ECO 5350 T. Fomby
Intro. Econometrics Fall 2016

EXERCISE 6
KEY

Purpose: To learn more about multiple regression prediction and prediction confidence intervals and marginal effects of variables in models with quadratic and interaction terms in them. **This exercise is due on Thursday, November 3.**

Work **word problem exercises 7 and 10, pp. 197 and 198** in your textbook. Work **computer exercises C8 and C10 on page 201** in your textbook. Each of these exercises will count 10 points each. I want you to provide "clean" answer sheets in the sense that I want your answers to be "typed up" using Microsoft Word or some equivalent document using a Mac word processor. I will be taking up all of the exercises.

7.

All three models have the same dependent variable so all we have to do is pick the equation that has the highest adjusted R2. That is the second equation.

10.

(i) The second equation has the *read4* score in it which is likely endogenously determined along with the *math4* score. That is, the unobserved factors represented by the error u probably affects the *math4* score at the same time it affects the read4 score. This endogeneity then invalidates the assumption MLR.4, p. 92 in our textbook. In this case, OLS provides inconsistent estimates of the coefficients. For this reason the first equation is preferred for policy analysis purposes. *Math4* is the percent of students scoring acceptably on the math test. Than a 10% in increase in expenditures per pupil is represented by a 0.10 change in *lexppp*. Thus, the implied change in the percentage of acceptable math scores is (0.1)(9.01) = 0.901 percent. That is not very much.

(ii) The *free* variable becomes insignificant while the *lmedinc* and *pctsgle* variables become significant.

(iii) Between two equations, one which provides "accurate" (consistent) information while the other one does not, you always go with the former equation even though it might have a smaller adjusted-R2.

C8. For the STATA key see Exercise 6_C8_Key.do

The code is

```
* Use Hprice1.dta
regress price lotsize sqrft bdrms
generate lotsize0 = lotsize - 10000
```

generate sqrft0 = sqrft - 2300
generate bdrms0 = bdrms - 4
*** Prediction at the designated point**
regress price lotsize0 sqrft0 bdrms0

```
. regress price lotsize sqrft bdrms
```

Source	SS	df	MS	Number of obs	=	88
				F(3, 84)	=	57.46
Model	617130.701	3	205710.234	Prob > F	=	0.0000
Residual	300723.805	84	3580.0453	R-squared	=	0.6724
				Adj R-squared	=	0.6607
Total	917854.506	87	10550.0518	Root MSE	=	59.833

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lotsize	.0020677	.0006421	3.22	0.002	.0007908	.0033446
sqrft	.1227782	.0132374	9.28	0.000	.0964541	.1491022
bdrms	13.85252	9.010145	1.54	0.128	-4.065141	31.77018
_cons	-21.77031	29.47504	-0.74	0.462	-80.38466	36.84405

$$\widehat{price} = \underset{(29.47504)}{-21.77031} + \underset{(0.0006421)}{0.0020677}\widehat{lotsize} + \underset{(0.0132374)}{0.1227782}sqrft + \underset{(13.85252)}{13.85252}bdrms$$

336.7066 = -21.77031 + 0.0020677(10,000) + 0.1227782(2,300) + 13.85252(4)

which represents $336,707.

(ii) The confidence interval for the intercept (prediction) is given directly by the output on the "translated" regression: [322.0417, 351.3716]. This is the confidence interval for the prediction of the **mean** of homes with the given characteristics.

```
. * Prediction at the designated point
. regress price lotsize0 sqrft0 bdrms0
```

Source	SS	df	MS	Number of obs	=	88
				F(3, 84)	=	57.46
Model	617130.701	3	205710.234	Prob > F	=	0.0000
Residual	300723.805	84	3580.0453	R-squared	=	0.6724
				Adj R-squared	=	0.6607
Total	917854.506	87	10550.0518	Root MSE	=	59.833

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lotsize0	.0020677	.0006421	3.22	0.002	.0007908	.0033446
sqrft0	.1227782	.0132374	9.28	0.000	.0964541	.1491022
bdrms0	13.85252	9.010145	1.54	0.128	-4.065141	31.77018
_cons	336.7067	7.374466	45.66	0.000	322.0417	351.3716

(iii) The standard error of the prediction of y_0 (an **individual** house price with the given characteristics) is going to be greater than the standard error of the prediction of the mean of homes with the given characteristics. This standard error is calculated by

$$se(\hat{y}_0) = sqrt(\widehat{var}\big(\hat{E}(y|X = x_0)) + \ MSresidual\big) = sqrt[7.374466^2 + 3580.0453]$$
$$= \ sqrt(54.383 + 3580.0453) = 60.286$$

Therefore, the 95% confidence interval for y_0 is $\hat{y}_0 \ \pm 1.96(60.286) = 336.7067 \pm 118.16$ $= [218.55, 454.87]$. This confidence interval is certainly wider than the previous confidence interval.

C10. For the STATA key see Exercise 6_C10_Key.do

The code is

```
* Use BWGHT2.dta
regress lbwght npvis npvissq
summarize npvis if npvis > 21
regress lbwght npvis npvissq mage magesq
summarize mage if mage > 30
* Getting the R^2 and SST from the bwght equation
regress bwght npvis npvissq mage magesq
* R^2_bwght = 0.0192, SST = 593759796
* Now get the predict values of birthweight from the lbwght equation
regress lbwght npvis npvissq mage magesq
predict lbwght_hat
generate bwght_tilda = exp(lbwght_hat + 0.0411052/2)
generate SSR_y_tilda = 1764*(bwght - bwght_tilda)^2
summarize SSR_y_tilda
* SSR_y_tilda = 5.85e+08
* R^2_y_tilda = 1 - SSR_y_tilda/SSt = 1 - 5.85e+8/593759796 = 0.0147531
*  Since the bwght equation had a higher R^2_bwght than the corresponding
* R^2_y_tilda implied from the lbwght equation, we go with the bwght specification.
```

(i)

```
. regress lbwght npvis npvissq
```

Source	SS	df	MS	Number of obs	=	1,764
				F(2, 1761)	=	19.12
Model	1.5771321	2	.788566048	Prob > F	=	0.0000
Residual	72.6282777	1,761	.041242634	R-squared	=	0.0213
				Adj R-squared	=	0.0201
Total	74.2054098	1,763	.04209042	Root MSE	=	.20308

lbwght	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
npvis	.0189167	.0036806	5.14	0.000	.0116979	.0261355
npvissq	-.0004288	.00012	-3.57	0.000	-.0006641	-.0001934
_cons	7.957883	.0273125	291.36	0.000	7.904314	8.011451

The quadratic term is significant as the p-value of its coefficient is less than 0.05.

$$\widehat{lbwght} = 7.95 + 0.0189npvis - 0.0004288npvissq$$

(0.027) (0.0036) (0.00012)

The calculus easily shows that $npvis^* = \frac{-\widehat{\beta}_1}{2\widehat{\beta}_2} = \frac{-0.0189167}{2(-0.0004288)} = 22.057.$

Only 21 women out of the 1,764 had 22 or more prenatal visits.

(iii) One would think that prenatal visits would add to birthweight but as we see at a decreasing rate. Probably a healthy women insisting on 22 prenatal visits to achieve maximal birthweight of their children is a little far-fetched. The quadratic term here is just indicating there are birthweight gain benefits of prenatal visits but this benefit is diminishing slightly with each additional prenatal visit.

(iv)

```
. regress lbwght npvis npvissq mage magesq
```

Source	SS	df	MS	Number of obs	=	1,764
				F(4, 1759)	=	11.56
Model	1.90136387	4	.475340968	Prob > F	=	0.0000
Residual	72.3040459	1,759	.0411052	R-squared	=	0.0256
				Adj R-squared	=	0.0234
Total	74.2054098	1,763	.04209042	Root MSE	=	.20274

lbwght	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
npvis	.0180374	.0037086	4.86	0.000	.0107636	.0253112
npvissq	-.0004079	.0001204	-3.39	0.001	-.0006441	-.0001717
mage	.025392	.0092542	2.74	0.006	.0072417	.0435423
magesq	-.0004119	.0001548	-2.66	0.008	-.0007154	-.0001083
_cons	7.583713	.1370568	55.33	0.000	7.314901	7.852524

The calculus easily shows that $mage^* = \frac{-\widehat{\beta}_3}{2\widehat{\beta}_4} = \frac{-0.025392}{2(-0.0004119)} = 30.82.$

Then average birthweight begins to fall beginning at age 31. 746 of the women in the sample of 1,764 are 31 years of age or greater.

(v)

Doing an F-test on the joint significance of the variables *npvis*, *npvissq*, *mage*, and *magesq* results in an overall F-statistic of 11.56 with a p-value of 0.0000. Thus, these variables, jointly speaking, are very significant is explaining birthweight.

(vi)

For the bwght equation, R^2_bwght = 0.0192. For the *lbwght* equation, the implied R^2 for bwght is R^2_y_tilda = 0.01475. Since, the R^2 of the *bwght* equation is higher than the R^2 implied from the *lbwght* equation, we go with the *bwgh*t equation.