

Sociology 63993
Exam 1 Answer Key
Revised February 26, 2007

I. True-False. (20 points) Indicate whether the following statements are true or false. If false, briefly explain why.

1. An outlier on Y will have the most effect on the regression line when its value for X is equal to the mean of X.

False. Such a case will have no leverage and will have no effect on the regression line, other than to change the intercept.

2. Serial correlation will not affect the unbiasedness or consistency of OLS estimators, but it does affect their efficiency.

True. This is a direct quote from the notes.

3. Increasing the number of items in a scale will always increase the value of Cronbach's Alpha.

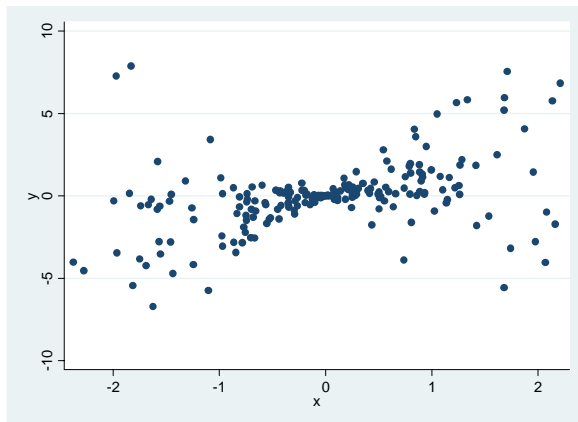
False. Items that don't belong in the scale can lower the overall reliability. The output for the Alpha command includes information on whether deleting an item from the scale will increase or decrease the overall Cronbach's Alpha.

4. Religion has four categories: Catholic, Protestant, Jewish and Other. The researcher wants to use Catholic as her reference category. Therefore, she could compute 3 dummy variables: Protestant, Jewish and Other. On each of these variables, Catholics should be coded as missing.

False. Catholics should be coded as zero.

5. A researcher conducts the following analysis:

```
. scatter y x
```



```
. quietly reg y x
```

```
. hettest
```

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance

Variables: fitted values of y

chi2(1) = 0.18

Prob > chi2 = 0.6721

Based on the above, she can conclude that heteroskedasticity is probably not a problem with her data.

False. The default version of the `hettest` command is testing for one type of heteroskedasticity, i.e. whether the error variances go up as \hat{y} goes up. But the graph suggests a different kind of heteroskedasticity, i.e. the error variances go up as X becomes more extreme in value, either positive or negative. White's general test for heteroskedasticity works better in this case:

```
. estat imtest, white
```

White's test for H_0 : homoskedasticity
against H_a : unrestricted heteroskedasticity

```
chi2(2)      =      49.53
Prob > chi2   =      0.0000
```

II. Short answer. Discuss all three of the following five problems. (15 points each, 45 points total.) In each case, the researcher has used Stata to test for a possible problem, concluded that there is a problem, and then adopted a strategy to address that problem. Explain (a) what problem the researcher was testing for, and why she concluded that there was a problem, (b) the rationale behind the solution she chose, i.e. how does it try to address the problem, and (c) one alternative solution she could have tried, and why. (NOTE: a few sentences on each point will probably suffice – you don't have to repeat everything that was in the lecture notes.)

II-1.

```
. reg y x1 x2 x3
```

Source	SS	df	MS	Number of obs =	3975
Model	7240.36972	3	2413.45657	F(3, 3971) =	24.14
Residual	397049.865	3971	99.9873746	Prob > F =	0.0000
				R-squared =	0.0179
				Adj R-squared =	0.0172
Total	404290.234	3974	101.733828	Root MSE =	9.9994

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
x1	.4211835	.8683167	0.49	0.628	-1.281205 2.123572
x2	.8435702	.8633128	0.98	0.329	-.8490077 2.536148
x3	1.025274	.7932184	1.29	0.196	-.5298791 2.580428
_cons	.2881428	.2493349	1.16	0.248	-.2006937 .7769793

```
. alpha x1 x2 x3, i gen(xscale)
```

Test scale = mean(unstandardized items)

Item	Obs	Sign	item-test correlation	item-rest correlation	average inter-item covariance	alpha
x1	3975	+	0.8006	0.6392	.3240937	0.8219
x2	3975	+	0.8079	0.6466	.3153546	0.8135
x3	3975	+	0.9928	0.9702	.0760545	0.4777
Test scale					.2385009	0.8263

```
. reg y xscale
```

Source	SS	df	MS	Number of obs = 3975		
Model	7175.70039	1	7175.70039	F(1, 3973) = 71.79		
Residual	397114.534	3973	99.9533184	Prob > F = 0.0000		
Total	404290.234	3974	101.733828	R-squared = 0.0177		
				Adj R-squared = 0.0175		
				Root MSE = 9.9977		

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
xscale	2.501135	.2951914	8.47	0.000	1.922395	3.079876
_cons	.3140862	.2443089	1.29	0.199	-.1648963	.7930686

The initial regression suggested multicollinearity was a problem. The global F was significant, but none of the individual T values were. The researcher decided to address the problem by creating a scale out of the three items. With only one independent variable in the subsequent regression, multicollinearity was obviously no longer a problem. The researcher presumably thought it made substantive sense to create a single scale from the three items. Alternative approaches could have included constraining the effects of the three variables to be equal (which would have the same effect as creating a scale), dropping one or more items, or just relying on the global F and not worrying about the individual T values.

//-2.

```
. reg y2 x11
```

Source	SS	df	MS	Number of obs = 3975		
Model	21109.1326	1	21109.1326	F(1, 3973) = 11.39		
Residual	7361172.91	3973	1852.79963	Prob > F = 0.0007		
Total	7382282.04	3974	1857.6452	R-squared = 0.0029		
				Adj R-squared = 0.0026		
				Root MSE = 43.044		

y2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x11	2.790641	.8267667	3.38	0.001	1.169714	4.411567
_cons	-.0881977	1.03719	-0.09	0.932	-2.121672	1.945276

```
. dfbeta
```

```
DFx11:  Dfbeta(x11)
```

```
. extremes DFx11 y2 x11
```

obs:	DFx11	y2	x11
2100.	-.0235065	-25.31333	2.568708
619.	-.0191516	33.0916	-.3147684
3124.	-.0170745	-23.67617	2.227267
3828.	-.0157568	-15.25209	2.531287
3739.	-.0153048	-20.67157	2.223721

2008.	.0140591	35.20511	2.00463
3950.	.0141953	-30.203	-.1235355
906.	.0145061	-26.45573	-.3324702
2546.	.0156778	-29.41599	-.286212
10.	6.147187	2643.918	2.15582

```
. replace y2 = y2/100 in 10
(1 real change made)
```

```
. reg y2 x11
```

Source	SS	df	MS	Number of obs =	3975
Model	7120.96787	1	7120.96787	F(1, 3973) =	71.23
Residual	397169.267	3973	99.9670945	Prob > F =	0.0000
Total	404290.234	3974	101.733828	R-squared =	0.0176
				Adj R-squared =	0.0174
				Root MSE =	9.9984

y2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
x11	1.620833	.1920425	8.44	0.000	1.244322 1.997344
_cons	.3580865	.2409198	1.49	0.137	-.1142516 .8304245

The researcher is concerned about outliers. The dfbeta for case 10 is far larger than 1, and the y2 value for case 10 is much larger than it is for other cases. The researcher apparently felt that the y2 value had been entered incorrectly, i.e. the decimal point was off by two places, so she divided case 10's y2 value by 100. Hopefully she had some additional reasons for doing this, e.g. she checked the original questionnaires to find out what the correct value was. Alternatively, she could have just dropped that case, or used a method like qreg or rreg that is designed to deal with outliers, or reported the results both with and without the outlier so readers could judge for themselves how important it was.

//-3.

. reg income skills scale01

Source	SS	df	MS	Number of obs =	10
Model	444.0924	2	222.0462	F(2, 7) =	9.55
Residual	162.827596	7	23.2610852	Prob > F =	0.0100
				R-squared =	0.7317
				Adj R-squared =	0.6551
Total	606.919997	9	67.4355552	Root MSE =	4.823

income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
skills	.6777237	.3230182	2.10	0.074	-.086093 1.44154
scale01	3.271439	.8299957	3.94	0.006	1.308811 5.234067
_cons	-8.71115	6.793964	-1.28	0.241	-24.77632 7.354021

. list

	income	black	skills	scale01	scale02
1.	5	black	9	1.741101	.
2.	9.7	black	18	3.031433	.
3.	28.4	black	21	5.278032	.
4.	8.8	black	12	5.278032	.
5.	21	black	14	5.278032	.
6.	26.6	black	16	6.309574	.
7.	25.4	black	16	7.300372	.
8.	23.1	black	9	7.300372	.
9.	22.5	black	18	7.300372	.
10.	19.5	black	5	7.300372	.
11.	21.7	white	7	.	7.300372
12.	24.8	white	9	.	7.783137
13.	30.1	white	12	.	8.258524
14.	24.8	white	17	.	8.258524
15.	28.5	white	19	.	8.727161
16.	26	white	6	.	8.727161
17.	38.9	white	17	.	9.189587
18.	22.1	white	1	.	9.189587
19.	33.1	white	10	.	9.646264
20.	48.3	white	17	.	11.42288

. sum

Variable	Obs	Mean	Std. Dev.	Min	Max
income	20	24.415	9.788354	5	48.3
black	20	.5	.5129892	0	1
skills	20	12.65	5.460625	1	21
scale01	10	5.611769	1.93984	1.741101	7.300372
scale02	10	8.850319	1.142792	7.300372	11.42288

. gen xscale01 = scale01
(10 missing values generated)

. replace xscale01 = 5.611769 if missing(scale01)
(10 real changes made)

. gen md = 0

. replace md = 1 if missing(scale01)
(10 real changes made)

```
. reg income skills xscale01 md
```

Source	SS	df	MS	Number of obs = 20		
Model	1242.10592	3	414.035306	F(3, 16) = 11.45		
Residual	578.319556	16	36.1449723	Prob > F = 0.0003		
Total	1820.42548	19	95.8118671	R-squared = 0.6823		
				Adj R-squared = 0.6228		
				Root MSE = 6.0121		

income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
skills	.7629064	.258854	2.95	0.009	.2141605	1.311652
xscale01	3.283386	1.033725	3.18	0.006	1.091988	5.474785
md	12.58468	2.753807	4.57	0.000	6.746875	18.42249
_cons	-9.953716	7.175127	-1.39	0.184	-25.1643	5.256874

The researcher observed that missing data was causing her to lose half her cases in her regression analysis. She then decided to use the old Cohen and Cohen missing data indicator method (which Allison calls “dummy variable adjustment”) to deal with this. She substituted the mean for the missing cases and then added a missing data dummy to her analysis. We now know that this is a bad idea in general (the estimates are biased) and it is probably an even worse idea in this case. We see from the listing of cases that scale01 is only coded for blacks while scale02 is only coded for whites. What the researcher has done, then, is assign the black mean to the white cases. Rather than just mindlessly use a method that is problematic to begin with, the researcher should try to find out why the data are missing in the first place. It may be, for example, that scale01 and scale02 are actually the same question, but whites and blacks got asked these questions at different points in the questionnaire. If so, a single scale could be created using the answers that whites actually provided. If, however, these really are different measures, then the researcher may just want to use listwise deletion, realizing that with scale01 she is only analyzing blacks. [NOTE: Some people said she would lose all her cases if she used listwise, but that would only be true if she tried to use both scales at the same time, which she isn’t in this case.]

III. Computation and interpretation. (35 points total)

Both Sociologists and Public Health researchers are interested in the determinants of self-reported health. The NHANES2F data, available from Stata’s web site, includes information on the following.

Variable	Description
health	Self-reported health. Values range from 1 (poor health) to 5 (excellent health)
age	Age in years
female	Coded 1 if female, 0 otherwise
black	Coded 1 if black, 0 otherwise
rural	Coded 1 if respondent lives in a rural area, 0 otherwise

[NOTE: These data are weighted and proper analysis should take that into account. In addition, the dependent variable is ordinal and would probably be better analyzed by ordinal regression methods that we will talk about later. For simplicity, we will ignore such details for now.]

An analysis of the data yields the following results.

```
. webuse nhanes2f, clear

. keep health age female black rural

. corr , means
(obs=10335)
```

Variable	Mean	Std. Dev.	Min	Max
health	3.413836	1.206196	1	5
age	47.56584	17.21752	20	74
female	.5250121	.4993982	0	1
black	.1050798	.3066711	0	1
rural	.3672956	.4820913	0	1

	health	age	female	black	rural
health	1.0000				
age	-0.3686	1.0000			
female	-0.0320	0.0090	1.0000		
black	-0.1286	-0.0321	0.0100	1.0000	
rural	-0.0827	0.0565	-0.0341	-0.1838	1.0000

```
. alpha female black rural, i gen(demscale)
```

Test scale = mean(unstandardized items)

Item	Obs	Sign	item-test correlation	item-rest correlation	average inter-item covariance	alpha
female	10337	+	0.6438	0.0316	.0271779	0.2855
black	10337	+	0.4966	0.1326	.0082063	0.0659
rural	10337	-	0.6892	0.1247	.0015225	0.0176
Test scale					.0123022	0.1704

```
. drop demscale
```

```
. pcorr2 health age female black rural
```

(obs=10335)

Partial and Semipartial correlations of health with

Variable	Partial	SemiP	Partial^2	SemiP^2	Sig.
age	-0.3730	-0.3675	0.1391	0.1350	0.000
female	-0.0331	-0.0303	0.0011	0.0009	0.001
black	-0.1664	-0.1542	0.0277	0.0238	0.000
rural	-0.0981	-0.0901	0.0096	0.0081	0.000

```
. reg health age female black rural
```

Source	SS	df	MS	Number of obs = 10335		
Model	2472.36904	[1]	618.092259	F(4, 10330) = 508.24		
Residual	12562.6523	10330	1.21613285	Prob > F = 0.0000		
Total	15035.0214	10334	[3]	R-squared = [2]		
				Adj R-squared = 0.1641		
				Root MSE = 1.1028		

health	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	-.0257926	.0006313	[4]	0.000	-.02703	-.0245552
female	-.0731412	.0217366	-3.36	0.001	-.1157492	-.0305331
black	-.6173391	.0359965	-17.15	0.000	-.6878992	-.5467791
rural	-.2296753	.0229357	-10.01	0.000	-.2746336	-.1847169
_cons	4.82831	.0349816	138.02	0.000	4.759739	4.89688

```
. test rural
```

```
( 1) rural = 0
```

```
F( 1, 10330) = [5]
Prob > F = 0.0000
```

a) (10 pts) Fill in the missing quantities [1] – [5].

[1] $DF_{Model} = K = 4$; or, it equals $2472.36904/618.092259 = 4$

[2] $R^2 = SS_{Model}/SS_{Total} = 2472.36904/15035.0214 = .1644$

[3] $MS_{Total} = SS_{Total}/DF_{Total} = 15035.0214/10334 = 1.4549$. Alternatively, $MS_{Total} = Var(health) = SD(health)^2 = 1.206196^2 = 1.4549$

[4] $T_{age} = b_{age}/se_{age} = -.0257926/.0006313 = -40.86$

[5] Wald chi-square = $T_{rural}^2 = -10.01^2 = 100.2$

b) (25 points) Answer the following questions about the analysis and the results, explaining how the printout supports your conclusions.

1. Briefly interpret the results, i.e. explain how race, gender, age and geographic location affect self-reported health. What types of people report the highest levels of health and which types of people report the lowest?

All 4 of the variables have negative effects on health. This means that people who are older, female, black, and/or live in rural areas tend to report lower levels of health than do people who are younger, male, nonblack, and/or live in cities.

2. Suppose that, 20 years from now, your friend decides to finally move away from the big city and live in a small farm out in the country. His mother thinks it is a terrible idea but at least she manages to talk him out of having a sex change operation too. According to the above model, how much higher/lower can your friend expect his health score to be then than it is now?

Your friend will be 20 years older, which will lower his expected health score by more than half a point, i.e. $20 * -.0257926 = -.515852$. The move to rural will further lower his expected score, by $-.2296753$. So, overall his expected health score will drop by about .745 points. Had he gone through with the sex change operation, becoming female

would have cost him an additional expected .0731412 points, which is no doubt why his mother did not want him to do it.

3. The researcher created a variable called demscale but then immediately deleted it. Why did he do this?

The researcher apparently wanted to see whether the dichotomous demographic items could be combined into a single scale. The Cronbach's Alpha was very low, only .17 (.80 or higher is considered good) so he gave up on the idea.

4. Suppose the researcher now ran backwards stepwise regression using the .05 level of significance, i.e. gave the command

```
. sw, pr(.05): reg health age female black rural
```

How would the results differ from the regression reported above?

They wouldn't. All variables in the model are statistically significant, so none would get dropped.

5. Suppose that in previous studies it has been found that, after controlling for other variables, on average women score a tenth of a point lower on health than do men. The researcher therefore decides to test

$$\begin{aligned} H_0: \beta_{\text{female}} &= -.10 \\ H_A: \beta_{\text{female}} &\neq -.10 \end{aligned}$$

Based on the results presented above and using the .05 level of significance, should the researcher reject or not reject the null hypothesis?

As the notes from the Review of Multiple Regression point out, "If the null hypothesis specifies a value that lies within the confidence interval, we will not reject the null." So, the easiest thing is just to note that -.10 falls within the 95% confidence interval (i.e. -.10 falls between -.1157492 and -.0305331), so you will not reject the null.

Alternatively, you can compute the T value for the test:

$$T_{N-K-1} = \frac{b_k - \beta_{k0}}{s_{b_k}} = \frac{-.0731412 - (-.10)}{.0217366} = \frac{0.0268588}{.0217366} = 1.2356$$

When using the .05 level of significance the critical value of T is 1.96, so we do not reject the null. Or, if we had Stata handy, we could just do

```
. test female = -.10
```

```
( 1) female = -.1
```

```
      F( 1, 10330) =      1.53  
      Prob > F =      0.2166
```

IV. Extra Credit (Up to 10 points.)

Y is coded 0 if failed, 1 if succeeded. X1, X2 and X3 are explanatory variables. An analysis of these data reveals the following:

```
. reg y x1 x2 x3
```

Source	SS	df	MS	Number of obs =	32
Model	3.1589065	3	1.05296883	F(3, 28) =	7.26
Residual	4.0598435	28	.144994411	Prob > F =	0.0009
				R-squared =	0.4376
				Adj R-squared =	0.3773
Total	7.21875	31	.232862903	Root MSE =	.38078

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x1	.0348776	.011257	3.10	0.004	.0118187	.0579365
x2	.0088082	.0191632	0.46	0.649	-.0304459	.0480622
x3	.3690164	.1365854	2.70	0.012	.089234	.6487988
_cons	-.6336808	.3923563	-1.62	0.118	-1.437386	.1700247

```
. predict yhat
```

(option xb assumed; fitted values)

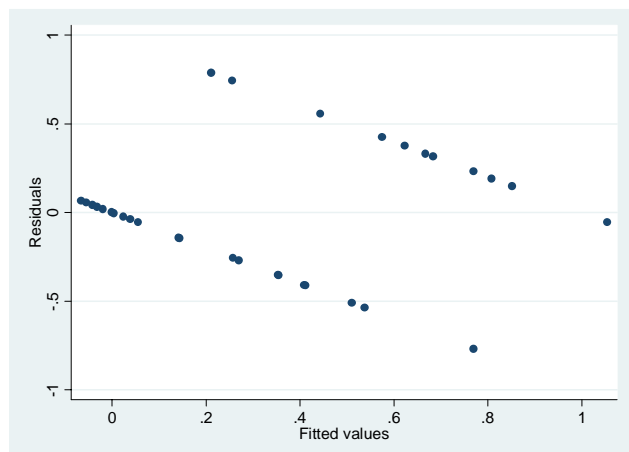
```
. predict e, resid
```

```
. extremes yhat e y
```

obs:	yhat	e	y
3.	-.0662836	.0662836	0
6.	-.0550311	.0550311	0
5.	-.0425547	.0425547	0
8.	-.0328912	.0328912	0
4.	-.0198185	.0198185	0

25.	.7690787	.2309213	1
27.	.769383	-.769383	0
26.	.8080336	.1919664	1
29.	.851562	.148438	1
31.	1.054008	-.0540077	1

```
. rvfplot
```



a) Why does the plot of residuals versus fitted values (i.e. \hat{y} versus e) look the way it does? [HINT: Any OLS regression using a binary dependent variable is going to look more or less like the above. Think about why that has to be.]

a. Recall that $e = y - \hat{y}$. Ergo, when y is a 0-1 dichotomy, it must be the case that either

$$e = -\hat{y} \text{ (which occurs when } y = 0\text{)}$$

or

$$e = 1 - \hat{y} \text{ (which occurs when } y = 1\text{)}.$$

These are equations for 2 parallel lines, which is what you see reflected in the residuals versus fitted plot. The lower line represents the cases where $y = 0$ and the upper line consists of those cases where $y = 1$. The lines slope downward because, as \hat{y} goes up, e goes down.

Whenever y is a 0-1 dichotomy, the residuals versus fitted plot will look something like this; the only thing that will differ are the points on the lines that happen to be present in the data, e.g. if, in the sample, \hat{y} only varies between .3 and .6 then you will only see those parts of the lines in the plot.

Note that this also means that, when y is a dichotomy, for any given value of \hat{y} , only 2 values of e are possible. So, for example, if $\hat{y} = .3$, then e is either $-.3$ or $.7$. This is in sharp contrast to the case when y is continuous and can take on an infinite number of values (or at least a lot more than two).

b) Does anything in the above analysis suggest any problems with the use of OLS regression with binary dependent variables, e.g. are any OLS assumptions being violated, are the estimates sensible? [HINT: the \hat{y} values can be interpreted as the predicted probability of success given the values of the x variables, e.g. a \hat{y} value of .73 implies a 73% chance of success.]

The above results suggest several potential problems with OLS regression using a binary dependent variable. A residuals versus fitted plot in OLS ideally looks like a random scatter of points. Clearly, the above plot does not look like this. This suggests that heteroskedasticity may be a problem. Indeed, as we will prove more formally later on, the assumption of homoskedastic errors is violated with a binary DV.

Also, OLS assumes that, for each set of values for the k independent variables, the residuals are normally distributed. This is equivalent to saying that, for any given value of \hat{y} , the residuals should be normally distributed. This assumption is also clearly violated, i.e. you can't have a normal distribution when the residuals are only free to take on two possible values.

These first two problems suggest that the estimated standard errors will be wrong when using OLS with a dichotomous dependent variable. However, the results from the predict commands also suggest that there may be problems with the plausibility of the model and/or its coefficient estimates. As noted in the hint, \hat{y} can be interpreted as

the estimated probability of success. Probabilities can only range between 0 and 1. However, in OLS, there is no constraint that the \hat{y} estimates fall in the 0-1 range; indeed, \hat{y} is free to vary between negative infinity and positive infinity. In this particular example, the \hat{y} values include both negative numbers (implying probabilities of success that are less than zero) and values greater than 1 (implying that success is more than certain).

As we will more formally show later, there are a number of reasons for believing that OLS regression on a dichotomy produces results that are not plausible. Out of range predictions are an obvious problem but there are other problems that may actually be more serious in practice. Anyone who wants a more detailed explanation can look ahead to the notes on logistic regression, where these and other problems are discussed more fully.

Appendix: Stata Commands for Exam 1. Here are the commands I used to generate the Stata output on the exam. In some cases, I just created fake data that met the conditions I wanted. In other cases, I took an existing data set, and manipulated it in some way to get what I wanted. You should be able to reproduce everything in the exam so long as get the necessary data files off of the web.

*** Problem I-5.

```
set seed 123
set obs 200
corr2data x e
gen y = x + 2*abs(x)*e
reg y x
scatter y x
hettest
```

*** Problem II-1.

```
use "D:\SOC63993\Statafiles\anomia.dta", clear
clonevar x1 = anomia6
clonevar x2 = anomia9
corr2data e1 e2
gen x3 = x1 + x2 + e1*.20
gen y = x1 + x2 + x3 + 10*e2
reg y x1 x2 x3
alpha x1 x2 x3, i gen(xscale)
reg y xscale
```

*** Problem II-2: Same code as 2-1, followed by

```
clonevar x11 = x3
clonevar y2 = y
replace y2 = y2*100 in 10
reg y2 x11
dfbeta
extremes DFx11 y2 x11
drop DFx11
replace y2 = y2/100 in 10
reg y2 x11
dfbeta
extremes DFx11 y2 x11
```

*** Problem II-3.

```
use "D:\SOC63993\Statafiles\reg01.dta", clear
clonevar black = race
clonevar skills = jobexp
gen scale01 = educ^.8 if black
gen scale02 = educ^.8 if !black
keep income black skills scale01 scale02
reg income skills scale01
list
sum
gen xscale01 = scale01
replace xscale01 = 5.611769 if missing(scale01)
gen md = 0
replace md = 1 if missing(scale01)
reg income skills xscale01 md
```

*** Problem III.

```
webuse nhanes2f, clear
keep health age female black rural
order health
corr , means
alpha female black rural, i gen(demscale)
drop demscale
```

```
pcorr2 health age female black rural
reg health age female black rural
test rural
```

*** Problem 4 - Extra credit.

```
use "D:\SOC63993\Statafiles\logist.dta", clear
ren grade y
ren gpa x1
ren tuce x2
ren psi x3
replace x1 = x1^2.5
reg y x1 x2 x3
predict yhat
predict e, resid
extremes yhat e y
rvfplot
```