Universidade de São Paulo / Faculdade de Filosofia, Letras e Ciências Humanas
Departamento de Ciência Política
FLP-0468 & FLS-6183
2º semester / 2018

Problem Set# 3

Answer Key

Please submit a write-up and do file at the beginning of class on October 4, 2018. Please hand in the problem set printed and only send the do-file.

1. For the first exercise, we will return to the first and second cases we analyzed last week. In case 1, we will suppose X and Z are not correlated. In case 2, we will suppose that X and Z are correlated such that corr(X,Z)=0.75 as are the variables that are generated in the simulation in part b of the do file such that:

$$Y \sim (100, 20)$$
$$X \sim (7, 8)$$
$$Z \sim (20, 2)$$
$$Corr(Y, X, Z) = \begin{pmatrix} 1 & 0.7 & 0.3 \\ 0.7 & 1 & 0.75 \\ 0.3 & 0.75 & 1 \end{pmatrix}$$

Based on the formulas for Omitted Variable Bias (OMB) suggested in Mastering Metrics (MM) calculate the OMB when Z is omitted in Case 1 and Case 2. Please make sure to try to replicate the exercise and use the formulas as outlined in MM.

MM propose a formula for comparing omitted variable bias such that:

$$y = \alpha + \beta_1^{short} x + u$$
$$y = \alpha + \beta_1^{long} x + \beta_2 z + u$$

$$OVB = \beta^{short} - \beta^{long} .$$

Case 1: x and z are not correlated

```
.
------------------------------------------------------------------------------
                              (1)                   (2)
                              y                     y
------------------------------------------------------------------------------
x                         1.828***              1.864***
                      [1.675,1.982]         [1.724,2.004]

z                                               2.937***
                                            [2.373,3.501]

_cons                     88.00***              28.28***
                      [86.33,89.68]         [16.72,39.85]
------------------------------------------------------------------------------
N                             500                   500
------------------------------------------------------------------------------
95% confidence intervals in brackets
* p<0.05, ** p<0.01, *** p<0.001
```

In a three variable case:

Long: $y=\alpha+\beta_1^{long}x+\beta_2 z+u$

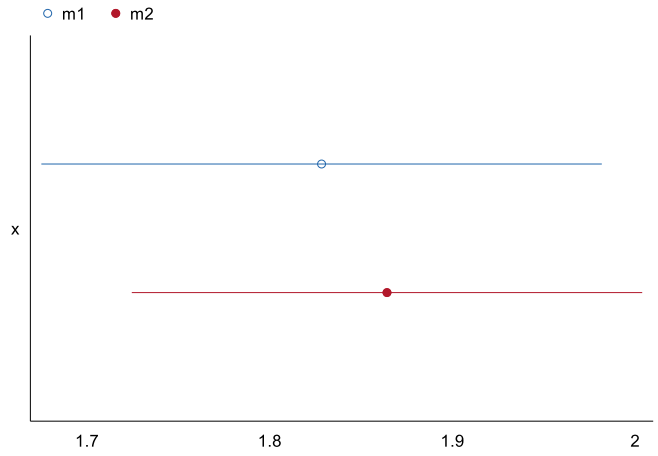
Short: $y=\beta_0+\beta_1^{short}x+e$

$z=\tau+\gamma x+\upsilon$

$OVB=\beta_1^{short}-\beta_1^{long}=\beta_2\times\gamma$

$OVB=1.828 - 1.864= - 0.036$

Note: This is an insightful formula, but in addition to the point estimates, we also need to think about the standard errors and confidence intervals. As the plot below suggests, the effect of x on y is slightly underestimated in the short form. However, it seems that the difference is not statistically significant from 0 as both confidence intervals overlap.



Case 2: x and z are correlated

```
. esttab m3 m4, se
```

	(1) y	(2) y
x	1.828***	2.695***
	(0.0781)	(0.101)
z		-5.000***
		(0.423)
_cons	88.00***	182.4***
	(0.852)	(8.028)
N	500	500

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

In a three variable case:

Long: $y=\alpha+\beta_1^{long}x+\beta_2 z+u$

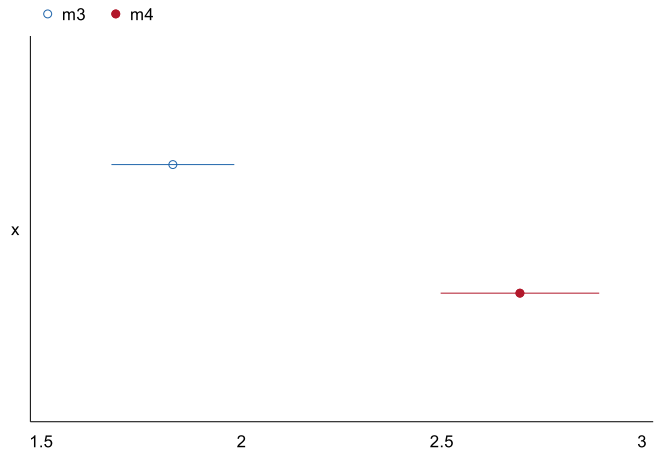
Short: $y=\beta_0+\beta_1^{short}x+e$

$z=\tau+\gamma x+\upsilon$

$OVB=\beta_1^{short}-\beta_1^{long}=\beta_2\times\gamma$

$OVB=1.828-2.695=-0.867$

As the plot below suggests, the effect of x on y is underestimated in the short form to a greater extent to Case 1. Moreover, it seems that the difference is statistically significant from 0 as the confidence intervals do not overlap.



When two CIs do NOT overlap: The two groups are significantly different.

When two CIs DO overlap: We do not know what the conclusion is – but we could, e.g., make a CI for the mean difference instead, to investigate. How do we do so? Let us use case 1.

Gelman and Stern (2007) propose we calculate z=b1-b2/sqrt(se1^2+se2^2) to compare betas, but this test is usually only considered appropriate where we are comparing coefficients from *independent samples*.

As in this case, we are comparing coefficient estimates from the same sample. The correct formula for the variance of (b1(long)-b1(short)) as summarized by Clogg, Petkova and Haritou (1995) is: $Var(a-b) = Var(a) + Var(b) - 2Cov(a,b)$. We have thus far estimated Var(b1(long)) and Var(b2(short)), but we have not estimate (Cov b1(short),b1(long).

Clogg, Petkova and Haritou (1995) propose calculating Clogg, Petkova and Haritou (1995) is:

$$Var(\beta_1^{short} - \beta_1^{long}) = V(\beta_1^{long}) - Var(\beta_1^{short})(\sigma_{long}^2 / \sigma_{short}^2)$$

$$se(\beta_1^{short} - \beta_1^{long}) = .00349861$$

Then, the t-statistics is $t = \frac{-0.03580677}{.00349861} = -\ 10.23457$

Note that this t-stat is very similar to the t-stat in the long-form as Clogg, Petkova and Haritou (1995) explain should be expected.

In this case, the difference is negative and statistically different from zero. We can see this by also calculating the 95% confidence interval.

$$\text{lower limit of diff} = -0.03580677 - .00349861*1.96 = -0.04266408$$

$$\text{upper limit of diff} = -0.03580677 + .00349861*1.96 = -0.02894952$$

Thus, we reject the null that the difference in the betas is equal to 0 with 99% confidence. The test statistic is telling us that the difference is statistically different from 0. Even though the correlation between X and Z is low, the test that both betas are the same (the null hypothesis is that the betas from both regressions are the same) is statistically significantly different from 0.

We need to be careful to not be fooled by overlapping confidence intervals and make sure to calculate correct comparisons as these samples *are not independent*. This is because as Cummins (2009) notes "If two means are in some other way correlated, the two CIs may not be used to assess the difference, because they do not reflect the correlation. For pre-test and post-test means, for example, the CI on the paired differences is needed [5]."

In the do file, I also show how to do this with a Hausman test.

An additional way to address this issue would be with bootstrapping. We could use a bootstrapping technique to run each of the two models multiple times, save the beta coefficients as cases and then run a t-test to determine whether the betas from each model are significantly different.

2. Now, let's build on the examples outlined in MM Chapter 2. For a sub-sample that is smaller than the entire number of cases (n<500), let's assume we have a "within group comparison" such as the case suggested in Table 2.2 of MM. Please run the regressions and report them in a separate table (e.g. one table for case 1 and one table for case 2). How do these models compare to the models estimated in 1? (Hint: The first step is to create some type of variable that allows you to select a subsample and within that subsample to have sub-groups analogous to the example discussed in the chapter).

Case 1

See do file for assumptions used to estimate groups. See discussion in Question 1 to understand how to compare models.

```
. esttab m1 m2 m5, ci
```

	(1) y	(2) y	(3) y
x	1.828***	1.864***	1.893***
	[1.675,1.982]	[1.724,2.004]	[1.677,2.109]
z		2.937***	
		[2.373,3.501]	
group2			2.572
			[-4.260,9.403]
group3			8.032*
			[1.084,14.98]
group4			2.552
			[-4.261,9.364]
group5			4.467
			[-2.842,11.78]
group6			11.40**
			[4.359,18.44]
group7			8.463*
			[1.154,15.77]
group8			15.14***
			[7.927,22.36]
group9			16.33***
			[8.916,23.74]
group10			20.59***
			[13.54,27.63]
_cons	88.00***	28.28***	78.38***
	[86.33,89.68]	[16.72,39.85]	[73.40,83.36]
N	500	500	242

95% confidence intervals in brackets
* p<0.05, ** p<0.01, *** p<0.001

.

Case 2
See discussion in Question 1 to understand how to compare models.

```
. esttab m3 m4 m6, ci
```

	(1) y	(2) y	(3) y
x	1.828***	2.695***	2.608***
	[1.675,1.982]	[2.497,2.893]	[2.314,2.902]
z		-5.000***	
		[-5.832,-4.168]	
group2			-3.154
			[-10.77,4.466]
group3			-3.884
			[-11.55,3.778]
group4			-8.146*
			[-16.19,-0.0982]
group5			-5.334
			[-13.55,2.884]
group6			-12.13**
			[-20.36,-3.889]
group7			-15.72***
			[-24.28,-7.158]
group8			-18.76***
			[-27.70,-9.813]
group9			-20.84***
			[-29.86,-11.82]
group10			-29.66***
			[-39.34,-19.98]
_cons	88.00***	182.4***	93.22***
	[86.33,89.68]	[166.6,198.2]	[87.48,98.95]
N	500	500	242

95% confidence intervals in brackets
* p<0.05, ** p<0.01, *** p<0.001

.

3. Now, let's continue to build on the example outlined in MM Chapter 2. To do so, let's introduce three group dummies in such a manner that the group dummies are correlated with Z (our treatment effect) in the entire sample (n=500). Please estimate a table similar to the "self-revelation" model reported in Table 2.3 (one table for case 1 and one table for case 2). Can you re-run the calculations above and observe what the OVB is from excluding Z? (Hint: The first step is to create some type of variable to identify the different groups).

Case 1

```
. esttab m1 m2 m5 m7 m8, ci
```

	(1) y	(2) y	(3) y	(4) y	(5) y
x	1.828***	1.864***	1.893***	1.868***	1.866***
	[1.675,1.982]	[1.724,2.004]	[1.677,2.109]	[1.726,2.010]	[1.726,2.006]
z		2.937***			2.538***
		[2.373,3.501]			[1.315,3.761]
group2			2.572	5.237***	-0.310
			[-4.260,9.403]	[2.375,8.099]	[-4.194,3.574]
group3			8.032*	13.53***	2.239
			[1.084,14.98]	[10.66,16.40]	[-3.891,8.369]
group4			2.552		
			[-4.261,9.364]		
group5			4.467		
			[-2.842,11.78]		
group6			11.40**		
			[4.359,18.44]		
group7			8.463*		
			[1.154,15.77]		
group8			15.14***		
			[7.927,22.36]		
group9			16.33***		
			[8.916,23.74]		
group10			20.59***		
			[13.54,27.63]		
_cons	88.00***	28.28***	78.38***	81.47***	35.70**
	[86.33,89.68]	[16.72,39.85]	[73.40,83.36]	[79.17,83.78]	[13.53,57.87]
N	500	500	242	500	500

95% confidence intervals in brackets
* p<0.05, ** p<0.01, *** p<0.001

Case 2

```
.
. esttab m3 m4 m6 m9 m10, ci
```

	(1) y	(2) y	(3) y	(4) y	(5) y
x	1.828***	2.695***	2.608***	2.402***	2.686***
	[1.675,1.982]	[2.497,2.893]	[2.314,2.902]	[2.210,2.593]	[2.487,2.884]
z		-5.000***			-5.232***
		[-5.832,-4.168]			[-6.667,-3.797]
group2			-3.154	-7.106***	2.249
			[-10.77,4.466]	[-10.32,-3.892]	[-1.747,6.244]
group3			-3.884	-17.51***	1.367
			[-11.55,3.778]	[-21.38,-13.64]	[-4.988,7.722]
group4			-8.146*		
			[-16.19,-0.0982]		
group5			-5.334		
			[-13.55,2.884]		
group6			-12.13**		
			[-20.36,-3.889]		
group7			-15.72***		
			[-24.28,-7.158]		
group8			-18.76***		
			[-27.70,-9.813]		
group9			-20.84***		
			[-29.86,-11.82]		
group10			-29.66***		
			[-39.34,-19.98]		
_cons	88.00***	182.4***	93.22***	92.08***	185.9***
	[86.33,89.68]	[166.6,198.2]	[87.48,98.95]	[90.05,94.12]	[160.1,211.7]
N	500	500	242	500	500

95% confidence intervals in brackets
* p<0.05, ** p<0.01, *** p<0.001

4. Table 1.1 in MM compares the health and demographic characteristics of insured and uninsured couples in the NHIS (NHIS2009_clean.dta). Execute the Stata code in NHIS2009_hicompare.do through line 35 to make sure that you use the same selection criteria that were used to produce Table 1.1.

The sample we are analyzing is married husbands who are at least 26 and not older than 59 where at least one spouse is working. We are removing single people households (e.g. adltempl>=1). By health insurance status, we can see that 87.66% of husbands have health insurance.

```
. * count of husbands by HI status
.         tab hi if  fml==0 [ aw=perweight ]

          hi |      Freq.     Percent        Cum.
-------------+-----------------------------------
           0 | 1,833.9341       12.34       12.34
           1 | 13,033.066       87.66      100.00
-------------+-----------------------------------
       Total |     14,867      100.00

.
```

V1	V2	V3	V4	V5	V6	V7	V8	NOTES_TITLES
VARIABLES	(1) hlth	(2) nwhite	(3) age	(4) yedu	(5) famsize	(6) empl	(7) inc	Robust standard errors in parentheses
HI	0.08	-0.04	10.18	2.38	-0.72	-0.03	48,967.13	
	(0.03)	(0.01)	(0.33)	(0.09)	(0.04)	(0.01)	(1,117.84)	
CONSTANT	3.72	0.18	41.23	11.61	3.83	0.75	43,727.74	
	(0.03)	(0.01)	(0.29)	(0.08)	(0.04)	(0.01)	(952.41)	
OBSERVATIONS	14,867	14,867	14,867	14,867	14,867	14,867	14,867	
R-SQUARED	0.00	0.00	0.05	0.07	0.03	0.00	0.08	
MEAN NO HI	3.7	.18	41	12	3.8	.75	43728	
SD NO HI	1	.38	11	3.3	1.6	.44	36386	
MEAN HI	3.8	.14	51	14	3.1	.71	92695	
SD HI	1.1	.35	15	2.8	1.3	.45	56365	

a. Panel A compares the health across husbands in this sample with and without health insurance. Use the sum command to calculate average health separately for husbands with and without health insurance. What is the difference in average health by insurance status? Is this difference statistically significant at the 5% level? Construct a 95% confidence interval for the difference.

There are two formulas we can use to calculate this test statistic depending on our assumption of whether the variance is equal (the same) in both groups, or unequal.

If it reasonable to assume that two groups have the same standard deviation, then the test statistic can be calculated as follows (see FPSR p. 160):

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{se(\bar{Y}_1 - \bar{Y}_2)}$$

$$se(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \times \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\text{degrees of freedom} = n_1 + n_2 - 2$$

However, if we assume that the two groups do not have the same variance, then the t-statistic would be:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{se(\bar{Y}_1 - \bar{Y}_2)}$$

$$se(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$

$$\text{degrees of freedom} = \textbf{using the } t(k) \textbf{ distribution where } k \textbf{ represents the } \textit{smaller} \textbf{ of } n_1 - 1 \textbf{ and } n_2 - 1.$$

```
. sum hlth if fml==0 & hi==1 [ w=perweight ]
(analytic weights assumed)

    Variable |     Obs      Weight      Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------------
        hlth |   7,866    29464737  4.008899    .928802         1          5

. sum hlth if fml==0 & hi==0 [ w=perweight ]
(analytic weights assumed)

    Variable |     Obs      Weight      Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------------
        hlth |   1,529     4653826  3.695654   1.012124         1          5
```

Under either the assumption of unequal or equal variances, there is a statistically significant difference in the health status of insured husbands as compared to uninsured health husbands. Under the equal variance assumption, the 95% confidence interval is that this difference ranges between 0.12 to 0.02 lower for insured husbands.

```
. ttesti `N1' `av1' `sd1' `N2' `av2' `sd2' , level (95)

Two-sample t test with equal variances
```

	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
x	7,866	4.008899	.0104724	.928802	3.988371	4.029428
y	1,529	3.695654	.0258839	1.012124	3.644883	3.746426
combined	9,395	3.95792	.0097998	.9498727	3.93871	3.97713
diff		.3132452	.026352		.2615895	.3649009

```
    diff = mean(x) - mean(y)                                      t =  11.8870
Ho: diff = 0                                     degrees of freedom =     9393

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000


.
. ttesti `N1' `av1' `sd1' `N2' `av2' `sd2' , unequal

Two-sample t test with unequal variances
```

	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
x	7,866	4.008899	.0104724	.928802	3.988371	4.029428
y	1,529	3.695654	.0258839	1.012124	3.644883	3.746426
combined	9,395	3.95792	.0097998	.9498727	3.93871	3.97713
diff		.3132452	.0279222		.2584865	.3680039

```
    diff = mean(x) - mean(y)                                      t =  11.2185
Ho: diff = 0                     Satterthwaite's degrees of freedom =  2058.48

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

b. Panel B of Table 1.1 shows that husbands with and without health insurance differ along many demographic dimensions. It is possible that the difference in health between the "Some HI" and "No HI" groups may be smaller if we compare across groups that are more homogeneous. To investigate this, please test if the difference between the health of husbands with Some and No HI significantly different from zero if you restrict to men who:

    i. are employed?

```
. ttesti `N1' `av1' `sd1' `N2' `av2' `sd2' , level (95)

Two-sample t test with equal variances
```

	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
x	7,255	4.06562	.0102976	.8771135	4.045434	4.085807
y	1,303	3.756331	.026857	.9694604	3.703644	3.809019
combined	8,558	4.018529	.0097139	.8986285	3.999488	4.037571
diff		.3092889	.0268321		.2566915	.3618863

```
    diff = mean(x) - mean(y)                                      t =  11.5268
Ho: diff = 0                                     degrees of freedom =     8556

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

ii. are employed and have at least 12 years of education?

```
. ttesti `N1' `av1' `sd1' `N2' `av2' `sd2' , level (95)

Two-sample t test with equal variances
```

	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
x	6,804	4.087774	.0104995	.866062	4.067192	4.108356
y	820	3.814704	.0335701	.9613001	3.74881	3.880597
combined	7,624	4.058404	.0100876	.8808072	4.038629	4.078178
diff		.27307	.0324115		.2095346	.3366054

```
    diff = mean(x) - mean(y)                                      t =   8.4251
Ho: diff = 0                                     degrees of freedom =     7622

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

iii. are employed, have at least 12 years of education, and earn income of at least $80,000?

```
.
. ttesti `N1' `av1' `sd1' `N2' `av2' `sd2' , level (95)

Two-sample t test with equal variances
```

	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
x	4,480	4.182719	.0121169	.8110198	4.158964	4.206474
y	121	4.055813	.0798512	.8783636	3.897714	4.213913
combined	4,601	4.179381	.0119859	.8130132	4.155883	4.20288
diff		.1269055	.0748865		-.019908	.2737191

```
    diff = mean(x) - mean(y)                                      t =   1.6946
Ho: diff = 0                                     degrees of freedom =     4599

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9549         Pr(|T| > |t|) = 0.0902          Pr(T > t) = 0.0451
```

c. Use the NHIS data to construct a variable such that a regression of health on this variable reproduces the difference calculated in question (a), above. Compare the difference, t-statistic, and confidence interval for your regression estimate of differences in health with those you computed in (a).

The difference in means is the same, but the t-statistic and confidence intervals are slightly different.

```
. reg hlth hi if fml==0 [ w=perweight ], robust
(analytic weights assumed)
(sum of wgt is 34,118,563)

Linear regression                               Number of obs     =      9,395
                                                F(1, 9393)        =      84.68
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0129
                                                Root MSE          =      .9406

------------------------------------------------------------------------------
             |               Robust
        hlth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          hi |   .3132452   .0340396     9.20   0.000     .2465202    .3799702
       _cons |   3.695654   .0316859   116.63   0.000     3.633543    3.757765
------------------------------------------------------------------------------

. ttesti `N1' `av1' `sd1' `N2' `av2' `sd2' , level (95)

Two-sample t test with equal variances

------------------------------------------------------------------------------
             |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   7,866    4.008899   .0104724    .928802    3.988371    4.029428
           y |   1,529    3.695654   .0258839    1.012124   3.644883    3.746426
-------------+----------------------------------------------------------------
    combined |   9,395     3.95792   .0097998    .9498727    3.93871     3.97713
-------------+----------------------------------------------------------------
        diff |            .3132452    .026352                .2615895    .3649009
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =  11.8870
Ho: diff = 0                                     degrees of freedom =     9393

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000

.
. ttesti `N1' `av1' `sd1' `N2' `av2' `sd2' , unequal

Two-sample t test with unequal variances

------------------------------------------------------------------------------
             |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   7,866    4.008899   .0104724    .928802    3.988371    4.029428
           y |   1,529    3.695654   .0258839    1.012124   3.644883    3.746426
-------------+----------------------------------------------------------------
    combined |   9,395     3.95792   .0097998    .9498727    3.93871     3.97713
-------------+----------------------------------------------------------------
        diff |            .3132452   .0279222                .2584865    .3680039
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =  11.2185
Ho: diff = 0                     Satterthwaite's degrees of freedom =  2058.48

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

d. In (b) of this Problem Set, we showed that some of the difference in average health between those with and without health insurance in the NHIS can be attributed to the fact that the insured differ from the uninsured along many relevant dimensions. We can also show this using regressions. Starting with your regression from part d above, sequentially add controls for age (age), years of education (yedu), and income (inc). Does any set of controls eliminate the difference in health between insured and uninsured? Explain how the results change as you add controls and what changes in the estimates as you add more controls might mean.

Table 1

	(1)		(2)		(3)		(4)	
Health insurance	0.313	***	0.366	***	0.145	***	0.019	
	(0.03)		(0.03)		(0.04)		(0.04)	
Age			-0.019	***	-0.019	***	-0.021	***
			(0.00)		(0.00)		(0.00)	
Years of Education					0.080	***	0.056	***
					(0.00)		(0.00)	
Income							0.000	***
							(0.00)	
Constant	3.696	***	4.495	***	3.553	***	3.772	***
	(0.03)		(0.06)		(0.08)		(0.08)	
R squared	0.0129		0.0443		0.0942		0.1212	
No. of obs.	9395		9395		9395		9395	
Root Mean Square Error	0.9406		0.9256		0.9011		0.8876	

Robust standard errors in parenthesis.
* p<0.10, ** p<0.05, *** p<0.01

As we add additional controls, we see that the observed differences in health status are less than appeared initially. In the final model that is presented, the coefficient on health insurance (the "treatment" effect) is no longer statistically different from zero.

5. Regression application: The effects of class size. The Angrist data archive (http://econ-www.mit.edu/faculty/angrist/data1) contains data from the following article (posted on Stellar): J. Angrist and V. Lavy, "Using Maimonides Rule to estimate the Effect of Class Size on Student Achievement," The Quarterly Journal of Economics, May 1999. This article uses the fact that Israeli class size is capped at 40 to estimate the effects of class size on test scores with an Instrumental Variables / Regression Discontinuity research design. But for now, we'll use the data to explore regression basics.

   (a) Read the article through Section I (at least), download the data, and construct the descriptive stats in Table 1 for 5th graders. From here you should be able to mostly tell what's what as far as variable names go (note that the unit of observation is the class average). Note that enrollment is called c_size and percent disadvantaged in called tipuach. To exactly reproduce the numbers in Table 1, you must follow footnote 11 and restrict the sample to schools with enrollment of at least 5 and classes of size less than 45. There are also a couple of non-obvious data corrections. There is an average math

(avgmath) score and an average verbal (avgverb) score greater than 100 due to a data entry error. The correct values of these scores are 87.606 and 81.246 (not 187.606 and 181.246). Finally, there is a non-missing math score for an observation with mathsize==0 (i.e. no math test takers). This is impossible. Replace avgmath=. if mathsize==0.

```
. tabstat classize c_size tipuach verbsize mathsize avgverb avgmath , stat(mean sd p10 p25 p50 p75 p90) long col(stat) save
```

variable	mean	sd	p10	p25	p50	p75	p90
classize	29.95416	6.598022	21	26	31	35	38
c_size	77.86841	39.06074	31	50	72	100	129
tipuach	14.09118	13.48489	2	4	10	19	35
verbsize	27.32806	6.61587	19	23	28	32	36
mathsize	27.72381	6.675614	19	23	28	33	36
avgverb	74.39608	7.680949	64.16267	69.855	75.43	79.848	83.34029
avgmath	67.30686	9.598965	54.86546	61.15	67.82	74.11	79.40897

(b) Economists and educators have long debated whether it's worth paying the extra labor costs (i.e., teachers' wages) required to reduce class size. What should the sign of the achievement/class-size relationship be if the investment is worthwhile?

The sign should be negative.

(c) Regress average math and verbal scores on class size. What is the sign of this relationship? Is it significantly different from zero? How does it look so far for the class size optimists?

```
.  regress avgverb classize
```

Source	SS	df	MS		
				Number of obs	= 409
				F(1, 407)	= 2.69
Model	244.950729	1	244.950729	Prob > F	= 0.1016
Residual	37024.626	407	90.9695971	R-squared	= 0.0066
				Adj R-squared	= 0.0041
Total	37269.5767	408	91.3470018	Root MSE	= 9.5378

avgverb	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
classize	.1122432	.068402	1.64	0.102	-.0222221	.2467086
_cons	69.86754	1.671272	41.81	0.000	66.58214	73.15295

```
.  regress avgmath classize
```

Source	SS	df	MS		
				Number of obs	= 408
				F(1, 406)	= 4.30
Model	491.3761	1	491.3761	Prob > F	= 0.0388
Residual	46408.6773	406	114.307087	R-squared	= 0.0105
				Adj R-squared	= 0.0080
Total	46900.0534	407	115.233546	Root MSE	= 10.691

avgmath	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
classize	.1590961	.0767342	2.07	0.039	.0082501	.309942
_cons	59.86392	1.875903	31.91	0.000	56.17622	63.55161

.

Based on these models, it does not very good for those who advocate that smaller class sizes have an influence on math and verbal learning. In the case of the first model, class size does not have a statistically significant impact effect on the average verbal score based on 90% confidence intervals. In the case of the second model, class size has a statistically significant impact effect on the average math score based on 95% confidence intervals however the effect appears to be substantively minimal if we examine the distribution of average math scores.

```
. sum avgmath
```

Variable	Obs	Mean	Std. Dev.	Min	Max
avgmath	408	63.59526	10.73469	27.69	93.27