

**Problem Set 3**  
**ANSWERS**

1. An economist estimates the following model:  $\ln(\text{price}) = \beta_0 + \beta_1 \ln(\text{assess}) + \beta_2 \ln(\text{lotsize}) + \beta_3 \ln(\text{sqrft}) + \beta_4 \text{bdrms} + u$  and plans to use it to predict house prices in his town. He gets predicted  $\ln(\text{price})$ , and then exponentiates it to get a predicted price in dollars. This predicted dollar price:

**a) is definitely an underestimate, on average**

Because OLS fits exactly at the mean, predicted  $\ln(\text{price})$  is exactly right, on average. However,  $\exp(\text{predicted } \ln(\text{price}))$  will be systematically underestimated by  $\exp(u)$ .

2. An economist estimates the following model:  $\ln(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{tenure} + u$  on a sample of 550 workers. He obtains the residuals, *what*, and regresses them on *educ*, *exper*, *tenure*, *exper* squared and *tenure* squared. He is carrying out

**d) an LM test of the null hypothesis that squared terms for *exper* and *tenure* are not necessary**

Since  $n=550$ , an LM is a valid alternative to an F-test for exclusion restrictions. The restricted model was estimated, the unrestricted model (not estimated) would be:  $\ln(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{tenure} + \beta_4 \text{expersq} + \beta_5 \text{tenuresq} + u$  and the null hypothesis is  $\beta_4 = 0, \beta_5 = 0$ .

3. Suppose an economist wants to estimate the following simple regression model:  $y_i = \beta_0 + \beta_1 x_i + u_i$ ,

but decides instead to estimate the following, transformed model:  $\frac{y_i}{\sqrt{h_i}} = \beta_0 \frac{1}{\sqrt{h_i}} + \beta_1 \frac{x_i}{\sqrt{h_i}} + \frac{u_i}{\sqrt{h_i}}$ .

What must the econometrician think  $\text{Var}(u_i|x_i)$  is?

**b)  $\sigma^2 h_i$**

If  $\text{Var}(u|x) = E(u^2)$  is  $\sigma^2 h$ , then the variance of the transformed error term is  $E(u^2)/h = \sigma^2 h/h = \sigma^2$  and thus the transformed model is homoskedastic and we can estimate the proper standard errors with OLS.

4. As sample size,  $n$ , goes to infinity, then

**d) all of the above**

This just indicates that OLS is consistent, that the usual t-statistic can be considered to be drawn from a standardized normal distribution, and that the t-distribution is asymptotically identical to a standardized normal. These are all just asymptotic properties of the OLS estimator.

5. Wages tend to be lower in the south than other regions, as seen by the following estimated model: predicted  $\ln(\text{wage}) = 1.56 + .05 \text{north central} + .15 \text{west} + .12 \text{northeast}$ . If I instead include south and drop north central, then the coefficient on south will be  $-.05$  and

**the last a) the intercept increases by .05 and the other dummy coefficients decrease by .05**  
(Also known as d, sorry about labeling them all a!)

The intercept must increase by  $.05$ , since it now refers to a north central worker who's  $\ln(\text{wage})$  was predicted to be  $1.56 + .05$ . The other coefficients must decrease by  $.05$  because we know that the western worker has a  $\ln(\text{wage})$  of  $.15$  greater than a southern worker, but the north central worker, who is  $.05$  higher than the southern worker, is now the comparison group. Thus, the western worker should have a  $\ln(\text{wage})$  that is  $.10$  greater than the north central worker.

## Part II. Stata Problems.

### 1. Start by looking at the data set ceosal1.dta

```
. desc

Contains data from C:\Documents and Settings\Patricia_Anderson\My Documents\ECO
> N 20\InClass\ceosal1.dta
  obs:          209
  vars:          13          13 Sep 2000 15:27
  size:         7,942 (100.0% of memory free)
```

```
-----
variable name      storage  display  value  variable label
                  type     format   label
-----
salary             int     %9.0g   1990 salary, thousands $
pcsalary           int     %9.0g   % change salary, 89-90
sales              float   %9.0g   1990 firm sales, millions $
roe                float   %9.0g   return on equity, 88-90 avg
pcroe              float   %9.0g   % change roe, 88-90
ros                int     %9.0g   return on firm's stock, 88-90
indus              byte    %9.0g   =1 if industrial firm
finance            byte    %9.0g   =1 if financial firm
consprod           byte    %9.0g   =1 if consumer product firm
utility            byte    %9.0g   =1 if transport. or utilities
lsalary            float   %9.0g   natural log of salary
lsales             float   %9.0g   natural log of sales
ros_sq             float   %9.0g
```

Sorted by:

Note: dataset has changed since last saved

```
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
salary	209	1281.12	1372.345	223	14822
pcsalary	209	13.2823	32.63392	-61	212
sales	209	6923.793	10633.27	175.2	97649.9
roe	209	17.18421	8.518509	.5	56.3
pcroe	209	10.80048	97.2194	-98.9	977
ros	209	61.80383	68.17705	-58	418
indus	209	.3205742	.4678178	0	1
finance	209	.2200957	.4153057	0	1
consprod	209	.2870813	.4534861	0	1
utility	209	.1722488	.3785031	0	1
lsalary	209	6.950386	.5663741	5.407172	9.603868
lsales	209	8.292265	1.01316	5.165928	11.48914
ros_sq	209	8445.584	20350.86	1	174724

a) We need to divide ros by 100.

```
. gen ros100=ros/100
```

b) To be able to allow for the return on stock to possibly be increasing at a decreasing rate, we will need to add ros squared. To estimate a constant elasticity, both salary and sales need to be in logs. The log variables are already available. The estimated elasticity is .27, and the effect of return on stock does seem to increase at a decreasing rate. The turning point is where the derivative is equal to zero,

so where  $.287 - .08 * \text{ros100} * 2 = 0$ , or where  $\text{ros100} = 1.79$ . This is a high return, but within data. If we don't think a negative effect of very high returns makes sense we may be concerned that we have omitted some important variable that is correlated with returns and CEO salary or otherwise don't quite have the model right.

```
. gen ros100sq=ros100^2
```

```
. reg lsalary lsales ros100 ros100sq
```

Source	SS	df	MS	Number of obs = 209		
Model	15.7516839	3	5.25056131	F( 3, 205)	=	21.12
Residual	50.970486	205	.248636517	Prob > F	=	0.0000
-----				R-squared	=	0.2361
-----				Adj R-squared	=	0.2249
Total	66.7221699	208	.320779663	Root MSE	=	.49863
-----						
lsalary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lsales	.270637	.0365194	7.41	0.000	.1986352	.3426388
ros100	.2871359	.110933	2.59	0.010	.0684201	.5058517
ros100sq	-.0785453	.0370208	-2.12	0.035	-.1515357	-.0055549
_cons	4.595068	.3173818	14.48	0.000	3.969317	5.220819

c) The implied model is  $\text{lsalary} = \beta_0 + \beta_1 \text{lsales} + \beta_2 \text{ros100} + \beta_3 \text{ros100sq} + \beta_4 \text{utility} + \beta_5 \text{utility} * \text{lsales} + \beta_6 \text{utility} * \text{ros100} + \beta_7 \text{utility} * \text{ros100sq} + u$  and the null hypothesis is  $\beta_4 = 0, \beta_5 = 0, \beta_6 = 0, \beta_7 = 0$ .

d) To do the Chow test we just need to run the restricted model separately for utility companies and other companies. We get an F-statistic of 8.81 and can reject the null at the 0% level. A standard F-test confirms that I carried out the Chow test correctly.

```
. reg lsalary lsales ros100 ros100sq if utility==1
```

Source	SS	df	MS	Number of obs = 36		
Model	1.73742661	3	.579142203	F( 3, 32)	=	5.42
Residual	3.41758033	32	.106799385	Prob > F	=	0.0039
-----				R-squared	=	0.3370
-----				Adj R-squared	=	0.2749
Total	5.15500694	35	.147285913	Root MSE	=	.3268
-----						
lsalary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lsales	.4224513	.1073437	3.94	0.000	.2037993	.6411033
ros100	-.113937	.2865899	-0.40	0.694	-.6977015	.4698275
ros100sq	.0935198	.1610335	0.58	0.565	-.2344947	.4215343
_cons	3.115521	.8848032	3.52	0.001	1.313236	4.917806

```
. reg lsalary lsales ros100 ros100sq if utility==0
```

Source	SS	df	MS	Number of obs = 173		
Model	13.2299438	3	4.40998125	F( 3, 169)	=	18.66
Residual	39.9467967	169	.236371578	Prob > F	=	0.0000
-----				R-squared	=	0.2488
-----				Adj R-squared	=	0.2355
Total	53.1767405	172	.309167096	Root MSE	=	.48618
-----						
lsalary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	

```

    lsales |   .2381558   .0369272    6.45  0.000   .1652578   .3110538
    ros100 |   .3594834   .1144971    3.14  0.002   .1334546   .5855122
    ros100sq | -.1124389   .0375136   -3.00  0.003  -.1864945  -.0383833
    _cons |   4.933535   .3237368   15.24  0.000   4.294446   5.572624
-----

```

```

. display ((50.97-(3.42+39.95))/(3.42+39.95))*(201/4)
8.805626

```

```

. display Ftail(4,201,8.81)
1.423e-06

```

```

. gen utilros=utility*ros100

```

```

. gen utilrosq=utility*ros100sq

```

```

. gen utilsales=utility*lsales

```

```

. reg lsalary lsales ros100 ros100sq utility utilsales utilros utilrosq

```

Source	SS	df	MS	Number of obs = 209		
Model	23.3577928	7	3.33682755	F( 7, 201)	=	15.47
Residual	43.3643771	201	.215743169	Prob > F	=	0.0000
				R-squared	=	0.3501
				Adj R-squared	=	0.3274
Total	66.7221699	208	.320779663	Root MSE	=	.46448

lsalary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lsales	.2381558	.0352791	6.75	0.000	.1685913	.3077204
ros100	.3594834	.1093869	3.29	0.001	.1437903	.5751765
ros100sq	-.1124389	.0358393	-3.14	0.002	-.1831082	-.0417697
utility	-1.818014	1.295041	-1.40	0.162	-4.371624	.735595
utilsales	.1842955	.1565928	1.18	0.241	-.12448	.493071
utilros	-.4734204	.4217608	-1.12	0.263	-1.305064	.3582229
utilrosq	.2059587	.2316651	0.89	0.375	-.2508469	.6627644
_cons	4.933535	.309288	15.95	0.000	4.32367	5.543401

```

. test utility utilsales utilros utilrosq

```

```

( 1) utility = 0.0
( 2) utilsales = 0.0
( 3) utilros = 0.0
( 4) utilrosq = 0.0

```

```

      F( 4, 201) =      8.81
      Prob > F =      0.0000

```

```

. clear

```

## 2. Start by taking a look at the data.

```

. use nbasal

```

```

. desc

```

```

Contains data from nbasal.dta

```

```

obs:                269

```

```
vars:          22          25 May 2002 11:13
size:         12,912 (100.0% of memory free)
```

---

variable name	storage type	display format	value label	variable label
marr	byte	%1.0f		=1 if married
wage	float	%9.2f		annual salary, thousands \$
exper	byte	%2.0f		years as a professional player
age	byte	%2.0f		age in years
coll	byte	%1.0f		years playing at college
games	byte	%6.0f		average games per year
minutes	int	%6.0f		minutes per season
guard	byte	%1.0f		=1 if guard
forward	byte	%1.0f		=1 if forward
center	byte	%1.0f		=1 if center
points	float	%2.0f		points per game
rebounds	float	%2.0f		rebounds per game
assists	float	%2.0f		assists per game
draft	int	%3.0f		draft number
allstar	byte	%1.0f		all-star player
avgmin	float	%9.2f		minutes per game
lwage	float	%9.2f		natural log of wage
black	byte	%1.0f		=1 if black
children	byte	%1.0f		=1 if has children
expersq	int	%3.0f		exper^2
agesq	int	%9.0g		age^2
marrblck	byte	%1.0f		marr*black

---

Sorted by:

```
. sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
marr	269	.4423792	.4975945	0	1
wage	269	1423.828	999.7741	150	5740
exper	269	5.118959	3.400062	1	18
age	269	27.39405	3.391292	21	41
coll	269	3.717472	.7544096	0	4
games	269	65.72491	18.85111	3	82
minutes	269	1682.193	893.3278	33	3533
guard	269	.4200743	.4944905	0	1
forward	269	.4089219	.4925512	0	1
center	269	.1710037	.377214	0	1
points	269	10.21041	5.900667	1.2	29.8
rebounds	269	4.401115	2.892573	.5	17.3
assists	269	2.408922	2.092986	0	12.6
draft	240	20.2	18.73582	1	139
allstar	269	.1152416	.3199085	0	1
avgmin	269	23.97925	9.731177	2.888889	43.08537
lwage	269	6.952296	.881376	5.010635	8.655214
black	269	.8066914	.3956289	0	1
children	269	.3457249	.4764905	0	1
expersq	269	37.72119	46.53702	1	324
agesq	269	761.8922	195.1494	441	1681
marrblck	269	.33829	.4740096	0	1

a) The appropriate model is  $\text{points} = \beta_0 + \beta_1\text{guard} + \beta_2\text{forward} + \beta_3\text{exper} + \beta_4\text{coll} + \beta_5\text{draft} + u$  and we want to test  $H_0: \beta_2 = 0$  against  $H_1: \beta_2 > 0$ . Since center is the omitted group,  $\beta_2$  tells us the difference in points between forwards and centers. Since the bet is about forwards scoring **more** than centers, a one-sided test is probably most appropriate. A one-sided p-value is obtained by dividing the reported 2-sided value in half. Thus, at the 5% level, Rick would win the bet. We can reject the null at the 3.65% level.

```
. reg points guard forward exper coll draft
```

Source	SS	df	MS			
Model	1550.88942	5	310.177884	Number of obs =	240	
Residual	6606.54642	234	28.2331044	F( 5, 234) =	10.99	
Total	8157.43584	239	34.1315307	Prob > F =	0.0000	
				R-squared =	0.1901	
				Adj R-squared =	0.1728	
				Root MSE =	5.3135	

  

points	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
guard	3.26752	.9758967	3.35	0.001	1.344854	5.190187
forward	1.761988	.9780607	1.80	0.073	-.1649413	3.688918
exper	.3123559	.1003041	3.11	0.002	.1147415	.5099703
coll	-1.249331	.4855555	-2.57	0.011	-2.20595	-.2927116
draft	-.1036279	.0186794	-5.55	0.000	-.1404292	-.0668266
_cons	13.80728	2.000939	6.90	0.000	9.865119	17.74943

b) In order to do a Breusch-Pagan test, we need to get the residuals and square them. The squared residuals are regressed on all the x variables in the model. The overall F-test reports that we can reject that they have no explanatory power at the .5% level. Thus, we reject the null of homoskedasticity. We should estimate the model with robust standard errors. (Note that it looks like there is a higher variance in scoring across centers compared to guards and forwards.)

```
. predict uhat, r
(29 missing values generated)
```

```
. gen uhatsq=uhat^2
(29 missing values generated)
```

```
. reg uhatsq guard forward exper coll draft
```

Source	SS	df	MS			
Model	34450.4866	5	6890.09732	Number of obs =	240	
Residual	467185.401	234	1996.51881	F( 5, 234) =	3.45	
Total	501635.888	239	2098.89493	Prob > F =	0.0050	
				R-squared =	0.0687	
				Adj R-squared =	0.0488	
				Root MSE =	44.682	

  

uhatsq	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
guard	-25.18376	8.206562	-3.07	0.002	-41.35194	-9.015571
forward	-23.52945	8.224759	-2.86	0.005	-39.73349	-7.325414
exper	1.47733	.8434821	1.75	0.081	-.1844591	3.139119
coll	1.297293	4.083159	0.32	0.751	-6.747157	9.341744
draft	-.2578956	.15708	-1.64	0.102	-.5673673	.051576
_cons	40.0768	16.8264	2.38	0.018	6.926211	73.2274

c) We just need to rerun the model using the robust option. Now, while the point estimate is still that forwards score 1.8 more points per game than centers, the lowest level that we can reject the null is the 7.45% level. Thus, with the robust standard errors, Rick would lose the bet to Shaquille.

```
. reg points guard forward exper coll draft, robust
```

```
Regression with robust standard errors
```

Number of obs =	240
F( 5, 234) =	12.17
Prob > F =	0.0000
R-squared =	0.1901
Root MSE =	5.3135

```
-----
```

points		Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
guard		3.26752	1.148519	2.84	0.005	1.004762 5.530279
forward		1.761988	1.218103	1.45	0.149	-.6378622 4.161839
exper		.3123559	.1141083	2.74	0.007	.087545 .5371668
coll		-1.249331	.3791934	-3.29	0.001	-1.9964 -.5022613
draft		-.1036279	.0225603	-4.59	0.000	-.1480751 -.0591807
_cons		13.80728	1.803133	7.66	0.000	10.25483 17.35973

```
-----
```

```
. log close
```