Assignment #2 KEY

Question 1. The data set CEOSAL1.DTA contains information on 209 CEOs for the year 1990; these data were obtained from Business Week (5/6/1991). To study effect of firm performances and types of industry where CEOs work on CEO compensation, the CEO salary regression is proposed as follows:

$$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 \mathrm{ROE}_i + \beta_3 finance_i + \beta_4 consprod_i + \beta_5 utility_i + u_i$$

where $\log(salary_i)$ = logarithm of CEO annual salary (in 1,000 USD)
$\log(sales_i)$ = logarithm of firms' sale (in 1 million USD)
ROE_i = average return on equity for the CEO's firm for the previous three years (Return on equity is defined in terms of net income as a percentage of common equity)
$finance_i$ = 1 if in financial industry, = 0 otherwise
$consprod_i$ = 1 if in consumer product industry, = 0 otherwise
$utility_i$ = 1 if in utility industry, = 0 otherwise

($finance_i$, $consprod_i$, and $utility_i$ are binary variables indicating the financial, consumer products, and utilities industries. The omitted industry is transportation.

Using STATA, the estimation result is shown below. Answer the following questions.

```
      Source |       SS       df       MS              Number of obs =     209
-------------+------------------------------           F(  5,   203) =   22.53
       Model |  23.8109943     5  4.76219887           Prob > F      =  0.0000
    Residual |  42.9111689   203  .211385068           R-squared     =  0.3569
-------------+------------------------------           Adj R-squared =  0.3410
       Total |  66.7221632   208  .320779631           Root MSE      =  .45977

------------------------------------------------------------------------------
     lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lsales |   .2571917   .0320348     8.03   0.000     .0194282    .3203553
         roe |   .0111517   .3342996     2.59   0.010     .0026742    .0196293
     finance |   .1579564   .0890017     1.77   0.077    -.0175299    .3334426
    consprod |   .1808917   .0847683     2.13   0.034     .0137524    .3480311
     utility |  -.2830015   .0992337    -2.85   0.005    -.4786624   -.0873405
       _cons |   4.588101   .2950221    15.55   0.000       4.0064    5.169801
------------------------------------------------------------------------------
```

a. Write out the estimated regression equation for $\log(salary_i)$. Interpret the estimated coefficient associated with $\log(sales_i)$.

$$\widehat{\log(salary_i)} = 4.5881 + 0.2572 \log(sales_i) + 0.0112\mathrm{ROE}_i + 0.1580 finance_i + 0.1809 consprod_i - 0.2830 utility_i + \hat{u}_i$$

› Both terms (lsalary and lsales) are in natural logarithmic form, therefore, the estimated coefficient can be interpreted as elasticity or percentage change.
› For this case, we can say that when firms' sale increase by 1 percent, CEO salary increases by 0.257 percent in average.

b. What is the overall significance of the regression? What test do you use? Which of the coefficients are individually statistically significant at the 5 percent level? State **the critical value** for hypothesis testing to receive full points.

To test the overall significance, we rely on F-test.
› H0: All the β_k are simultaneously equal to zero.
Ha: Otherwise.
› Calculated F-stat is $F_{cal} = \frac{ESS/k-1}{RSS/n-k} = \frac{23.8109943/5}{42.9111689/203} = 22.528549$
› Critical value for $F_{cri\ (0.05;5,203)} \approx 2.26$
› Since $F_{cal} > F_{cri}$, therefore we can reject the null hypothesis or we can say that all the coefficients are not simultaneously zero.

c. Compute the approximate percentage difference in estimated salary between the utility and transportation sector, holding $sales_i$ and ROE$_i$ fixed.

› This topic was NOT COVERED in class. However, please learn that there are two ways to achieve this.

1) Approximate interpretation: We can rely on the utility coefficient directly. This is a log-lin model so to interpret how much the CEO in utility sector earns compared to the transport sector (reference case) is to find 100 x (-0.2830) = -28.30 percent. In other words, when sales and ROE are kept constant, CEOs in utility sector earn less than CEOs in transport sector in average 28.3 percent.

2) Exact interpretation: Holding others constant, consider transport and utility sector as follows.
- Transport sector (reference case or all dummies are 0): $\log(\widehat{salary_i}|trans) = 4.5881$
- Utility sector: $\log(\widehat{salary_i}|util) = 4.5881 - 0.2830$
- To find the difference, we must anti-log

$$\frac{e^{4.5881} - e^{4.5881-0.2830}}{e^{4.5881}} = 1 - \frac{e^{4.5881-0.2830}}{e^{4.5881}} = 1 - (1)e^{-0.2830}$$

- We can simplify this difference using

$$1 - e^{\beta_k} \text{ or } e^{\beta_k} - 1$$

- Then, if we want to find the difference in percentage term, we multiply the result with 100, we get 100 x $(e^{-0.2830} - 1) = -24.65$

- So, the CEO in utility sector gains less than CEO in transportation sector, in average, about 24.65 percent.

d. Why can't we put all the sector dummies (i.e. $finance_i$, $consprod_i$, $utility_i$ and $transport_i$) in the equation? What would happen if we put all the sector dummies in the equation and use STATA run the regression anyway?

› We can’t do so since adding all the dummies into the model will cause perfect collinearity. STATA will reject one of the dummies (randomly).

e. In the above model, is there any benefit if we add interaction terms between roe and sector dummies, i.e. $\text{ROE}_i * finance_i$ and/or $\text{ROE}_i * consprod_i$ and/or $\text{ROE}_i * utility_i$?

› We have to see if the added interaction terms are statistically significant from zero or not. If they are, it can reveal the different effects that ROE has on CEO earning between sectors.

Question 2. Birth weight has been used by officials as one of the main determinants of health. Data set BWGHT.DTA contains data on infant birth weights in ounces ($bwght_i$), average number of cigarettes mother smoked per day during pregnancy (cigs), family income ($faminc_i$), father's year of education ($fathereduc_i$), and mother's year of education ($mothereduc_i$). The following two regressions were estimated using data on n = 1191 births:

Model 2.1: $bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + u_i$

```
regress  bwght  cigs   faminc

    Source |       SS       df       MS              Number of obs =    1191
-----------+------------------------------           F(  2,  1188) =   18.44
     Model |  14536.9538     2  7268.47691           Prob > F      =  0.0000
  Residual |  468209.738  1188  394.115941           R-squared     =  0.0301
-----------+------------------------------           Adj R-squared =  0.0285
     Total |  482746.692  1190  405.669489           Root MSE      =  19.852

------------------------------------------------------------------------------
     bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
      cigs |  -.5876985   .1090181
    faminc |   .0624684   .0324438       Omitted for the purpose of this exam.
     _cons |   118.5568   1.234278
------------------------------------------------------------------------------
```

Model 2.2: $bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 faminc_i + \beta_3 fathereduc_i + \beta_4 mothereduc_i + u_i$

```
regress  bwght  cigs   faminc fatheduc motheduc

    Source |       SS       df       MS              Number of obs =    1191
-----------+------------------------------           F(  4,  1186) =   10.05
     Model |  15827.6593     4  3956.91482           Prob > F      =  0.0000
  Residual |  466919.033  1186   393.69227           R-squared     =  0.0328
-----------+------------------------------           Adj R-squared =  0.0295
     Total |  482746.692  1190  405.669489           Root MSE      =  19.842

------------------------------------------------------------------------------
     bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
      cigs |  -.5894954   .1106172       Omitted for the purpose of this exam.
    faminc |   .0538254   .0366502
  fatheduc |   .4936695   .2832896
  motheduc |  -.4379234   .3197377
     _cons |   118.0741   3.500291
------------------------------------------------------------------------------
```

where $bwght_i$ = birth weight, ounces
$cigs_i$ = average number of cigarettes the mother smoked per day while pregnant
$faminc_i$ = 1988 family income, \$1000s
$fathereduc_i$ = father's years of education
$mothereduc_i$ = mother's years of education

Answer the following questions.

a. Based on **Model 2.1**, test whether smoking has an impact on birth weight. Show your work. (use $\alpha = 0.05$)

› $t_{cal} = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = -\frac{0.5876985}{0.1090181} = -5.39083418$
› $t_{cri} = \pm 1.96$; we can reject the null hypothesis or β_2 (smoking) is significantly different from zero.

b. Based on **Model 2.1**, construct a 99% confidence interval for β_2.

› The 99% CI is $P\left(\hat{\beta}_2 - t_{\frac{\alpha}{2}} \cdot se(\hat{\beta}_2) > \beta_2 > \hat{\beta}_2 + t_{\frac{\alpha}{2}} \cdot se(\hat{\beta}_2)\right) = 0.99$
› $t_{\frac{\alpha}{2}} = 2.576$ therefore
› The lower bound is $\hat{\beta}_2 - t_{\frac{\alpha}{2}} \cdot se(\hat{\beta}_2) = 0.0624684 - (2.576 * 0.0324438) = -0.021106$
› The upper bound is $\hat{\beta}_2 - t_{\frac{\alpha}{2}} \cdot se(\hat{\beta}_2) = 0.0624684 + (2.576 * 0.0324438) = 0.1460436288$

c. Would your conclusion in a) change if you use the result from **Model 2.2**? Show your work. (use $\alpha = 0.05$)

› No, because $t_{cal} = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = -\frac{0.5894954}{0.1106172} = -5.32914773$ which is still exceeding critical value.

d. What is the overall significance of the regression from **Model 2.2**? What test do you use? Which of the coefficients are individually statistically significant at the 5 percent level? State the critical value for hypothesis testing to receive full points.

› **Testing the overall significance relies on F-test.** The hypotheses are
› H0: All the β_k are simultaneously equal to zero.
Ha: Otherwise.
› Calculated F-stat is $F_{cal} = \frac{ESS/df}{RSS/df} = \frac{15{,}827.6593/4}{466{,}919.033/1186} = 10.05078108$
› Critical value for $F_{cri\,(0.05;4,1186)} = 2.37$
› Since $F_{cal} > F_{cri}$, therefore we can reject the null hypothesis or we can say that all the coefficients are not simultaneously equal to zero.

› **Testing the individual significance relies on t-test.** The hypotheses are
› H0: $\beta_k = 0$
Ha: $\beta_k \neq 0$; for every coefficient
› $t_{cri} = \pm 1.96$; as usual
› $t_{cal} = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{118.0741}{3.500291} = 33.7326525$; reject H0
› $t_{cal} = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} = -\frac{0.5894954}{0.1106172} = 5.32914773$; reject H0
› $t_{cal} = \frac{\hat{\beta}_3 - \beta_3}{se(\hat{\beta}_3)} = \frac{0.538254}{0.366502} = 1.4693617$; cannot reject H0

› $t_{cal} = \frac{\widehat{\beta}_4 - \beta_4}{se(\widehat{\beta}_4)} = \frac{0.4936695}{0.2832896} = 1.74263192$; cannot reject H_0

› $t_{cal} = \frac{\widehat{\beta}_5 - \beta_5}{se(\widehat{\beta}_5)} = -\frac{0.4379234}{0.3197377} = -1.369633$; cannot reject H_0

Only the constant and smoking coefficient that is statistically different from zero.

e. If we are interested in testing whether "**parents' education**" has an impact on birth weight at all, what kind of null/alternative hypothesis would we be testing? Perform the test and discuss your finding. (use $\alpha = 0.05$)

› **Testing for the marginal contribution of "parents' education" relies on F-test.** The hypotheses are

› H_0: $\beta_4 = \beta_5 = 0$

H_a: Otherwise.

› $F_{cal} = \frac{R^2_{new} - R^2_{old} / number\ of\ new\ regressor}{1 - R^2_{new} / n - k_{new}} = \frac{0.0328 - 0.0301/2}{1 - 0.0328/1191 - 5} = 1.65539702$

› Critical value for $F_{cri\ (0.05;2,1186)} = 3$

› We cannot reject the null hypothesis, therefore, we cannot make sure that, 95 percent of the time, parent's education has an impact on birth weight.

Question 3. A model of wage equation is given by

$$lawge_i = \beta_1 + \beta_2 exp_i + +\beta_3 expsq_i + \beta_4 educ_i + \beta_5 age_i + \beta_6 kid6_i + \beta_7 kid18_i + u_i$$

where $lwage_i$ = natural log of hourly wage of married women
exp_i = years of experience
$expsq_i$ = years of experience squared
$educ_i$ = years of education
age_i = age
$kid6_i$ = number of children aged 0-6 in a household
$kid18_i$ = number of children aged 6-18 in a household

The regression result from OLS is shown in the table below and answer the following questions.

```
   Source |       SS           df       MS            Number of obs =      428
----------+------------------------------            F(___,______) =    13.19
    Model |  __________       ___  __________         Prob > F      =   0.0000
 Residual |  __________       ___  .446526442         R-squared     =   0.1582
----------+------------------------------            Adj R-squared =   ______
    Total |  223.327441       ___  __________         Root MSE      =   .66823

---------------------------------------------------------------------------
    lwage |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
----------+----------------------------------------------------------------
    exper |    .039819    .013393     2.97   0.003     .0134936    .0661444
  expersq |  -.0007812   .0004022    -1.94   0.053    -.0015718    9.37e-06
     educ |   .1078319   .0144021     7.49   0.000      .079523    .1361409
      age |  -.0014653   .0052925    -0.28   0.782    -.0118682    .0089377
  kidslt6 |  -.0607106   .0887626    -0.68   0.494    -.2351836    .1137625
  kidsge6 |   -.014591   .0278981    -0.52   0.601     -.069428    .0402459
    _cons |  -.4209078    .316905    -1.33   0.185    -1.043821    .2020053
---------------------------------------------------------------------------
```

a) Figure out all the degrees of freedom in this model.

› DF for Model part is k-1. k is 7 so it is 6.
› DF for Residual part is n-k. n is 428 so it is 421.
› DF for Total part is total DF which is 421+6 = 427.

b) Figure out all the sum of squares (ESS and RSS) and mean squares in this model.

› MS is SS/DF so Total MS = 223.327441 / 427 = 0.523015084.
› We can calculate residual SS from MS * DF so RSS = 0.446526442 * 421 = 187.987632.
› ESS is TSS – RSS so it is 223.327441 – 187.987632 = 35.339809.
› Explained MS is ESS / DF = 35.339809 / 6 = 5.88996817.

c) Figure out the adjusted R-squared ($\bar{R}^2$)

› Adjusted R-square can be retrieved from $\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} = 0.1462$

d) Given that the model above is called '**Model 3.1**', there is another competing model called '**Model 3.2**' which **an explanatory variable is excluded**, compared to '**Model 3.1**'. Though the result of estimating '**Model 3.2**' is not shown here, **what is the maximum value of R^2 from 'Model 3.2'** which will make you conclude that the excluded variable has a significant contribution in '**Model 3.1**', at the significance level of 0.05. (**Hint:** the critical value of the F-test at the significance level of 0.05 is $F_{1,421} = 3.84$)

› The essence here is to figure out that the excluded variable has any marginal contribution to this Model 3.1 or not.

› We firstly define Model 3.1 as "New model" and Model 3.2 as "Old model" according to marginal contribution chapter in the book. (The "New" model has more variables compared to the "Old" model)

› In order for the excluded variable to have marginal contribution $F_{cal} > F_{cri}$, or $F_{cal} > 3.84$.

› Comparing model 3.1 and 3.2 using F-test

› $F_{cal} = \frac{R^2_{new} - R^2_{old} / number\ of\ new\ regressor}{1 - R^2_{new} / n - k_{new}} > 3.84$; replacing known values

› $\frac{0.1582 - R^2_{old}/1}{1 - 0.1582/428 - 7} > 3.84$; solve the equation to get R^2_{old}

› $-R^2_{old} > [3.84 * 0.001995] - 0.1582$

› $-R^2_{old} > -0.1505$

› $R^2_{old} < 0.1505$ Therefore, this is the maximum value of R^2 from Model 3.2.

e) As you can see from the result, age is not significantly different from zero. In other words, age does not determine how much hourly wage would be. Does this make economic sense in your opinion? What do you think cause this insignificance?

› It is very likely that age and experience may be linearly correlated. Though age is supposed to be significant determining wage by economic sense, it is not which may be due to multicollinearity problem. We might say that age does not add any further information to this model since experience and square of experience already do.