

**Problem Set 4**  
**ANSWERS**

1. Suppose that our variable for consumption is measured with error, so  $cons = \text{consumption} + e_0$ , where  $e_0$  is uncorrelated with  $inc$ ,  $educ$  and  $size$ . We would expect that:

**c) all our standard errors will be bigger than they would be without the measurement error**

This is random error in the dependent variable, which we don't expect to be much of a problem. The new compound error term means the estimated model has a larger error variance, and thus larger estimated standard errors. The reason it's not d) is that while  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  will all be unbiased,  $\beta_0$  will be biased if the mean of  $e_0$  is not 0.

2. Now suppose that consumption is fine, but that our variable for income is measured with error, so  $inc = \text{income} + e_1$ , where  $e_1$  is uncorrelated with **true** income,  $educ$  and  $size$ . Assuming that **true** income is uncorrelated with  $educ$  and  $size$ , we would expect that:

**d) all of the above**

This is the classical errors in variables assumption, which leads to attenuation bias.

3. Suppose the researcher gets the predicted value,  $yhat$ , and then runs a regression of  $yhat$  on  $inc$ ,  $educ$  and  $size$ . We can

**a) be absolutely certain that the  $R^2$  is equal to 1**

This is a twist on one of the midterm problems. In that one, we regressed  $uhat$  on all of the x's. Here, we are taking  $yhat$ , a linear function of  $inc$ ,  $educ$  and  $size$  and regressing it on  $inc$ ,  $educ$  and  $size$ . Needless to say, we can predict  $yhat$  perfectly in this manner, so  $R^2$  is equal to 1.

4. Suppose the data was collected through a telephone survey, and the last 4 digits of the households telephone number was accidentally named  $assets$  and included in the regression. We would expect

**b) our estimates of  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  and their standard errors to stay about the same**

This one is also a twist on a problematic midterm question. By adding a random number to the regression, we would expect no change in our results, since it is not correlated with either y or the x's.

5. Suppose the following population model meets the Gauss-Markov assumptions:

$pct \text{ of students passing a reading exam} = \beta_0 + \beta_1 \ln(\text{expenditures}) + \beta_2 \ln(\text{enrollment}) + \beta_3 \text{ pct of students living in poverty} + u$ , but we do not have data on the percentage of students living in poverty and plan to use the percentage of students eligible for the federally funded school lunch program as a proxy. Our estimates of  $\beta_1$  and  $\beta_2$  will be consistent if

**d) all of the above must be true**

This is just a list of the properties of a good proxy variable. If they are all true, then the new compound error term will be uncorrelated with all of the x's.

**Part II. Stata Problems.**

1.. desc

Contains data from youth.dta

```
obs:      13,601
vars:      29                               5 Nov 2002 13:56
size:      1,550,514 (99.0% of memory free)
```

---

variable name	storage type	display format	value label	variable label
---------------	--------------	----------------	-------------	----------------

```

-----
greg          float  %9.0g      greg          geographic region (ne, midwest,
                south or west)
metrost       float  %9.0g      metrost       metropolitan status (urban,
                suburban, rural)
ageyrs        float  %9.0g      age in years
male          float  %9.0g      =1 if male
grade         float  %9.0g      grade in school
white         float  %9.0g      =1 if self describe as white
black         float  %9.0g      =1 if self describe as black
hispanic      float  %9.0g      =1 if self describe as at least
                part hispanic
other         float  %9.0g      =1 if self describe as asian,
                native or other
hit           float  %9.0g      =1 if ever hit/slapped by
                boy/girlfriend
forced        float  %9.0g      =1 if ever forced to have sex
consider      float  %9.0g      =1 if ever considered suicide
attempt       float  %9.0g      =1 if ever attempted suicide
smoked        float  %9.0g      =1 if ever smoked
hvsex         float  %9.0g      =1 if ever had sex
partners      float  %9.0g      number of sexual partners
tvhrs         float  %9.0g      number of hours of tv watched
                on school day
teams         float  %9.0g      number of sports team played on
drank         float  %9.0g      =1 if drank alcohol
pot           float  %9.0g      =1 if smoked marijuana
coke          float  %9.0g      =1 if used cocaine
glue          float  %9.0g      =1 if sniffed glue
heroin        float  %9.0g      =1 if used heroin
steroids      float  %9.0g      =1 if used steroids
speed         float  %9.0g      =1 if used methamphetamines
thinkover     float  %9.0g      =1 if think slightly or very
                overweight
bmi           float  %9.0g      body mass index (weight in
                kg/height in meters sq)
ovwt          byte   %8.0g      bmi>85th percentile, so at-risk
                overweight
obese         byte   %8.0g      bmi>95th percentile, so
                overweight
-----

```

```
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
greg	13601	3.034262	.8740555	1	4
metrost	13601	1.705022	.6368897	0	3
ageyrs	13574	16.18447	1.214857	12	18
male	13601	.4854055	.4998053	0	1
grade	13525	10.50048	1.108478	9	12
white	13601	.4658481	.4988506	0	1
black	13601	.1921918	.3940374	0	1
hispanic	13601	.2723329	.4451767	0	1
other	13601	.0696272	.2545271	0	1
hit	13601	.0992574	.2990183	0	1
forced	13601	.0776413	.2676162	0	1
consider	13601	.1774869	.3820943	0	1
attempt	11959	.0918973	.288893	0	1
smoked	13601	.6259834	.4838857	0	1

hvsex		13601	.4683479	.4990155	0	1
partners		12731	1.410494	1.945986	0	6
tvhrs		13246	2.756002	1.723232	0	5.5
teams		13206	.9672876	1.076191	0	3
drank		12143	.7853908	.4105679	0	1
pot		13338	.4417454	.4966134	0	1
coke		13467	.0991312	.2988492	0	1
glue		13039	.1276171	.3336758	0	1
heroin		13461	.0285269	.1664786	0	1
steroids		13531	.0453773	.2081379	0	1
speed		13523	.0906604	.2871361	0	1
thinkover		13476	.1898931	.392231	0	1
bmi		12614	23.17292	4.793146	13.57768	54.98256
ovwt		12614	.1508641	.3579306	0	1
obese		12614	.0765023	.2658106	0	1

a) . reg consider ageyrs white black hisp male partners drank smoked pot coke glue  
> heroin steroids speed, robust

Regression with robust standard errors

Number of obs = 10947  
F( 14, 10932) = 64.21  
Prob > F = 0.0000  
R-squared = 0.0872  
Root MSE = .36958

consider	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
ageyrs	-.0070456	.0030238	-2.33	0.020	-.0129727	-.0011185
white	-.0261966	.0160016	-1.64	0.102	-.0575627	.0051694
black	-.0549422	.01723	-3.19	0.001	-.088716	-.0211684
hisp	-.0363146	.0166541	-2.18	0.029	-.0689597	-.0036695
male	-.1105346	.0070699	-15.63	0.000	-.1243929	-.0966763
partners	.0071329	.0023248	3.07	0.002	.0025758	.0116899
drank	.0500709	.008749	5.72	0.000	.0329214	.0672204
smoked	.0315671	.0088458	3.57	0.000	.0142278	.0489065
pot	.0269106	.0095519	2.82	0.005	.0081871	.045634
coke	.0318064	.0181427	1.75	0.080	-.0037565	.0673693
glue	.1358375	.0142393	9.54	0.000	.1079259	.1637491
heroin	.0853489	.0347715	2.45	0.014	.0171905	.1535074
steroids	.1081753	.0231407	4.67	0.000	.0628153	.1535352
speed	.0679094	.0195819	3.47	0.001	.0295253	.1062935
_cons	.2646026	.0503327	5.26	0.000	.1659414	.3632638

All else equal, it looks like females are more likely to consider suicide, as are younger students. Blacks are least likely to, while each additional sexual partner increases the probability of considering suicide by just under 1 percentage point. Using drugs definitely seems to increase the probability of considering suicide. All types are individually significant at a minimum 5% level, except for cocaine use, which is only significant at the 8% level.

```
. predict probl
(option xb assumed; fitted values)
(2654 missing values generated)
. sum probl
```

Variable	Obs	Mean	Std. Dev.	Min	Max
probl	10947	.1828812	.1141753	-.0276948	.7241641

The problem with the predicted probabilities is that some are negative.

b) I used dprobit (instead of probit) to get the marginal effects that we can interpret in the same manner as the coefficients from the linear probability model. There are no big differences – we would still conclude that all else equal, females are more likely to consider suicide, as are younger students. Blacks are least likely to, while each additional sexual partner increases the probability of considering suicide by just under 1 percentage point. Using drugs still seems to increase the probability of considering suicide. All types are individually significant at a minimum 5% level, and even cocaine is now significant at the 5.4% level.

```
. dprobit consider ageyrs white black hisp male partners drank smoked pot coke
> glue heroin steroids speed
Iteration 0:   log likelihood = -5207.8635
Iteration 1:   log likelihood = -4762.3115
Iteration 2:   log likelihood = -4756.9509
Iteration 3:   log likelihood = -4756.9441
Probit estimates
```

```
Number of obs = 10947
LR chi2(14)    = 901.84
Prob > chi2    = 0.0000
Pseudo R2     = 0.0866
```

```
Log likelihood = -4756.9441
```

consider	dF/dx	Std. Err.	z	P> z	x-bar	95% C.I.
ageyrs	-.0062072	.003088	-2.01	0.044	16.2004	-.01226 -.000155
white*	-.0270195	.0154156	-1.75	0.080	.495478	-.057233 .003194
black*	-.0527346	.0148682	-3.26	0.001	.170732	-.081876 -.023593
hisp*	-.034664	.0151065	-2.22	0.027	.275144	-.064272 -.005056
male*	-.1114642	.0072335	-15.09	0.000	.478213	-.125642 -.097287
partners	.0064321	.0021524	2.99	0.003	1.44642	.002213 .010651
drank*	.0639168	.0099506	5.84	0.000	.794007	.044414 .08342
smoked*	.0337063	.0093989	3.50	0.000	.65945	.015285 .052128
pot*	.0290302	.0091886	3.17	0.002	.470723	.011021 .04704
coke*	.0272083	.0146637	1.93	0.054	.105782	-.001532 .055949
glue*	.1214521	.0132883	10.26	0.000	.130355	.095408 .147497
heroin*	.0566366	.0274512	2.24	0.025	.024025	.002833 .11044
steroids*	.0948439	.0213361	4.99	0.000	.043025	.053026 .136662
speed*	.0517956	.0164404	3.37	0.001	.092719	.019573 .084018
obs. P	.1828812					
pred. P	.1635944	(at x-bar)				

(\*) dF/dx is for discrete change of dummy variable from 0 to 1  
z and P>|z| are the test of the underlying coefficient being 0

```
. predict prob2
(option p assumed; Pr(consider))
(2654 missing values generated)
```

```
. sum prob*
Variable | Obs      Mean      Std. Dev.      Min      Max
-----+-----
```

prob1	10947	.1828812	.1141753	-.0276948	.7241641
prob2	10947	.1828718	.1155977	.030521	.7843404

Now the predicted values are all between 0 and 1, as expected. Note also that the means across the two models are similar.

```
c) . dprobit consider ageyrs white black hisp male partners drank smoked pot coke
> glue heroin steroids speed grade partnersq
```

```
Iteration 0: log likelihood = -5169.8956
Iteration 1: log likelihood = -4730.115
Iteration 2: log likelihood = -4724.7181
Iteration 3: log likelihood = -4724.7114
```

```
Probit estimates
Number of obs = 10900
LR chi2(16) = 890.37
Prob > chi2 = 0.0000
Pseudo R2 = 0.0861
Log likelihood = -4724.7114
```

consider	dF/dx	Std. Err.	z	P> z	x-bar	[	95% C.I.	]
ageyrs	.0113987	.0059647	1.91	0.056	16.2044	-.000292	.023089	
white*	-.0269806	.0154506	-1.74	0.081	.495229	-.057263	.003302	
black*	-.0554786	.014754	-3.44	0.001	.170917	-.084396	-.026561	
hisp*	-.0357415	.0151033	-2.29	0.022	.275505	-.065343	-.00614	
male*	-.1115064	.0072573	-15.04	0.000	.478165	-.12573	-.097282	
partners	.0207214	.006636	3.12	0.002	1.4456	.007715	.033728	
drank*	.0624747	.0099945	5.69	0.000	.794128	.042886	.082063	
smoked*	.0297598	.0094761	3.08	0.002	.659633	.011187	.048333	
pot*	.0274141	.0092549	2.97	0.003	.471009	.009275	.045553	
coke*	.0260468	.0146186	1.85	0.064	.105413	-.002605	.054699	
glue*	.1199742	.0133053	10.12	0.000	.12945	.093896	.146052	
heroin*	.0562259	.0275257	2.22	0.026	.02367	.002277	.110175	
steroids*	.0886679	.0211788	4.68	0.000	.042477	.047158	.130178	
speed*	.0532473	.0165069	3.46	0.001	.092294	.020894	.0856	
grade	-.0230241	.0064092	-3.59	0.000	10.518	-.035586	-.010462	
partne~q	-.0025526	.0011178	-2.28	0.022	5.88083	-.004744	-.000362	
obs. P	.1819266							
pred. P	.1626007	(at x-bar)						

(\*) dF/dx is for discrete change of dummy variable from 0 to 1  
z and P>|z| are the test of the underlying coefficient being 0

Now age has a positive effect, while grade is negative. Thus, while younger (i.e. lower grade) students are still less likely to consider suicide, those who are old for their grade (perhaps having been left back) are more likely to. Number of partners now seems to have a nonlinear effect. It is .021 – 2\*.003\*partners. Thus, the turning point is 3.5, and about 15% of the observations do have more than 3 partners. I don't think I'd use this result to recommend teen sexual promiscuity as the solution to teen suicides!

2. . desc

```
Contains data from C:\Documents and Settings\Patricia_Anderson\My Documents\ECO
> N 20\InClass\vote1.dta
```

```
obs: 173
vars: 12 13 Sep 2000 15:39
size: 6,574 (100.0% of memory free)
```

variable name	storage type	display format	value label	variable label
state	str2	%9s		state postal code
district	byte	%9.0g		congressional district
democA	byte	%9.0g		=1 if A is democrat
voteA	byte	%9.0g		percent vote for A

```

expendA      float   %9.0g      campaign expends. by A, $1000s
expendB      float   %9.0g      campaign expends. by B, $1000s
prtystrA     byte    %9.0g      % vote for president
lexpendA     float   %9.0g      log(expendA)
lexpendB     float   %9.0g      log(expendB)
shareA       float   %9.0g      100*(expendA/(expendA+expendB))

```

Sorted by:

```
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
state	0				
district	173	8.83815	8.768823	1	42
democA	173	.5549133	.498418	0	1
voteA	173	50.50289	16.78476	16	84
expendA	173	310.6111	280.9857	.3	1470.67
expendB	173	305.0884	306.278	.93	1548.19
prtystrA	173	49.75723	9.98365	22	71
lexpendA	173	5.025556	1.601602	-1.197328	7.293476
lexpendB	173	4.944369	1.571143	-.0725707	7.344844
shareA	173	51.07618	33.48389	.09	99.5

a) I'll call the model with share of spending model A, and the model with separate expenditure measures model B

```
. reg voteA shareA democA prtystrA
```

Source	SS	df	MS	Number of obs =	173
Model	42017.405	3	14005.8017	F( 3, 169) =	367.55
Residual	6439.84356	169	38.1055832	Prob > F =	0.0000
Total	48457.2486	172	281.728189	R-squared =	0.8671
				Adj R-squared =	0.8647
				Root MSE =	6.173

voteA	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
shareA	.4358856	.0163758	26.62	0.000	.4035581 .468213
democA	2.374906	1.156002	2.05	0.041	.0928419 4.65697
prtystrA	.2158898	.0578588	3.73	0.000	.1016708 .3301088
_cons	16.17958	2.986474	5.42	0.000	10.28398 22.07517

```
. predict yhata
```

(option xb assumed; fitted values)

```
. reg voteA lexpendA lexpendB democA prtystrA
```

Source	SS	df	MS	Number of obs =	173
Model	38822.1768	4	9705.5442	F( 4, 168) =	169.23
Residual	9635.07174	168	57.3516175	Prob > F =	0.0000
Total	48457.2486	172	281.728189	R-squared =	0.8012
				Adj R-squared =	0.7964
				Root MSE =	7.5731

voteA	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lexpendA	5.779294	.3918197	14.75	0.000	5.00577 6.552819
lexpendB	-6.237836	.3974596	-15.69	0.000	-7.022495 -5.453178
democA	3.792944	1.40652	2.70	0.008	1.016213 6.569674
prtystrA	.2519175	.0712925	3.53	0.001	.1111729 .3926622

```

      _cons |    37.66142    4.736036    7.95    0.000    28.3116    47.01123
-----+-----

```

```

. predict yhatb
(option xb assumed; fitted values)

```

b) I'll do the Davidson-MacKinnon test both ways. In the first, the null is that model A is good, and I cannot reject the null since the p-value for yhatb is .419. In the second, the null is that model B is good, and I can reject the null since the p-value for yhata is .000. If we had not been able to reject one, or had rejected both, we could have seen if one was better based on the adjusted  $R^2$ , or just decided which was easier to interpret. Model A would probably win on both of those counts as well.

```

. reg voteA shareA democA prtystRA yhatb

```

Source	SS	df	MS	Number of obs = 173		
Model	42042.5027	4	10510.6257	F( 4, 168)	=	275.27
Residual	6414.74587	168	38.1830111	Prob > F	=	0.0000
-----+-----				R-squared	=	0.8676
-----+-----				Adj R-squared	=	0.8645
Total	48457.2486	172	281.728189	Root MSE	=	6.1792
-----+-----						
voteA	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
shareA	.4757247	.0518013	9.18	0.000	.3734593	.5779901
democA	2.772508	1.256808	2.21	0.029	.2913363	5.253679
prtystRA	.2458244	.0686856	3.58	0.000	.1102263	.3814225
yhatb	-.102613	.126567	-0.81	0.419	-.3524798	.1472538
_cons	17.6169	3.475651	5.07	0.000	10.75532	24.47848

```

. reg voteA lexpendA lexpendB democA prtystRA yhata

```

Source	SS	df	MS	Number of obs = 173		
Model	42047.8057	5	8409.56115	F( 5, 167)	=	219.11
Residual	6409.44282	167	38.3798971	Prob > F	=	0.0000
-----+-----				R-squared	=	0.8677
-----+-----				Adj R-squared	=	0.8638
Total	48457.2486	172	281.728189	Root MSE	=	6.1952
-----+-----						
voteA	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lexpendA	-.6857767	.7746341	-0.89	0.377	-2.215114	.843561
lexpendB	.5742858	.8110877	0.71	0.480	-1.027021	2.175593
democA	-.2296507	1.231428	-0.19	0.852	-2.660823	2.201522
prtystRA	-.0145895	.0651645	-0.22	0.823	-.1432419	.1140628
yhata	1.093195	.1192456	9.17	0.000	.8577723	1.328619
_cons	-3.246343	5.909451	-0.55	0.584	-14.9132	8.420513