

Part II. Stata Problems

1. a)

```
. sum score89 pcy, detail
```

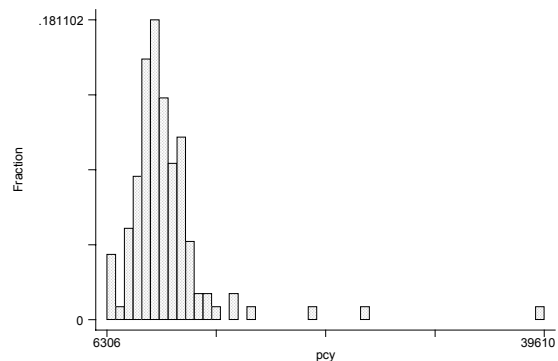
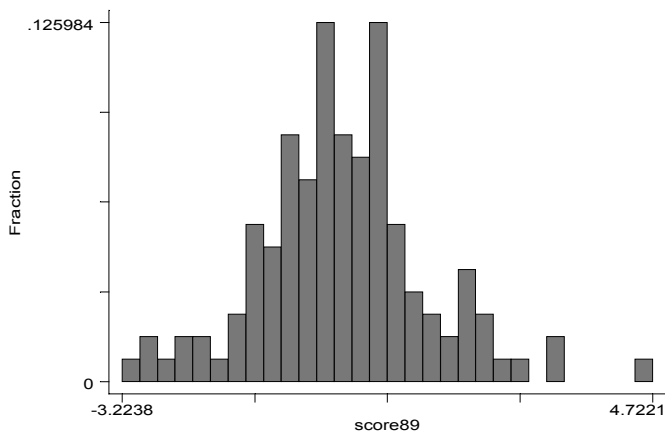
score89					
Percentiles		Smallest			
1%	-2.804	-3.2238			
5%	-1.9984	-2.804			
10%	-1.2417	-2.7727	Obs		127
25%	-.69525	-2.5743	Sum of Wgt.		127
			Mean		.0839172
50%	.016168		Std. Dev.		1.266603
75%	.72278	Largest			
		2.669	Variance		1.604283
90%	1.8124	3.1895	Skewness		.3415457
95%	2.2796	3.2187	Kurtosis		4.177176
99%	3.2187	4.7221			

pcy					
Percentiles		Smallest			
1%	6552	6306			
5%	7844.8	6552			
10%	8164	6848	Obs		127
25%	9234	6879	Sum of Wgt.		127
			Mean		10790.68
50%	10127		Std. Dev.		3613.81
75%	11542	Largest			
		17582	Variance		1.31e
90%	12806	21961	Skewness		4.879524
95%	14374	25940	Kurtosis		35.87324
99%	25940	39610			

Given the above, the mean of the test score is 0.084, and the median is 0.016 – note the 50<sup>th</sup> percentile is the median! The mean of per capita income is \$10,791, and the median is \$10,127.

b) I chose bin sizes of 30 and 50, respectively. You may have made a different choice.

```
. graph score89, bin(30)
. graph pcy, bin(50)
```



c) From part a) we know that median per capita income is 10127, so

```
. sum score89 pcy if pcy >= 10127
```

Variable	Obs	Mean	Std. Dev.	Min	Max
score89	64	.6253451	1.216266	-1.5618	4.7221
pcy	64	12585.69	4317.745	10127	39610

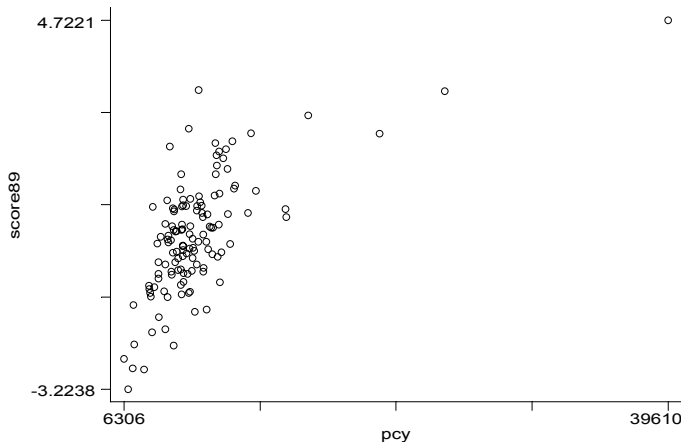
```
. sum score89 pcy if pcy < 10127
```

Variable	Obs	Mean	Std. Dev.	Min	Max
score89	63	-.4661048	1.071052	-3.2238	2.0024
pcy	63	8967.171	944.517	6306	10110

The average test score in districts with above median per capita income is 0.625 and in districts with below median per capita income it is -0.466. This is a difference of 1.091 standard deviation units. The average per capita income in the high-income districts is \$12,586, while in the low-income districts it is \$8,967. This is a difference of \$3,619.

d) You may have graphed this the other way. While it doesn't really matter, it is more natural to think of the test score as the outcome, and thus as the Y variable.

```
. graph score89 pcy
```



e)

```
. reg score89 pcy
```

Source	SS	df	MS	Number of obs = 127		
Model	84.8861871	1	84.8861871	F( 1, 125)	=	90.49
Residual	117.253472	125	.938027778	Prob > F	=	0.0000
Total	202.139659	126	1.60428301	R-squared	=	0.4199
				Adj R-squared	=	0.4153
				Root MSE	=	.96852

score89	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pcy	.0002271	.0000239	9.513	0.000	.0001799	.0002744
_cons	-2.366932	.2715919	-8.715	0.000	-2.904446	-1.829418

The above regression shows that there is a positive relationship between per capita income and test scores. An extra \$1000 in per capita income in a district will increase test scores in the district by 0.227 standard deviation units. This should not surprise us, since the graph in part d) appeared to reflect a positive relationship. Also, in part c) we saw a difference of \$3619 between high- and low-income districts, and a difference of 1.09 in test scores. Our regression would predict a difference of about 0.822 in test scores for a \$3619 difference in scores. This is in the same ballpark. Of course, this simple regression does not control for the many other things that will affect test scores that may be correlated with per capita income.

2. a)

. reg TDs attempts

Source	SS	df	MS	Number of obs = 60		
Model	2634.62634	1	2634.62634	F( 1, 58)	=	247.12
Residual	618.356991	58	10.6613274	Prob > F	=	0.0000
-----				R-squared	=	0.8099
Total	3252.98333	59	55.1353107	Adj R-squared	=	0.8066
-----				Root MSE	=	3.2652
TDs	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
attempts	.0373786	.0023778	15.72	0.000	.0326189	.0421382
_cons	-.2807747	.6655973	-0.42	0.675	-1.613112	1.051563

Based on this regression, 100 more attempts would imply 3.7 more touchdowns in the season.

b)

. reg TDs attempts completions

Source	SS	df	MS	Number of obs = 60		
Model	2776.59711	2	1388.29856	F( 2, 57)	=	166.11
Residual	476.386219	57	8.35765297	Prob > F	=	0.0000
-----				R-squared	=	0.8536
Total	3252.98333	59	55.1353107	Adj R-squared	=	0.8484
-----				Root MSE	=	2.891
TDs	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
attempts	-.0314791	.016839	-1.87	0.067	-.0651986	.0022404
completions	.1160537	.028158	4.12	0.000	.0596683	.172439
_cons	.0249384	.5939655	0.04	0.967	-1.164457	1.214334

Now, all else equal, another 100 attempts would imply 3 fewer touchdowns! What is going on? The key here is the ceteris paribus interpretation of the coefficient on attempts. What does it mean to increase attempts, holding completions constant? It means the QB is throwing a bunch of incomplete passes. Presumably, QBs with more incompletions are worse QBs, so they have fewer touchdowns.

c) At first glance, we might just say that with 100 more completions, 11.6 more touchdowns will be thrown. This is true, all else equal – i.e. holding attempts constant. So, it's like comparing 2 QBs that throw the same number of passes, one of whom has 100 more completions. Thus, it's the answer to what happens if you increase the completion *rate* for a QB. Another interpretation might be to think about just adding 100 completions to what was already being done. In this case, since every completion is also an attempt, we would gain 11.6 touchdowns, but lose 3.1 from the additional attempts. Overall, then, we would expect 8.5 more touchdowns after this change (adding 100 attempts that were all completions).

