ECO 5350 T. Fomby
Intro. Econometrics Fall 2016

EXERCISE 8
KEY

Purpose: To learn more about using **indicator (dummy) variables** in multiple regression and the **linear probability model**. **This exercise is due on Thursday, November 10.**

Work **word problem exercises 2 and 4, pages 233 and 234** in your textbook. Work **computer exercises C3, C6, and C8 on pages 237, 238, and 239** 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.

2.

(i) $\Delta\log(\widehat{bwght}) = -0.0044\Delta cigs = -0.0044(10) = 0.044$ This implies that with increased consumption of 10 cigarettes per day, the baby's average birthweight drops by 4.4%.

(ii) The coefficient on the "white" indicator variable is 0.055. This means that a white child, on average, is expected to have a 5.5% higher birth weight than a non-white child.

(iii) The coefficient on the "motheduc" variable is an unexpected negative number, -0.0030. However, the coefficient is not statistically significant so one shouldn't put too much weight on this counter-intuitive result. This "weird" result might be resulting from the multicollinearity that exists between the *mothedu* and *fathedu* variables. It might be better to model the effect of parent's education on birth weight by forming a combined variable parent's education = *mothedu* + *fathedu*.

(iv) For some reason, probably some missing observations in the second equation, the two regression equations do not have the same number of observations (1,388 for the first equation and 1,191 observations for the second equation. This prevents us from being able to do a subset F- test on the joint significance of the *mothedu* and *fathedu* variables.

4.

(i) The coefficient of -0.283 on the utility variable indicates that the salaries of workers in the utility industry are 28.3 % lower than those in the transportation industry. This difference is statistically significant at least the 5% level because the t-ratio on the utility variable is -0.283/0.099 = - 2.8585 and its absolute value, 2.2585, is greater than 2.0 (the two-sigma rule). Again, one could use a p-value calculator and get a p-value for the statistic which is going to be less than 0.05 by quite a bit.

(ii) The exact percentage difference is calculated by the formula $100[\exp(-0.283) - 1] =$

-24.648 %. Thus, we can see for big percentage change the formula (7.10) might be a better one to use as compared to looking at the logarithmic changes.

(iii) The salaries of workers in the consumer product industries are, on average, 2.3 percent higher than in the finance industry, 0.181 – 0.158 = 0.023. The null hypothesis that the salaries of these two industries are not statistically significant would be $H_0: \beta_3 - \beta_4 = 0$. This can be tested two equivalent ways:

One can use the conventional t-test: $t = \frac{\widehat{\beta}_3 - \widehat{\beta}_4}{se(\widehat{\beta}_3 - \widehat{\beta}_4)} = \frac{\widehat{\beta}_3 - \widehat{\beta}_4}{sqrt[\widehat{var}(\widehat{\beta}_3) + \widehat{var}(\widehat{\beta}_4) - 2\,\widehat{cov}(\widehat{\beta}_3, \widehat{\beta}_4)]}$

or the F-test where the **unrestricted regression** is the one that is listed and the **restricted regression** is

$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \beta_3(finance + consprod) + \beta_5 utility + u.$

The corresponding F-statistic is

$$F_{q,N-K} = \frac{(SSR_r - SSR_{u)q}}{SSR_u/(N-K)} = \frac{(R_u^2 - R_r^2)/q}{(1-R_u^2)/(N-K)}.$$

C3. For STATA program key see **Exercise_C3_Ch7_Key.do**

The code is

```
* Use Mlb1.DTA
* The base group is outfielders.  Obviously, pitchers are not being considered.
regress lsalary years gamesyr bavg hrunsyr rbisyr runsyr fldperc allstar frstbase scndbase thrdbase shrtstop catcher
* Test that salaries do not differ by position
test frstbase scndbase thrdbase shrtstop catcher
```

(i) The null hypothesis that catchers and outfielders, on average, earn the same salaries is equivalent to testing the statistical significance of the catcher coefficient since outfielders is the base group. The t-ratio on the catcher variable is 1.93 with a two-sided p-value of 0.054. This ratio is significant at the 5% level if we use a one-sided alternative but not if we use a two-sided alternative.

(ii) Using the "test" statement we get
test frstbase scndbase thrdbase shrtstop catcher

```
 ( 1)  frstbase = 0
 ( 2)  scndbase = 0
 ( 3)  thrdbase = 0
 ( 4)  shrtstop = 0
 ( 5)  catcher = 0

       F(  5,   339) =    1.78
            Prob > F =    0.1168
```

Given this joint test, we conclude that there is not a significant difference in salaries across non-pitching positions.

(iii) The joint test of part (ii) is a very broad test where most of the salary differentials between positions are not statistically significant. That is, in general, it seems that position doesn't matter except in the case of catchers which tend to "wear out" earlier due to the grueling nature of the position.

C6. For STATA program key see **Exercise_C6_Ch7_Key.do**

The code is

```
* Use SLEEP75.dta
* Male regression equation
regress sleep totwrk educ age agesq yngkid if male == 1
* Female regression equation
regress sleep totwrk educ age agesq yngkid if male == 0
* Generate multiplicative dummies
generate totwrk_male = totwrk*male
generate educ_male = educ*male
generate age_male = age*male
generate agesq_male = agesq*male
generate yngkid_male = yngkid*male
* Additive/Multiplicative Dummy model (unrestricted regression)
regress sleep totwrk educ age agesq yngkid male totwrk_male educ_male age_male
agesq_male yngkid_male
* Chow test for difference between male and female groups
test male totwrk_male educ_male age_male agesq_male yngkid_male
* Restricted regression which assumes apriori that the groups are the same
regress sleep totwrk educ age agesq yngkid
* Testing the joint significance of the multiplicative dummies
regress sleep totwrk educ age agesq yngkid male totwrk_male educ_male age_male
agesq_male yngkid_male
test totwrk_male educ_male age_male agesq_male yngkid_male
* Model with only intercept shift
regress sleep totwrk educ age agesq yngkid male
```

```
. regress sleep totwrk educ age agesq yngkid if male == 1
```

Source	SS	df	MS	Number of obs	=	400
				F(5, 394)	=	14.59
Model	11806161.6	5	2361232.32	Prob > F	=	0.0000
Residual	63763979	394	161837.51	R-squared	=	0.1562
				Adj R-squared	=	0.1455
Total	75570140.6	399	189398.849	Root MSE	=	402.29

sleep	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
totwrk	-.1821232	.0244855	-7.44	0.000	-.2302618	-.1339846
educ	-13.05238	7.414218	-1.76	0.079	-27.62876	1.523996
age	7.156591	14.32037	0.50	0.618	-20.99731	35.31049
agesq	-.0447674	.1684053	-0.27	0.791	-.3758528	.2863181
yngkid	60.38021	59.02278	1.02	0.307	-55.65877	176.4192
_cons	3648.208	310.0393	11.77	0.000	3038.67	4257.747

Here is the regression for **females**:

```
. regress sleep totwrk educ age agesq yngkid if male == 0
```

Source	SS	df	MS	Number of obs	=	306
				F(5, 300)	=	6.50
Model	6201576.18	5	1240315.24	Prob > F	=	0.0000
Residual	57288575.9	300	190961.92	R-squared	=	0.0977
				Adj R-squared	=	0.0826
Total	63490152.1	305	208164.433	Root MSE	=	436.99

sleep	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
totwrk	-.1399495	.0276594	-5.06	0.000	-.1943806	-.0855184
educ	-10.20514	9.588848	-1.06	0.288	-29.07506	8.664787
age	-30.35657	18.53091	-1.64	0.102	-66.82361	6.110464
agesq	.3679406	.2233398	1.65	0.101	-.0715705	.8074516
yngkid	-118.2826	93.18757	-1.27	0.205	-301.6667	65.10154
_cons	4238.729	384.8923	11.01	0.000	3481.299	4996.16

The signs of the "yngkid" variable are different for males and females: The more young kids in the family, the less sleep women have (however, not statistically significant) while the number of young kids in the family does not seem to affect the number of hours of sleep very much for men. Also the signs of the coefficients of age and age^2 are different for males as compared to females. The quadratic age profile equation for males is concave and has a peak at 80.04 years (age* = 7.156/(2*0.0447) = 80.04) while the quadratic age profile equation for females is convex and has a trough at 41.36 years (age* = 30.356/(2*0.367) = 41.36). Sleep for males essentially increases throughout their lifetimes while, for women, their sleep decreases until age 41 and then increases thereafter.

(ii) Here is the Stata output for the Chow Test of significant different in the male and female sleep equations:

```
. regress sleep totwrk educ age agesq yngkid male totwrk_male educ_male age_male agesq_male yngki
> d_male

      Source |       SS           df       MS      Number of obs   =       706
-------------+----------------------------------   F(11, 694)      =      9.48
       Model |  18187280.8        11  1653389.17   Prob > F        =    0.0000
    Residual |   121052555       694  174427.313   R-squared       =    0.1306
-------------+----------------------------------   Adj R-squared   =    0.1168
       Total |   139239836       705  197503.313   Root MSE        =    417.64

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totwrk |  -.1399495   .0264349    -5.29   0.000    -.1918514   -.0880476
        educ |  -10.20514   9.164321    -1.11   0.266    -28.19826    7.787983
         age |  -30.35657   17.71049    -1.71   0.087    -65.12914    4.415998
       agesq |   .3679406   .2134519     1.72   0.085    -.0511483    .7870294
      yngkid |  -118.2826   89.06187    -1.33   0.185    -293.1456    56.58047
        male |  -590.5211   488.7916    -1.21   0.227    -1550.209    369.1665
 totwrk_male |  -.0421737    .036674    -1.15   0.251     -.114179    .0298317
   educ_male |  -2.847243   11.96795    -0.24   0.812    -26.34497    20.65048
    age_male |   37.51316   23.12332     1.62   0.105    -7.886888    82.91321
  agesq_male |  -.4127079   .2759136    -1.50   0.135    -.9544333    .1290175
 yngkid_male |   178.6628   108.1051     1.65   0.099     -33.5895     390.915
       _cons |   4238.729   367.8519    11.52   0.000     3516.493    4960.965
------------------------------------------------------------------------------

. test male totwrk_male educ_male age_male agesq_male yngkid_male

 ( 1)  male = 0
 ( 2)  totwrk_male = 0
 ( 3)  educ_male = 0
 ( 4)  age_male = 0
 ( 5)  agesq_male = 0
 ( 6)  yngkid_male = 0

       F(  6,   694) =    2.12
            Prob > F =    0.0495
```

The conclusion is that the male and female sleep equations are statistically different. The F-statistic for the Chow test is 2.12 with a p-value of 0.0495 < 0.05.

It should be noted that using the "separate regressions" approach will give you the same answer.

From the above two separate regressions we see $SSR_{male} = 63763979$ and $SSR_{female} = 57288575.9$. Therefore, from the separate equations we see the unrestricted sum-of-squared errors is $SSR_u = SSR_{male} + SSR_{female}$ = 121052555 as reported in the analysis-of-variance table in the unrestricted model (the additive/multiplicative dummy variable model).

The restricted regression assuming the two groups are the same is

```
. regress sleep totwrk educ age agesq yngkid

      Source |       SS           df       MS      Number of obs   =       706
-------------+----------------------------------   F(5, 700)       =     18.14
       Model |  15972384.7         5  3194476.94   Prob > F        =    0.0000
    Residual |   123267451       700  176096.359   R-squared       =    0.1147
-------------+----------------------------------   Adj R-squared   =    0.1084
       Total |   139239836       705  197503.313   Root MSE        =    419.64

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totwrk |  -.1460463   .0168809    -8.65   0.000    -.1791896   -.1129031
        educ |  -11.13772   5.890168    -1.89   0.059    -22.70223    .4267914
         age |  -8.123949   11.37049    -0.71   0.475     -30.4483     14.2004
       agesq |    .126287    .135186     0.93   0.351    -.1391317    .3917057
      yngkid |   17.15441   50.00839     0.34   0.732    -81.02999    115.3388
       _cons |   3825.375   240.2585    15.92   0.000     3353.661    4297.088
------------------------------------------------------------------------------
```

Then the Chow test F-statistic using this different approach is the same as reported above, namely,

$$F = \frac{(123267451 - 121052555)/6}{121052555/(706-12)} = 2.12$$

(iii) Now we do a joint test of the multiplicative dummies in the additive/multiplicative dummy model.

```
. regress sleep totwrk educ age agesq yngkid male totwrk_male educ_male age_male agesq_male yngki
> d_male
. test totwrk_male educ_male age_male agesq_male yngkid_male

 ( 1)  totwrk_male = 0
 ( 2)  educ_male = 0
 ( 3)  age_male = 0
 ( 4)  agesq_male = 0
 ( 5)  yngkid_male = 0

       F(  5,   694) =    1.26
            Prob > F =    0.2814
```

From this joint test, we see that the multiplicative dummies are not jointly significant because the joint F-test has a p-value = 0.2814 that is greater than 0.05. Maybe the simplest model we can entertain is to model the group differences as simply an additive (intercept) shift in the regression equation. This model is estimated to be

```
. regress sleep totwrk educ age agesq yngkid male

      Source |       SS           df       MS      Number of obs   =       706
-------------+----------------------------------   F(6, 699)       =     16.30
       Model |  17092058.6         6  2848676.43   Prob > F        =    0.0000
    Residual |   122147777       699  174746.462   R-squared       =    0.1228
-------------+----------------------------------   Adj R-squared   =    0.1152
       Total |   139239836       705  197503.313   Root MSE        =    418.03

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totwrk |  -.1634235   .0181634    -9.00   0.000    -.1990848   -.1277622
        educ |  -11.71327   5.871952    -1.99   0.046    -23.24205   -.1844947
         age |  -8.697402   11.32909    -0.77   0.443    -30.94053    13.54572
       agesq |   .1284415   .1346696     0.95   0.341    -.1359638    .3928469
      yngkid |  -.0228006   50.27641    -0.00   1.000    -98.73367    98.68807
        male |   87.75455   34.66794     2.53   0.012     19.68877    155.8203
       _cons |   3840.852   239.4139    16.04   0.000     3370.795    4310.909
------------------------------------------------------------------------------

. test age agesq yngkid

 ( 1)  age = 0
 ( 2)  agesq = 0
 ( 3)  yngkid = 0

       F(  3,   699) =    0.92
            Prob > F =    0.4292
```

From these results we can see that the variables *age*, *agesq*, and *yngkid* are not jointly significant since the F-statistic has a probability value that is greater than 0.05. This suggests that we should look at the model that has an additive dummy "male" and, in addition, only the variables *totwrk* and *educ*. This Final model is reported below. All of the variables in the model are statistically significant.

Final Model:

```
. regress sleep totwrk educ male

      Source |       SS           df       MS      Number of obs   =       706
-------------+----------------------------------   F(3, 702)       =     31.69
       Model |  16608173.4         3  5536057.79   Prob > F        =    0.0000
    Residual |   122631662       702  174688.978   R-squared       =    0.1193
-------------+----------------------------------   Adj R-squared   =    0.1155
       Total |   139239836       705  197503.313   Root MSE        =    417.96

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      totwrk |  -.1673395   .0179371    -9.33   0.000    -.2025562   -.1321228
        educ |  -13.88479   5.657573    -2.45   0.014    -24.99258      -2.777
        male |   90.96919   34.27441     2.65   0.008     23.67657    158.2618
       _cons |   3747.517   81.00609    46.26   0.000     3588.474     3906.56
------------------------------------------------------------------------------
```

C8. For STATA program key see **Exercise_C8_Ch7_Key.do**

The code is

```
* Use Loanapp.dta
* Simplest regression without additional controls
regress approve white
* Regression with controls
regress approve white hrat obrat loanprc unem male married dep sch cosign chist pubrec mortlat1 mortlat2 vr
* Regression with controls plus interaction term.
generate obrat_white = obrat*white
regress approve white hrat obrat loanprc unem male married dep sch cosign chist pubrec mortlat1 mortlat2 vr obrat_white
```

(i) The coefficient would be positive and significant.

(ii) Here is the simple LPM:

```
. regress approve white
```

Source	SS	df	MS	Number of obs	=	1,989
				F(1, 1987)	=	102.23
Model	10.4743407	1	10.4743407	Prob > F	=	0.0000
Residual	203.59303	1,987	.102462521	R-squared	=	0.0489
				Adj R-squared	=	0.0485
Total	214.067371	1,988	.107679764	Root MSE	=	.3201

approve	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
white	.2005957	.01984	10.11	0.000	.1616864	.239505
_cons	.7077922	.0182393	38.81	0.000	.6720221	.7435623

The coefficient on the white indicator variable is positive and statistically significant, preliminarily indicating discrimination in the loan market, **but**, we don't have any additional control variables. The coefficient of 0.20 is pretty large indicating that in going from a non-white applicant to a white applicant, the probability of approval increases by 20%.

(ii) Here is the LPM regression with the additional control variables:

```
. regress approve white hrat obrat loanprc unem male married dep sch cosign chist pubrec mortlat1 mortlat2 vr
```

Source	SS	df	MS	Number of obs	=	1,971
				F(15, 1955)	=	25.86
Model	35.4004787	15	2.36003192	Prob > F	=	0.0000
Residual	178.393534	1,955	.09124989	R-squared	=	0.1656
				Adj R-squared	=	0.1592
Total	213.794013	1,970	.10852488	Root MSE	=	.30208

approve	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
white	.1288196	.0197317	6.53	0.000	.0901223	.1675169
hrat	.001833	.0012632	1.45	0.147	-.0006444	.0043104
obrat	-.0054318	.0011018	-4.93	0.000	-.0075926	-.003271
loanprc	-.1473001	.0375159	-3.93	0.000	-.2208755	-.0737246
unem	-.0072989	.003198	-2.28	0.023	-.0135708	-.0010271
male	-.0041441	.0188644	-0.22	0.826	-.0411405	.0328523
married	.0458241	.0163077	2.81	0.005	.0138418	.0778064
dep	-.0068274	.0067013	-1.02	0.308	-.0199699	.0063151
sch	.0017525	.0166498	0.11	0.916	-.0309006	.0344057
cosign	.0097722	.0411394	0.24	0.812	-.0709094	.0904538
chist	.1330267	.0192627	6.91	0.000	.0952492	.1708043
pubrec	-.2419268	.0282274	-8.57	0.000	-.2972858	-.1865677
mortlat1	-.0572511	.050012	-1.14	0.252	-.1553336	.0408314
mortlat2	-.1137234	.0669838	-1.70	0.090	-.2450905	.0176438
vr	-.0314408	.0140313	-2.24	0.025	-.0589586	-.0039229
_cons	.9367312	.0527354	17.76	0.000	.8333077	1.040155

The coefficient on the white indicator variable is still positive and statistically significant but has fallen in magnitude to 0.128 from 0.20. There is still evidence of discrimination against non-whites.

(iii) Here is the LPM regression with the additional interaction term obrat_white.

```
. regress approve white hrat obrat loanprc unem male married dep sch cosign chist pubrec mortlat1 mortlat2 vr obrat_white
```

Source	SS	df	MS	Number of obs	=	1,971
				F(16, 1954)	=	25.17
Model	36.5318071	16	2.28323794	Prob > F	=	0.0000
Residual	177.262206	1,954	.090717608	R-squared	=	0.1709
				Adj R-squared	=	0.1641
Total	213.794013	1,970	.10852488	Root MSE	=	.30119

approve	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
white	-.1459751	.080263	-1.82	0.069	-.3033851	.0114349
hrat	.0017897	.0012596	1.42	0.156	-.0006806	.0042599
obrat	-.0122262	.0022155	-5.52	0.000	-.0165713	-.0078812
loanprc	-.1525356	.0374357	-4.07	0.000	-.2259537	-.0791175
unem	-.0075281	.0031893	-2.36	0.018	-.0137829	-.0012733
male	-.0060154	.0188167	-0.32	0.749	-.0429184	.0308875
married	.0455358	.0162603	2.80	0.005	.0136465	.0774251
dep	-.00763	.0066856	-1.14	0.254	-.0207417	.0054817
sch	.0017766	.0166011	0.11	0.915	-.0307812	.0343344
cosign	.0177091	.0410807	0.43	0.666	-.0628576	.0982757
chist	.1298548	.0192274	6.75	0.000	.0921464	.1675632
pubrec	-.240325	.0281486	-8.54	0.000	-.2955296	-.1851205
mortlat1	-.0627819	.0498906	-1.26	0.208	-.1606262	.0350624
mortlat2	-.1268446	.0668914	-1.90	0.058	-.2580306	.0043414
vr	-.0305396	.0139926	-2.18	0.029	-.0579816	-.0030975
obrat_white	.0080879	.0022903	3.53	0.000	.0035963	.0125796
_cons	1.180648	.0868076	13.60	0.000	1.010403	1.350894

Then

$$\frac{d(approve)}{d(white)} = \frac{d(approve)}{d(white)}(-0.1459751white + 0.0080879obrat * white)$$

$$= -0.1459751 + 2 * 0.0080879obrat$$

Then if *obra*t = 32, we have $\frac{d(approve)}{d(white)} = 0.1128377$.

The standard error of this marginal effect when *obrat* = 32 is

$$sqrt\left(2^2 32^2 [se(\beta(\widehat{obrat}))]^2\right) = 64 * 0.0022155 = 0.141792.$$

Therefore, the 95% confidence interval of this marginal effect of being white is

$$0.1128371 \pm 1.96 * 0.141792 = [-0.165, 0.39]$$

The 95% confidence interval for the effect of being white encompasses zero and, therefore, for the case *where obrat* = 32, there appears to be no significant discrimination against non-whites.