Assignment 1: Due Oct 9 (Monday) before Midnight

Name: Abhishek Arya

NETID: axa220149

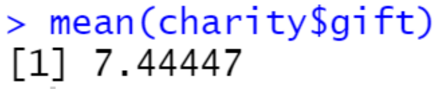
PROGRAMMING LANGUAGE USED = R

Directions: Respond appropriately to the following questions. Upload your final assignment onto e-learning by the due date. Insert all tables and images, if any, into your word file (or pdf) so that answers are all in one place. Copy and paste your code (no matter what statistical software you are using, including STATA) at the end of the assignment or submit a separate file. Part of your grade (5 points) will be based on the code, and the remaining will be based on your ability to follow directions and *fully* explain econometric models (75 points). This is an individual assignment. You must turn in your own word document (or pdf). Late submissions within 24 hours will receive 50% of the original points, late submissions within 48 hours will receive 25%, and so on.

1. (Lecture 2) (12 points) Use the data in CHARITY [obtained from Franses and Paap (2001)] to answer the following questions:

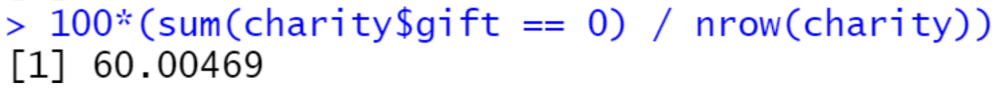
   i. (2 points) What is the average gift in the sample of 4,268 people (in Dutch guilders)? What percentage of people gave no gift?

   Ans : average gift in the sample of 4,268 people = 7.44



```
> mean(charity$gift)
[1] 7.44447
```

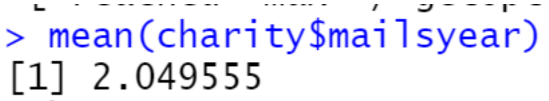
   Percentage of prople gave on gift = 60.00469%



```
> 100*(sum(charity$gift == 0) / nrow(charity))
[1] 60.00469
```

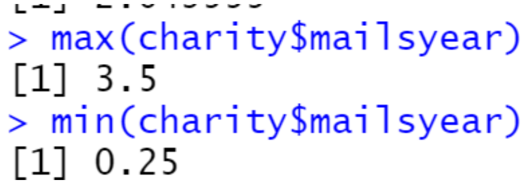
   ii. (2 points) What is the average mailing per year? What are the minimum and maximum values?

   Ans: Average mailings per year : 2.04955



```
> mean(charity$mailsyear)
[1] 2.049555
```

   Minimum and maximum values = 3.5, 0.25



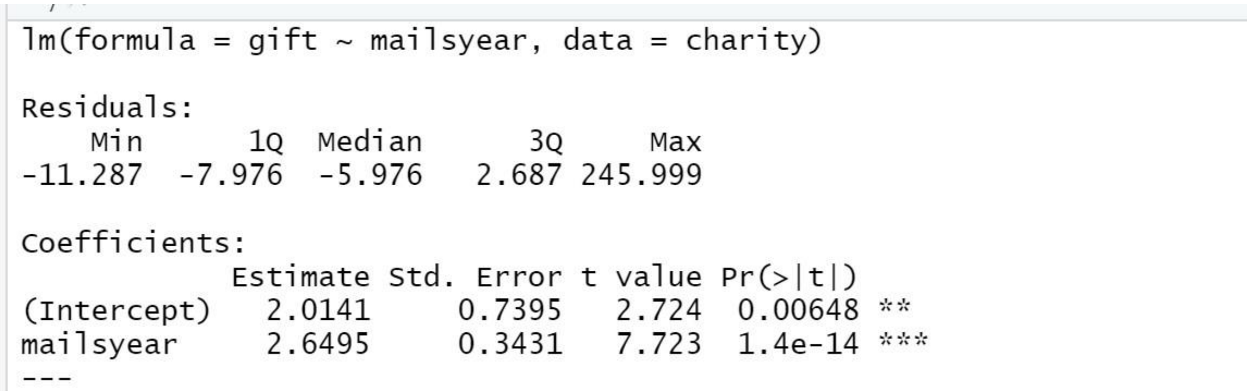
```
> max(charity$mailsyear)
[1] 3.5
> min(charity$mailsyear)
[1] 0.25
```

   iii. (2 points) Estimate the model

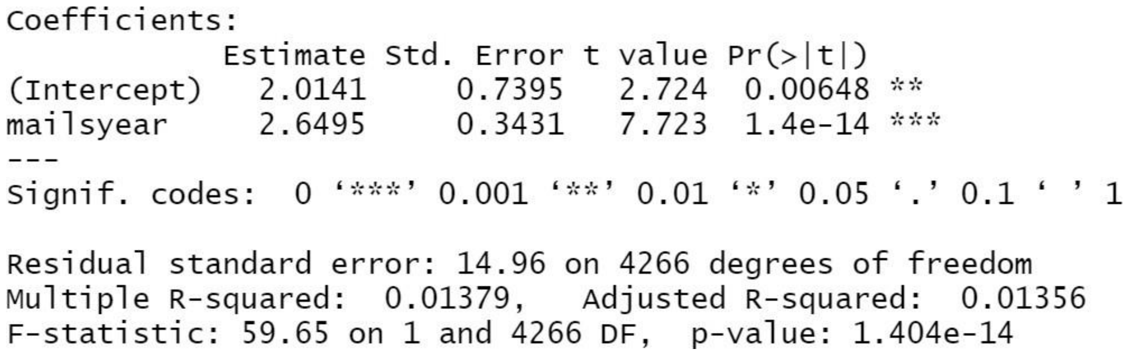
$$gift = \beta_0 + \beta_1 mailsyear + u$$

   By OLS and report the results in the usual way (as shown in the lectures), including the sample size and R-squared.

Ans: Linear regression summary



```
lm(formula = gift ~ mailsyear, data = charity)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.287  -7.976  -5.976   2.687 245.999 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.0141     0.7395   2.724  0.00648 ** 
mailsyear     2.6495     0.3431   7.723  1.4e-14 ***
---
```



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.0141     0.7395   2.724  0.00648 ** 
mailsyear     2.6495     0.3431   7.723  1.4e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.96 on 4266 degrees of freedom
Multiple R-squared:  0.01379,	Adjusted R-squared:  0.01356 
F-statistic: 59.65 on 1 and 4266 DF,  p-value: 1.404e-14
```

Sample size = df(degrees of freedom) + 2 = 4266 + 2 = 4268R-squared = 0.01379

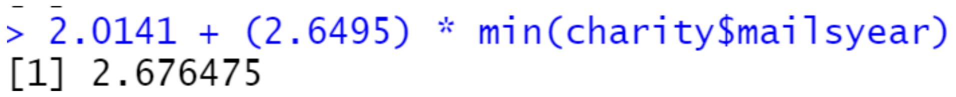
(4 points) Interpret the slope coefficient. If each mailing costs one guilder, is the charity expected to make a net gain on each mailing? Does this mean the charitymakes a net gain on every mailing? Explain.

Ans: The slope coefficient for mailsyear = 2.6495. For each additional mailing sent peryear, the charity can expect an increase of approximately 2.6495 Dutch guilders in theaverage gift received from donors. Yes, the charity makes a net gain of 2.64 on every mailing. Since cost is 1, the net surplus is 2.649-1 = 1.649. So, each mailing costs one guilder then the expected profit from each mailing is estimated to be 1.65 guilders.

But some mailings generate less than the mailing cost; other mailings generated muchmore than the mailing cost.

(2 points) What is the smallest predicted charitable contribution in the sample?Using this simple regression analysis, can you ever predict zero for gift?

Ans: smallest predicted charitable contribution in the sample = 2.0141(intercept) + (2.6495)[slope] * (min(mailsyear))



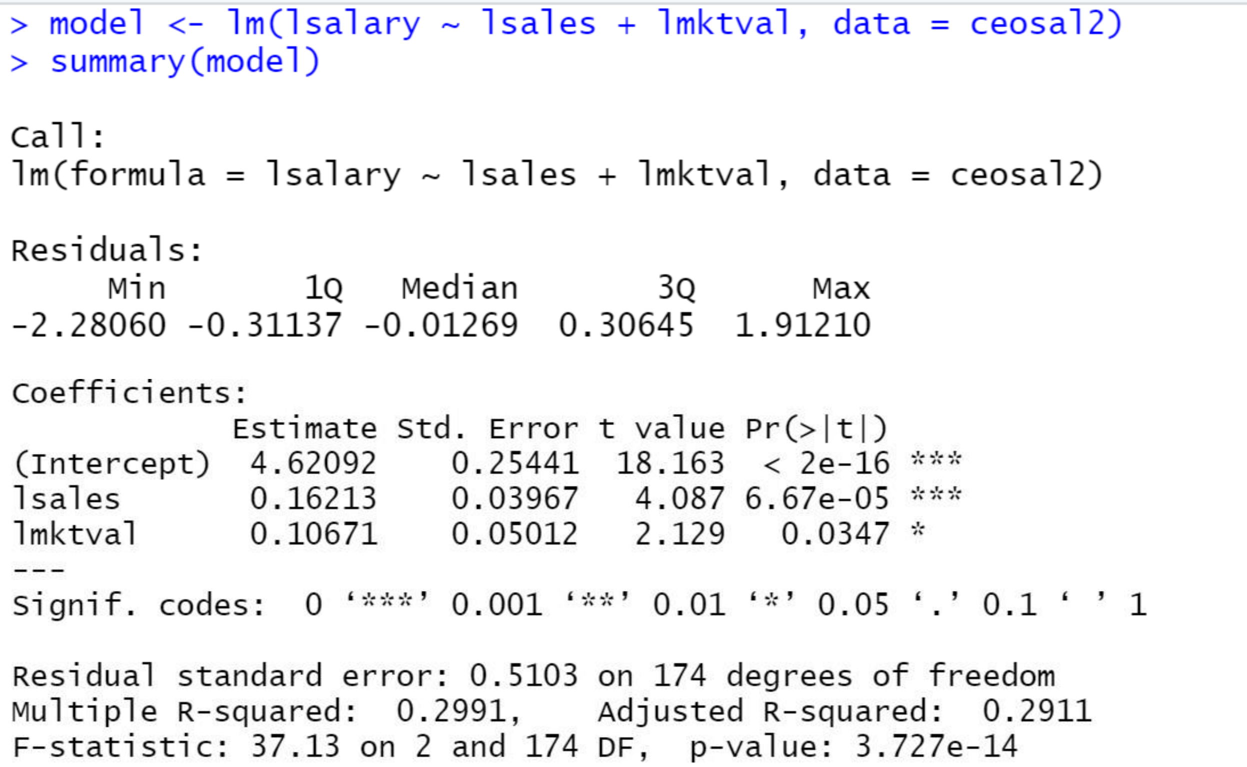
```
> 2.0141 + (2.6495) * min(charity$mailsyear)
[1] 2.676475
```

No we can never predict zero for gift. If mailsyear = 0 then the value of gifts will be equalto the intercept which is 2.014. Therefore, with this estimated equation, we can never predict zero for gifts.

2.(Lecture 3) (14 points) The file CEOSAL2 contains data on 177 chief executive officers andcan be used to examine the effects of firm performance on CEO salary.

(3 points) Estimate a model relating annual salary to firm sales and market value. Make the model of the constant elasticity variety for both independent variables.Report the results in the usual way.

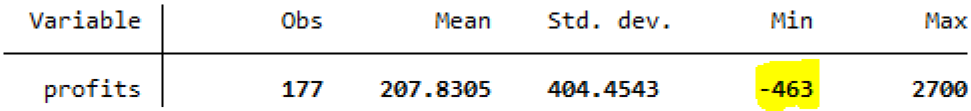
Ans :



```
> model <- lm(lsalary ~ lsales + lmktval, data = ceosal2)
> summary(model)

Call:
lm(formula = lsalary ~ lsales + lmktval, data = ceosal2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.28060 -0.31137 -0.01269  0.30645  1.91210 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.62092    0.25441  18.163  < 2e-16 ***
lsales       0.16213    0.03967   4.087 6.67e-05 ***
lmktval      0.10671    0.05012   2.129   0.0347 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5103 on 174 degrees of freedom
Multiple R-squared:  0.2991,	Adjusted R-squared:  0.2911 
F-statistic: 37.13 on 2 and 174 DF,  p-value: 3.727e-14
```

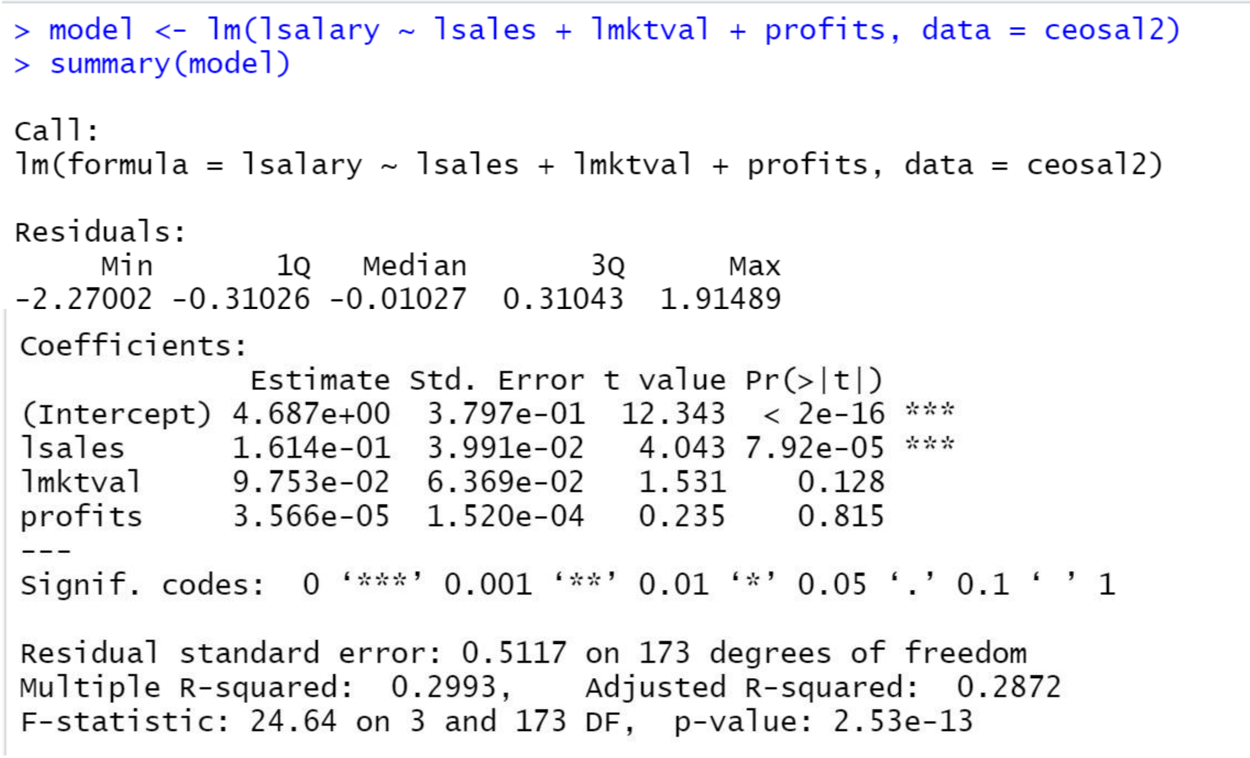
(4 points) Add *profits* to the model from part (i), re-estimate the model andreport the results in the usual way. Why can this variable not be included in

logarithmic form? Would you say that these firm performance variables explainmost of the variation in CEO salaries?
Ans :



```
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
     profits |        177    207.8305    404.4543       -463       2700
```

We cannot do log transformation to profits, since the minimum value present in the sample for profits column is -463, which strictly indicates that there are negative values. And also speaking generally, the variable "profits" cannot be included in logarithmic form because it is possible for a company to have negative profits. Taking the logarithm of a negative number or zero is undefined.

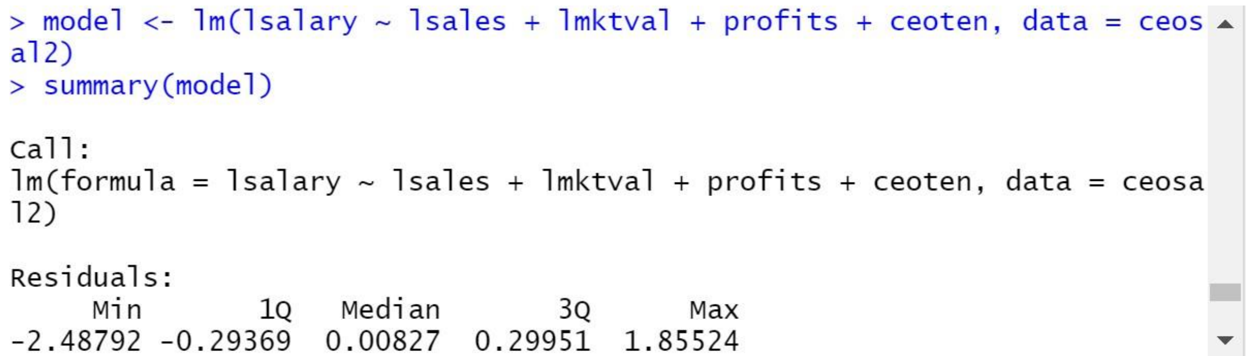


```
> model <- lm(lsalary ~ lsales + lmktval + profits, data = ceosal2)
> summary(model)

Call:
lm(formula = lsalary ~ lsales + lmktval + profits, data = ceosal2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.27002 -0.31026 -0.01027  0.31043  1.91489 
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.687e+00  3.797e-01  12.343  < 2e-16 ***
lsales      1.614e-01  3.991e-02   4.043 7.92e-05 ***
lmktval     9.753e-02  6.369e-02   1.531    0.128    
profits     3.566e-05  1.520e-04   0.235    0.815    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5117 on 173 degrees of freedom
Multiple R-squared:  0.2993,	Adjusted R-squared:  0.2872 
F-statistic: 24.64 on 3 and 173 DF,  p-value: 2.53e-13
```

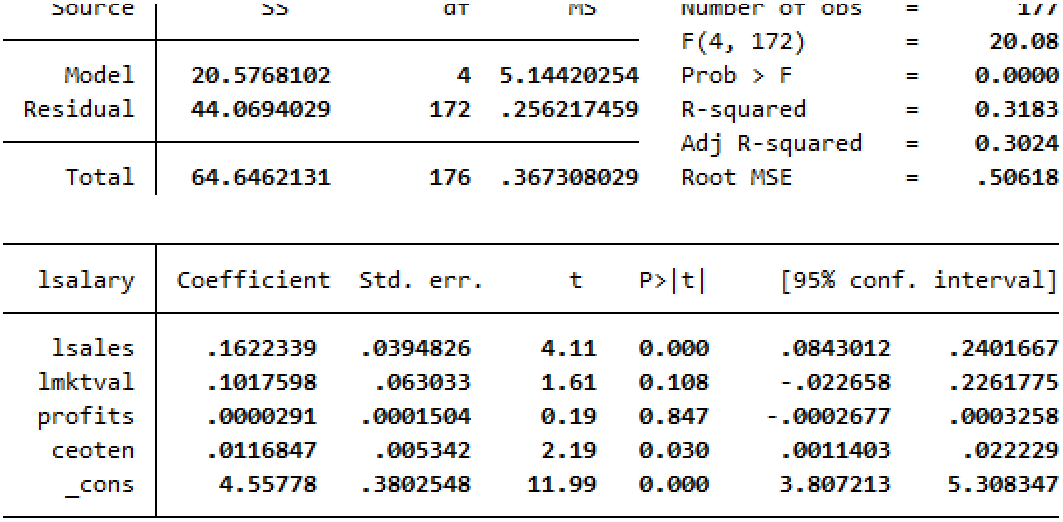
Without adding the profits variable, the adj-r-squared is 0.2911, with adding profitsvariable the adj-r-squared decreased to 0.2972. These firm performance variables explain most of the variation in CEO salaries. The coefficient profit is very small there, profits are measured in millions so if profits increase by $1 billion, the predicted salary increases by only 3.6%, with sales and market value held fixed.

Together, these variables explain almost 30% of the sample variation in log (salary). This iscertainly not most of the variation

(3 points) Add the variable *ceoten* to the model in part (ii), re-estimate the modeland report the results in the usual way. What is the estimated percentage return for another year of CEO tenure, holding other factors fixed?



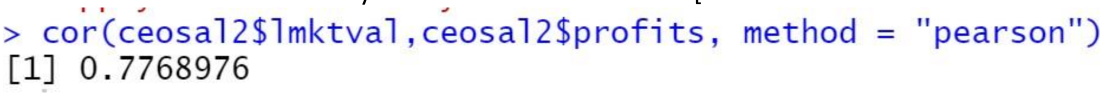
```
> model <- lm(lsalary ~ lsales + lmktval + profits + ceoten, data = ceos
al2)
> summary(model)

Call:
lm(formula = lsalary ~ lsales + lmktval + profits + ceoten, data = ceosa
l2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.48792 -0.29369  0.00827  0.29951  1.85524 
```



```
      Source |       SS           df       MS      Number of obs   =       177
-------------+----------------------------------   F(4, 172)       =     20.08
       Model |  20.5768102         4  5.14420254   Prob > F        =    0.0000
    Residual |  44.0694029       172  .256217459   R-squared       =    0.3183
-------------+----------------------------------   Adj R-squared   =    0.3024
       Total |  64.6462131       176  .367308029   Root MSE        =    .50618

------------------------------------------------------------------------------
     lsalary | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      lsales |   .1622339   .0394826     4.11   0.000     .0843012    .2401667
     lmktval |   .1017598    .063033     1.61   0.108     -.022658    .2261775
     profits |   .0000291   .0001504     0.19   0.847    -.0002677    .0003258
      ceoten |   .0116847    .005342     2.19   0.030     .0011403     .022229
       _cons |    4.55778   .3802548    11.99   0.000     3.807213    5.308347
------------------------------------------------------------------------------
```

The estimated % return for another year of CEO tenure holding other factors fixed isincreased by 1.2% for predicted salary.

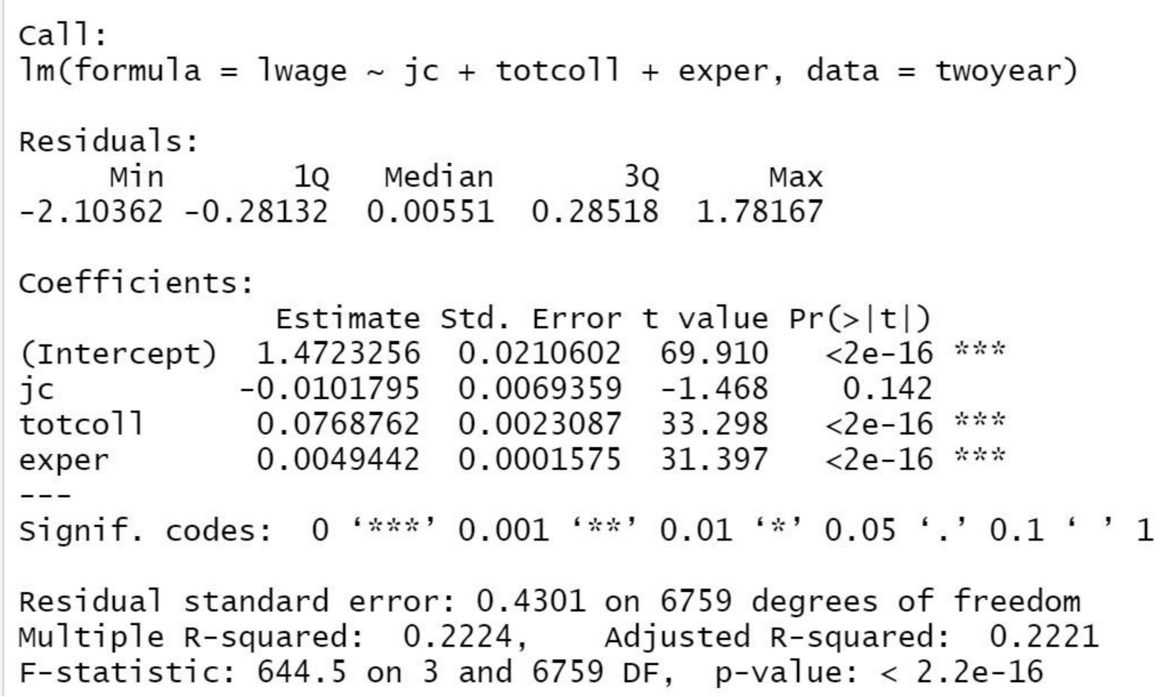
(4 points) Find the sample correlation coefficient between the variables log(*mktval*) and *profits*. Are these variables highly correlated? What does this sayabout the OLS estimators? [Hint: You can use the stata command **correlate**.]



```
> cor(ceosal2$lmktval,ceosal2$profits, method = "pearson")
[1] 0.7768976
```

The correlation for both log(mktval) and profits = 0.77, which is highly positivelycorrelated. High correlation (0.77) between log(mktval) and profits suggests a strong linear relationship between these variables, which can lead to multicollinearity in the above regression model.

3.(Lecture 4) (17 points) Refer to the example used in Lecture 4 to compare the returns toeducation at junior colleges and four-year colleges. The model after rearrangement is
$\log(wage) = \beta_0 + \theta_1 \text{jc} + \beta_2 totcoll + \beta_3 exper + u,$
where *totcoll* is total years of college. Use the data set TWOYEAR, which comes fromKane and Rouse (1995).
i. (5 points) Run the regression above and report the OLS estimates in the usualform, including the standard errors, sample size and R-squared. How do you interpret θ_1? Is it statistically significant?
Ans :

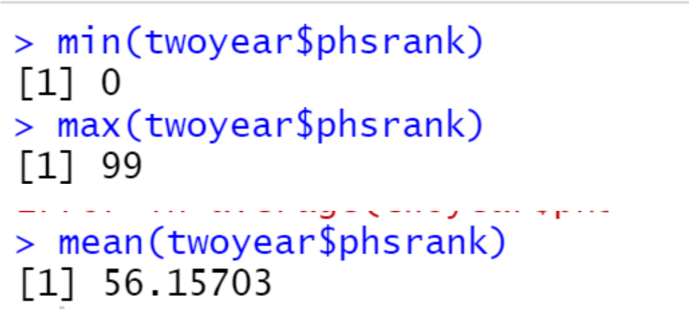


```
Call:
lm(formula = lwage ~ jc + totcoll + exper, data = twoyear)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.10362 -0.28132  0.00551  0.28518  1.78167 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.4723256  0.0210602  69.910   <2e-16 ***
jc          -0.0101795  0.0069359  -1.468    0.142    
totcoll      0.0768762  0.0023087  33.298   <2e-16 ***
exper        0.0049442  0.0001575  31.397   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4301 on 6759 degrees of freedom
Multiple R-squared:  0.2224,	Adjusted R-squared:  0.2221 
F-statistic: 644.5 on 3 and 6759 DF,  p-value: < 2.2e-16
```

Null hypothesis => theta1 = 0 Alternative Hypothesis => theta1 < 0

When coming to theta1, the p-value = 0.142 is greater than 0.05, so we cannot reject nullhypothesis. Theta1 is not statistically significant. We accept null hypothesis, i.e (Beta1 – Beta2) = 0. θ_1 = -0.0101795, This is the difference between the coefficients.P-value of θ_1 is 0.142 > C_0.005. The null hypothesis is true.
So, jc is not statistically significant.

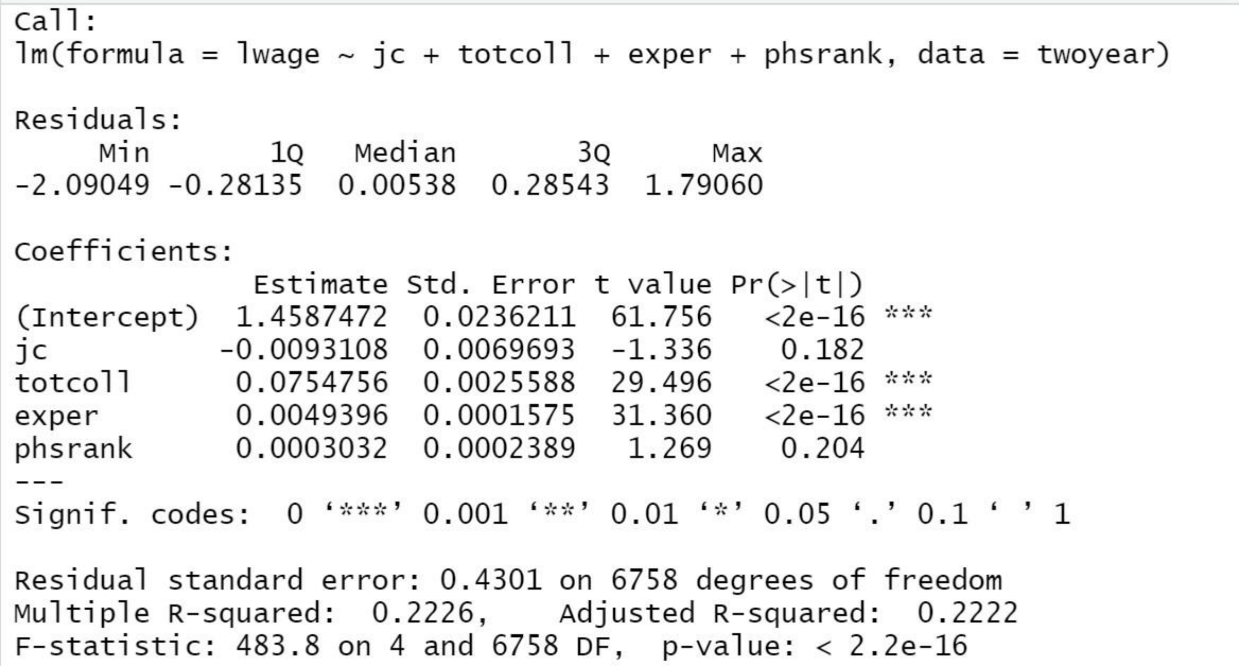
ii .)(2 points) The variable *phsrank* is the person’s high school percentile. (A higher number is better. Forexample, 90 means you are ranked better than 90 percent of your graduating class.) Find the smallest, largest, and average *phsrank* in the sample.
Ans :



```
> min(twoyear$phsrank)
[1] 0
> max(twoyear$phsrank)
[1] 99
> mean(twoyear$phsrank)
[1] 56.15703
```

Smallest = 0, largest = 99, average = 56.15703

iii.)(4 points) Add *phsrank* to the model and report the OLS estimates in the usual form. Is *phsrank* statistically significant? How much is 10 percentage points of high school rank worth in terms of wage?Ans :

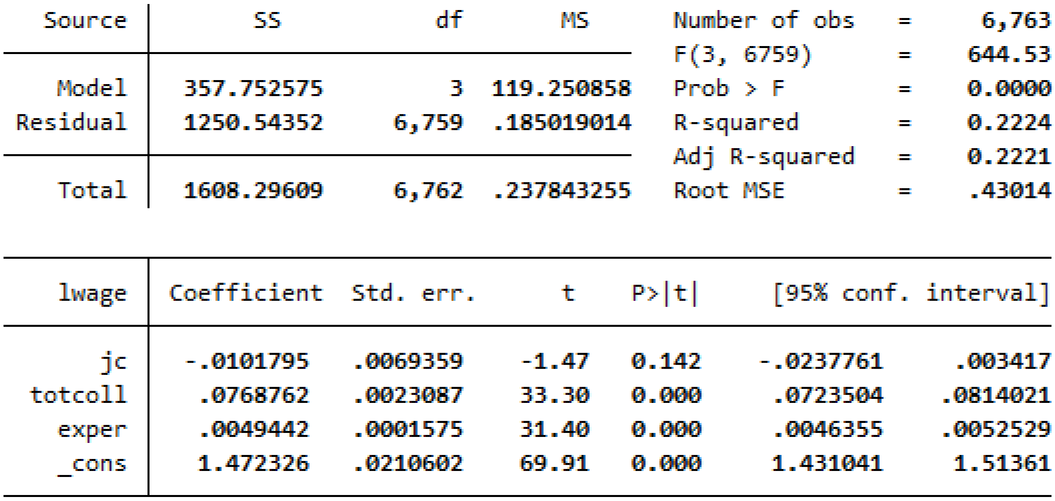


```
Call:
lm(formula = lwage ~ jc + totcoll + exper + phsrank, data = twoyear)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.09049 -0.28135  0.00538  0.28543  1.79060 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.4587472  0.0236211  61.756   <2e-16 ***
jc          -0.0093108  0.0069693  -1.336    0.182    
totcoll      0.0754756  0.0025588  29.496   <2e-16 ***
exper        0.0049396  0.0001575  31.360   <2e-16 ***
phsrank      0.0003032  0.0002389   1.269    0.204    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4301 on 6758 degrees of freedom
Multiple R-squared:  0.2226,	Adjusted R-squared:  0.2222 
F-statistic: 483.8 on 4 and 6758 DF,  p-value: < 2.2e-16
```

The p-value for phsrank is 0.204 which is greater than 0.05, null hypothesis cannot be rejected. Therefore, phsrank is not statistically significant. . If points of high school rank increase by 10% thenthere will be a 0.3% increase in wages.

To estimate how much a 10 percentage point increase in high school rank is worth in terms of wage, youcan multiply the coefficient by 10 (since we are interested in a 10 percentage point change):
Estimated change in wage = 0.0003032 * 10
Estimated change in wage ≈ 0.003032 * 100 = 0.3 % increase in wage

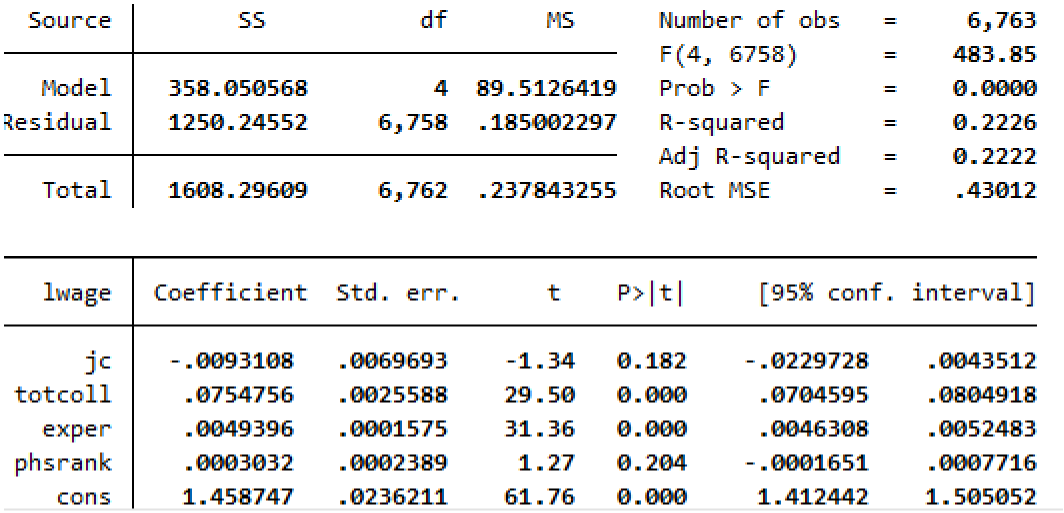
iv.) (3 points) Compare regression results in (i) and (iii), does adding *phsrank* to the model substantivelychange the conclusions on the returns to two- and four-year colleges? Explain.
Ans: Model (i) (Without phsrank):



```
      Source |       SS           df       MS      Number of obs   =     6,763
-------------+----------------------------------   F(3, 6759)      =    644.53
       Model |  357.752575         3  119.250858   Prob > F        =    0.0000
    Residual |  1250.54352     6,759  .185019014   R-squared       =    0.2224
-------------+----------------------------------   Adj R-squared   =    0.2221
       Total |  1608.29609     6,762  .237843255   Root MSE        =    .43014

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          jc |  -.0101795   .0069359    -1.47   0.142    -.0237761     .003417
     totcoll |   .0768762   .0023087    33.30   0.000     .0723504    .0814021
       exper |   .0049442   .0001575    31.40   0.000     .0046355    .0052529
       _cons |   1.472326   .0210602    69.91   0.000     1.431041     1.51361
------------------------------------------------------------------------------
```

totcoll coefficient estimate: 0.0768762
P-value for totcoll: < 2.2e-16 (highly significant)

Model 2 (With phsrank):



```
      Source |       SS           df       MS      Number of obs   =     6,763
-------------+----------------------------------   F(4, 6758)      =    483.85
       Model |  358.050568         4  89.5126419   Prob > F        =    0.0000
    Residual |  1250.24552     6,758  .185002297   R-squared       =    0.2226
-------------+----------------------------------   Adj R-squared   =    0.2222
       Total |  1608.29609     6,762  .237843255   Root MSE        =    .43012

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          jc |  -.0093108   .0069693    -1.34   0.182    -.0229728    .0043512
     totcoll |   .0754756   .0025588    29.50   0.000     .0704595    .0804918
       exper |   .0049396   .0001575    31.36   0.000     .0046308    .0052483
     phsrank |   .0003032   .0002389     1.27   0.204    -.0001651    .0007716
        cons |   1.458747   .0236211    61.76   0.000     1.412442    1.505052
```

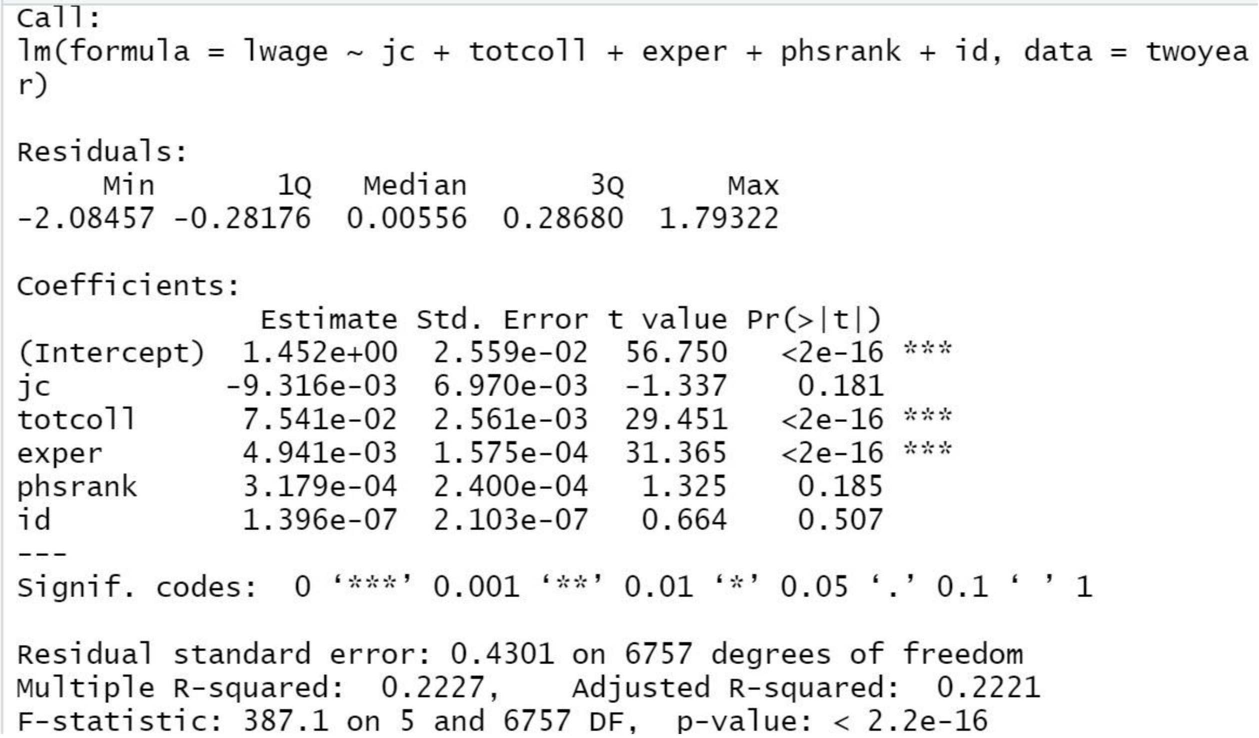
totcoll coefficient estimate: 0.0754756

P-value for totcoll: < 2.2e-16 (highly significant)

In both models, the coefficient estimate for totcoll is highly significant (p-value < 2.2e-16), and theestimated coefficient values are very close. The coefficient for totcoll in Model 2 (with phsrank) is slightly smaller than in Model 1 (without phsrank), but the difference is negligible.
Therefore, adding phsrank to the model does not substantively change the conclusion on the return of two and four-year colleges. T- statistic on jc gets even smaller in absolute value, about 1.33 whenphsrank is added. However, the coefficient magnitude is almost equal. Hence, we can say that the base point remains unchanged. This implies that even though the difference is not significant, and the levels are standard, the return to junior college is estimated to be somewhat smaller.

v.)(3 points) The data set contains a variable called *id*. Explain why if you add *id* to the model you expectit to be statistically insignificant. What is the two-sided p-value?

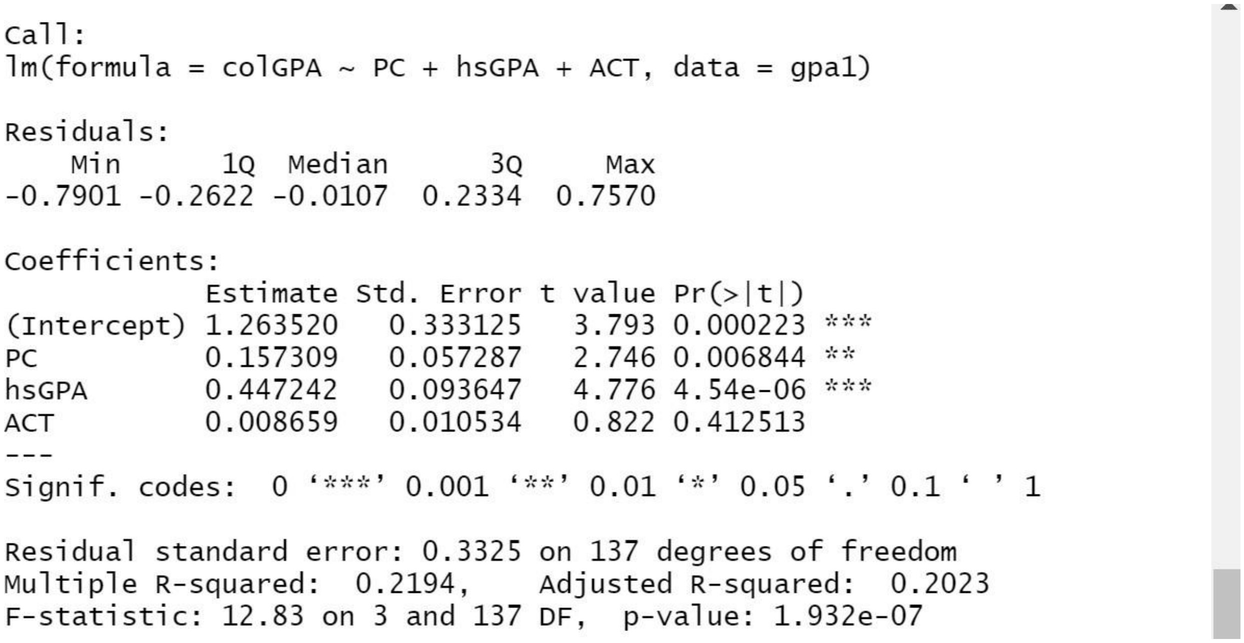
Ans: In most cases, the variable "id" (ID Number) is expected to be statistically insignificant when addedto a regression model for the following reasons:
Unique Identifier, No Meaningful Relationship, Low Variablity



```
Call:
lm(formula = lwage ~ jc + totcoll + exper + phsrank + id, data = twoyea
r)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.08457 -0.28176  0.00556  0.28680  1.79322 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.452e+00  2.559e-02  56.750   <2e-16 ***
jc          -9.316e-03  6.970e-03  -1.337    0.181    
totcoll      7.541e-02  2.561e-03  29.451   <2e-16 ***
exper        4.941e-03  1.575e-04  31.365   <2e-16 ***
phsrank      3.179e-04  2.400e-04   1.325    0.185    
id           1.396e-07  2.103e-07   0.664    0.507    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4301 on 6757 degrees of freedom
Multiple R-squared:  0.2227,    Adjusted R-squared:  0.2221 
F-statistic: 387.1 on 5 and 6757 DF,  p-value: < 2.2e-16
```

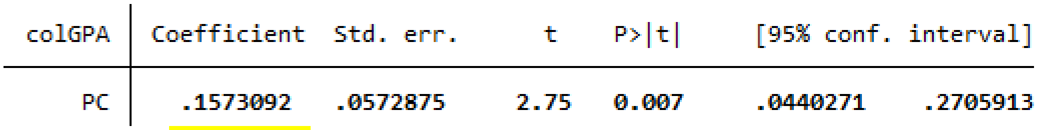
The two-sided p-value = 0.507

4.) i.)(Lecture 4) (9 points) Use the data set GPA1 to answer this question.
(3 points) Run the regression *colGPA* on *PC*, *hsGPA*, and *ACT* and obtain a 95% confidence interval for β_{PC}.Is the estimated coefficient statistically significant at the 5% level against a two-sided alternative?
Ans :



```
Call:
lm(formula = colGPA ~ PC + hsGPA + ACT, data = gpa1)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7901 -0.2622 -0.0107  0.2334  0.7570 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.263520   0.333125   3.793 0.000223 ***
PC          0.157309   0.057287   2.746 0.006844 ** 
hsGPA       0.447242   0.093647   4.776 4.54e-06 ***
ACT         0.008659   0.010534   0.822 0.412513    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3325 on 137 degrees of freedom
Multiple R-squared:  0.2194,	Adjusted R-squared:  0.2023 
F-statistic: 12.83 on 3 and 137 DF,  p-value: 1.932e-07
```

95% Confidence Interval for βPC = 0.157309 ± (tcritical×0.057287). tcritical at 95% level is 1.97



```
      colGPA | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          PC |   .1573092   .0572875     2.75   0.007     .0440271    .2705913
```

Lower bound = 0.04402791

Upper bound = 0.2705901

The p-value is 0.007 < 0.05, therefore it is statistically significant at 5% level

ii.) (3 points) discuss the statistical significance of the estimates $\hat{\beta}_{hsGPA}$ and $\hat{\beta}_{ACT}$ in part (i). Is *hsGPA* or *ACT* the more important predictor of *colGPA*? Explain.

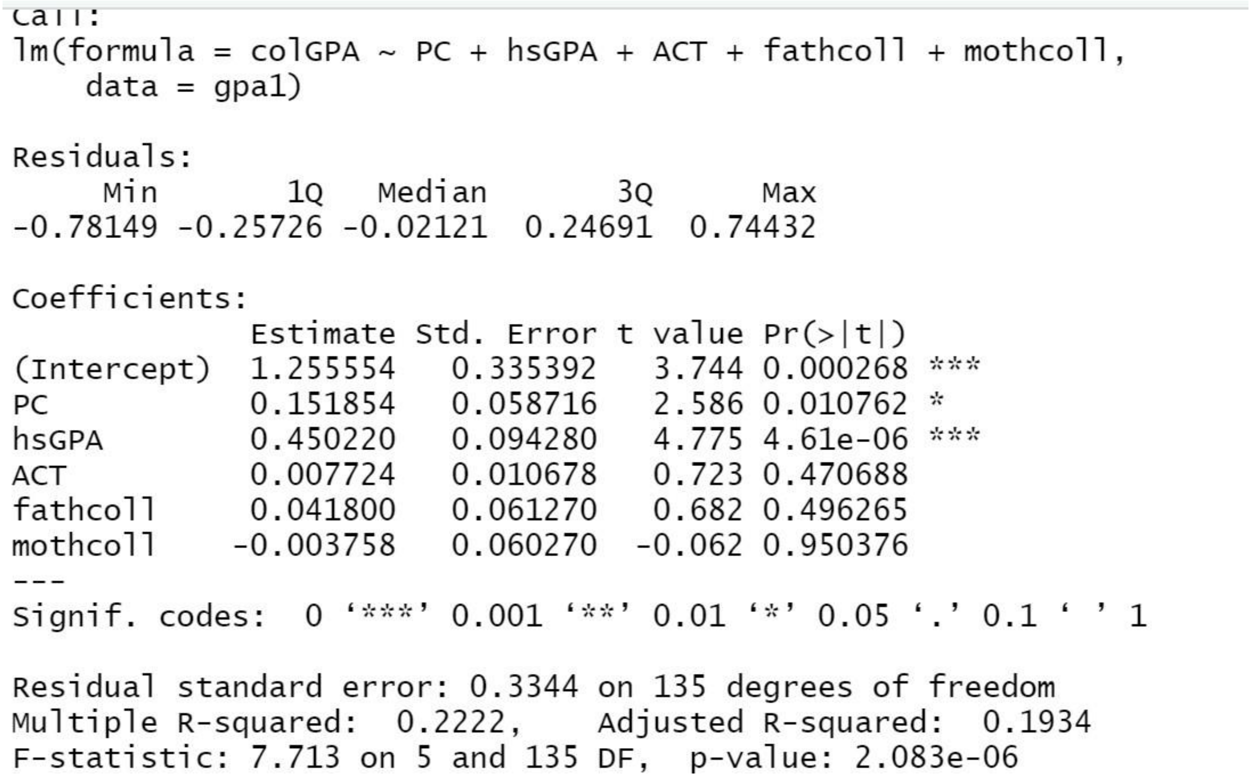
Ans : Betahsgpa p-value is 4.54e-06 < 0.05, therefore it is statistically significant.

BetaACT p-value is 0.412 > 0.05, it is not statistically significant.

hsGPA is the more important predictor of colGPA than ACT, because of its statistical significanceexplained above.

iii.) (3 points) Add the two indicators *fathcoll* and *mothcoll* to the regression in part (i). Is eitherindividually significant? Are they jointly statistically significant at the 5% level?

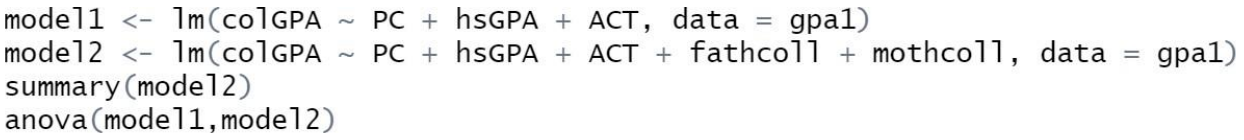
Ans : Individual significance



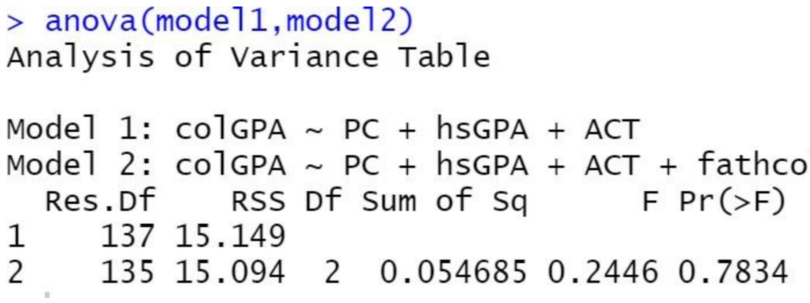
```
Call:
lm(formula = colGPA ~ PC + hsGPA + ACT + fathcoll + mothcoll,
    data = gpa1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.78149 -0.25726 -0.02121  0.24691  0.74432 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.255554   0.335392   3.744 0.000268 ***
PC           0.151854   0.058716   2.586 0.010762 *  
hsGPA        0.450220   0.094280   4.775 4.61e-06 ***
ACT          0.007724   0.010678   0.723 0.470688    
fathcoll     0.041800   0.061270   0.682 0.496265    
mothcoll    -0.003758   0.060270  -0.062 0.950376    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3344 on 135 degrees of freedom
Multiple R-squared:  0.2222,	Adjusted R-squared:  0.1934 
F-statistic: 7.713 on 5 and 135 DF,  p-value: 2.083e-06
```

Both fathcoll and mothcoll have p-values greater than 0.05, therefore they are not statisticallysignificant at 5% level.

Let’s do joint hypothesis testing



```
model1 <- lm(colGPA ~ PC + hsGPA + ACT, data = gpa1)
model2 <- lm(colGPA ~ PC + hsGPA + ACT + fathcoll + mothcoll, data = gpa1)
summary(model2)
anova(model1,model2)
```

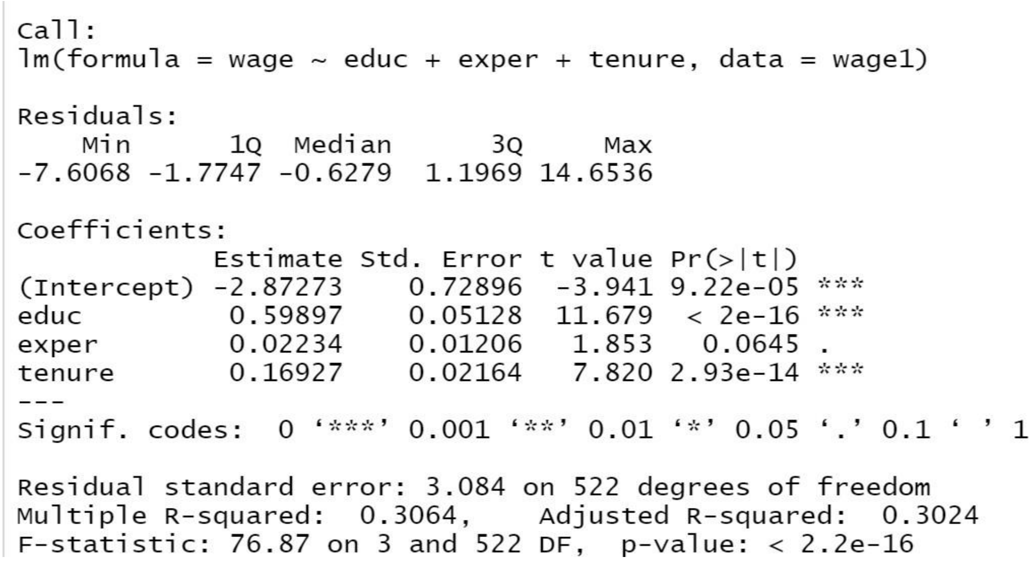


```
> anova(model1,model2)
Analysis of Variance Table

Model 1: colGPA ~ PC + hsGPA + ACT
Model 2: colGPA ~ PC + hsGPA + ACT + fathcoll + mothcoll
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    137 15.149
2    135 15.094  2  0.054685 0.2446 0.7834
```

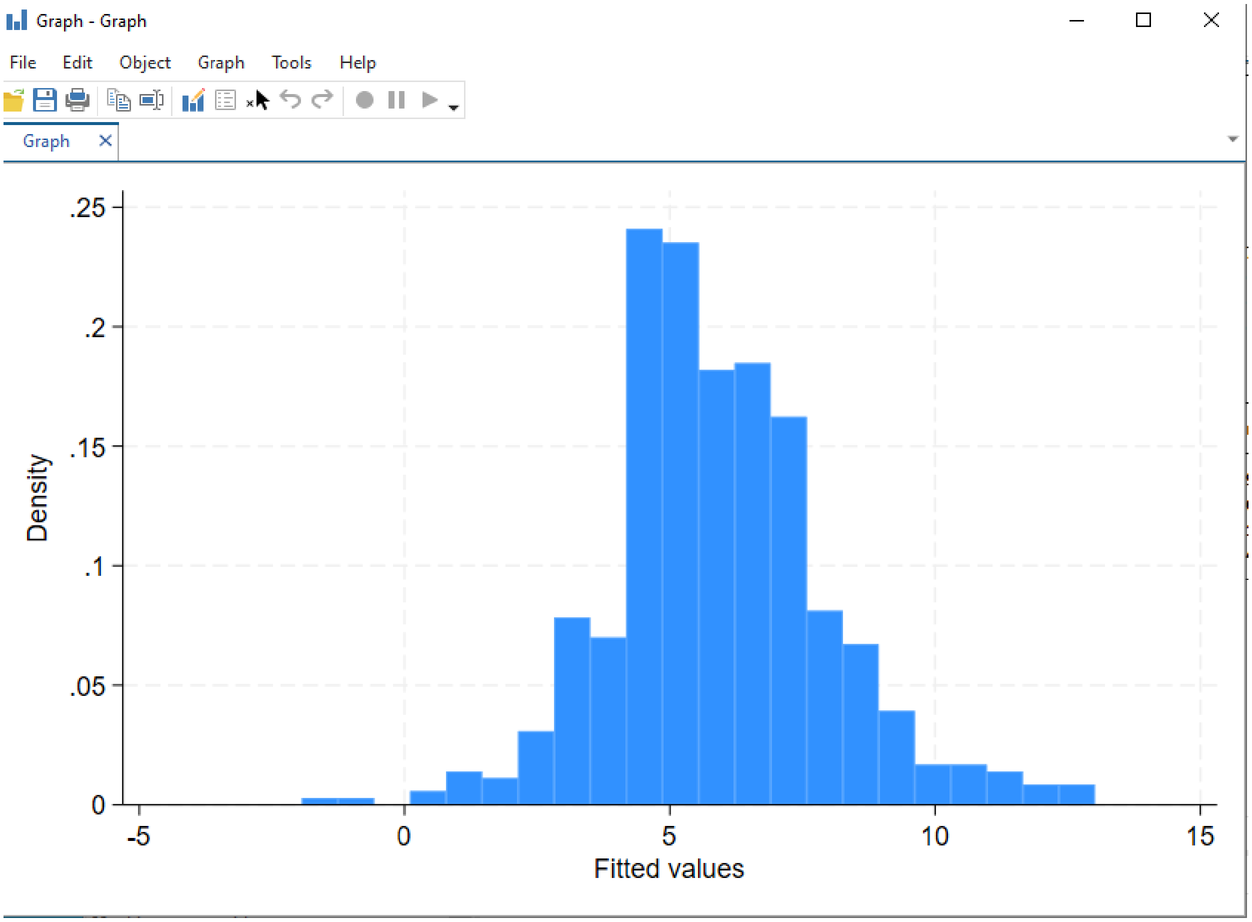
The estimated p-value is 0.7834, both(fathcoll and mothcoll) jointly are statistically insignificant at 5%level.

5.) (Lecture 5) (10 points) Use the data in WAGE1 for this exercise.i.) (4 points) Estimate the equation
$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$
and report the OLS estimates in the usual form. Save the residuals and plot a histogram.
[Hint: 1) You can obtain the residuals of each prediction by using the **residuals** command and storing these values in a variable named whatever you'd like, e.g., predict resid_wage, residuals. 2) You can usethe **histogram** command to plot a histogram, e.g., histogram resid_wage.]

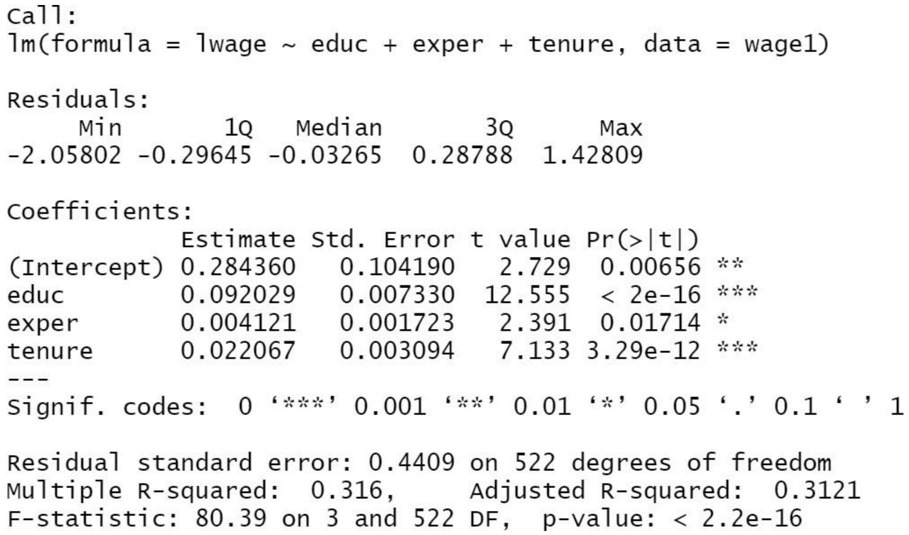
Ans :



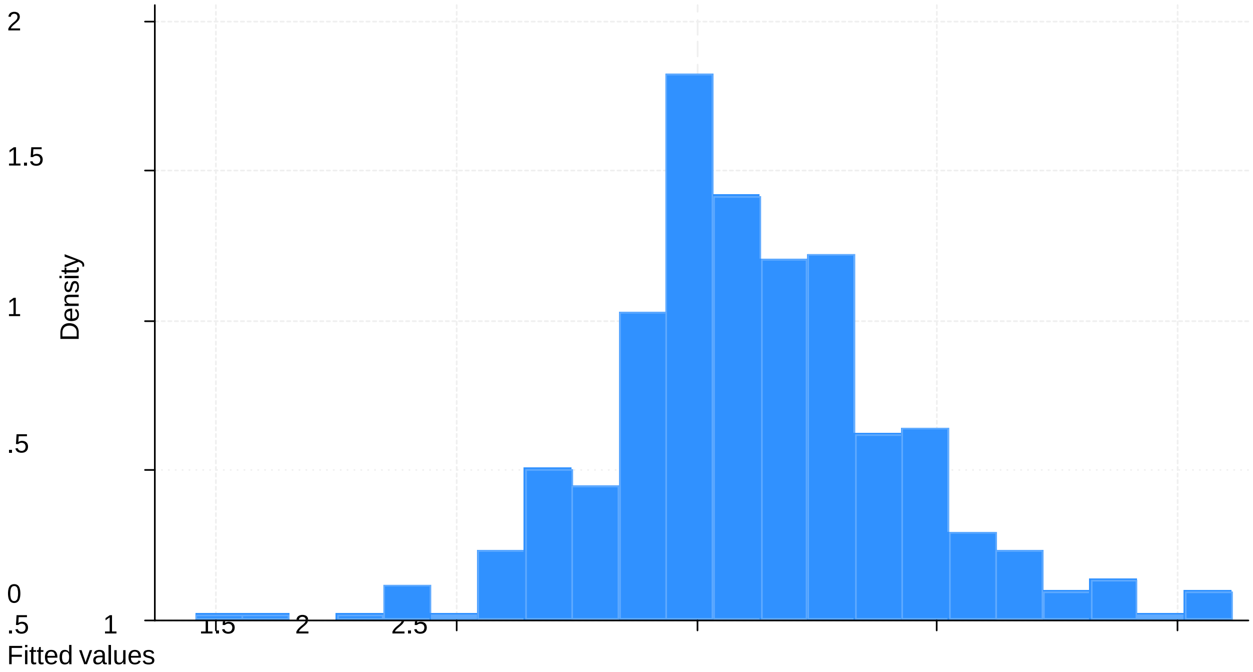
```
Call:
lm(formula = wage ~ educ + exper + tenure, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6068 -1.7747 -0.6279  1.1969 14.6536

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.87273    0.72896  -3.941 9.22e-05 ***
educ         0.59897    0.05128  11.679  < 2e-16 ***
exper        0.02234    0.01206   1.853   0.0645 .
tenure       0.16927    0.02164   7.820 2.93e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.084 on 522 degrees of freedom
Multiple R-squared:  0.3064,    Adjusted R-squared:  0.3024
F-statistic: 76.87 on 3 and 522 DF,  p-value: < 2.2e-16
```



ii.) (4 points) Repeat part (i), but with log(*wage*) as the dependent variable.



```
Call:
lm(formula = lwage ~ educ + exper + tenure, data = wage1)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.05802 -0.29645 -0.03265  0.28788  1.42809 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.284360   0.104190   2.729  0.00656 ** 
educ        0.092029   0.007330  12.555  < 2e-16 ***
exper       0.004121   0.001723   2.391  0.01714 *  
tenure      0.022067   0.003094   7.133 3.29e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4409 on 522 degrees of freedom
Multiple R-squared:  0.316,	Adjusted R-squared:  0.3121 
F-statistic: 80.39 on 3 and 522 DF,  p-value: < 2.2e-16
```



iii.) (2 points) Would you say that Assumption MLR.6 is closer to being satisfied for the level-level modelor the log-level model? Explain.

Ans : The residuals from the log(wage) i.e (ii) regression appear to be more normally distributed. Certainly the histogram in part (ii) fits under its comparable normal density better than in part (i), andthe histogram for the wage residuals is notably skewed to the right.

6.) (Lecture 5) (13 points) The model we used in class to explain the standardized outcome on a final exam (*stndfnl*) in terms of percentage of classes attended, prior college grade point average, and ACTscore is
$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA^2 + \beta_5 ACT^2 + \beta_6 priGPA \cdot atndrte +$
$u.$
i.) (2 points) Argue that
$\Delta stndfnl$
$= \beta_2 + 2\beta_4 priGPA + \beta_6 atndrte.$
$\Delta priGPA$

ANS :
If we hold everything else contant except priGPA then we get the equation to be

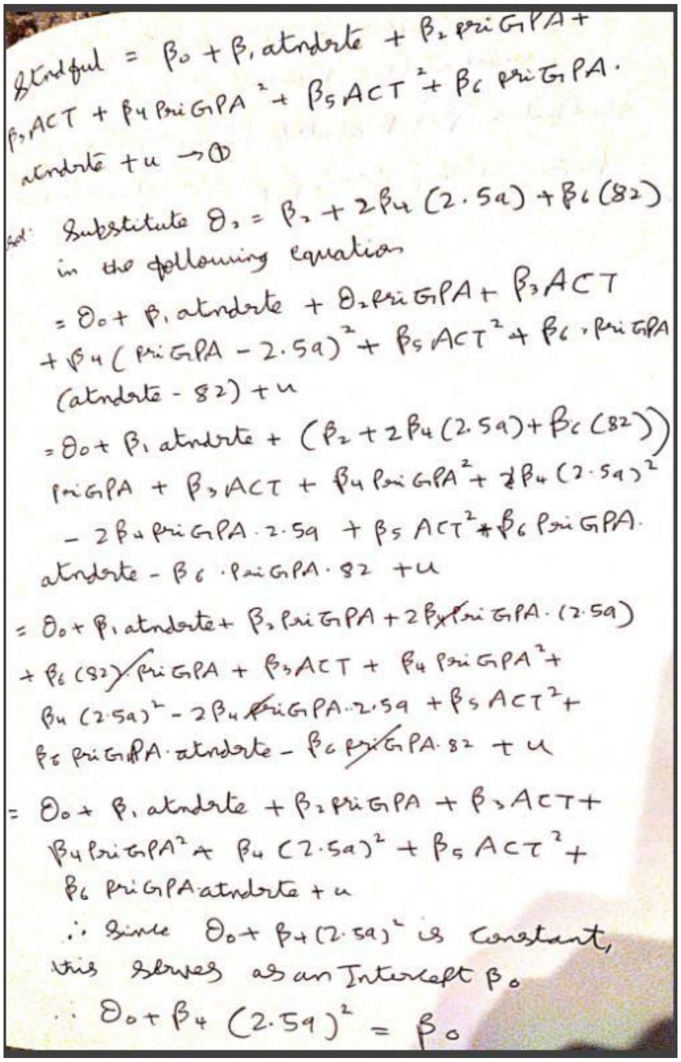
$$\Delta stndfnl = \beta_2 priGPA + \beta_4 priGPA^2 + \beta_6 priGPA \cdot atndrte.$$

Using the calculus assumption $\beta_4 priGPA^2 = 2\beta_4 priGPA$ andUsing partial effect of interaction terms we get:-

$\Delta stndfnl$
$= \beta_2 + 2\beta_4 priGPA + \beta_6 atndrte$
$\Delta priGPA$
ii.) (3 points) Use the equation above to estimate the partial effect of *priGPA* on *stndfnl* when*priGPA* is at its mean value 2.59, and *atndrte* is also at it mean value 82. Interpret your estimate. [Hint: The estimated OLS equation can be found in Lecture 5.]
Ans :
Beta2 = -1.63, Beta4 = 0.296, Beta6 = 0.0056 [Taken from the lecture-5 slides-12,13]
Now substitute this values in the above equation, with mean priGPA value = 2.59, atndrte = 82,
-1.63 + 2*(0.296)*(2.59) + (0.0056)*(82) = 0.36428
A 10 percentage points increase in priGPA increases stndfnl by 3.6 standard deviations from the meanfinal exam score

iii.)(4 points) Show that the equation can be re-written as $stndfnl = \theta_0 + \beta_1 atndrte + \theta_2 priGPA + \beta_3 ACT + \beta_4(priGPA - 2.59)^2 + \beta_5 ACT^2 + \beta_6 priGPA \cdot (atndrte - 82) + u$,where $\theta_2 = \beta_2 + 2\beta_4(2.59) + \beta_6(82)$. How do you interpret$\theta_2$?

Ans:



$stndful = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 PriGPA^2 + \beta_5 ACT^2 + \beta_6 priGPA \cdot atndrte + u \rightarrow (1)$

Sol: Substitute $\theta_2 = \beta_2 + 2\beta_4(2.59) + \beta_6(82)$ in the following equation

$= \theta_0 + \beta_1 atndrte + \theta_2 priGPA + \beta_3 ACT + \beta_4(priGPA - 2.59)^2 + \beta_5 ACT^2 + \beta_6 \cdot priGPA (atndrte - 82) + u$

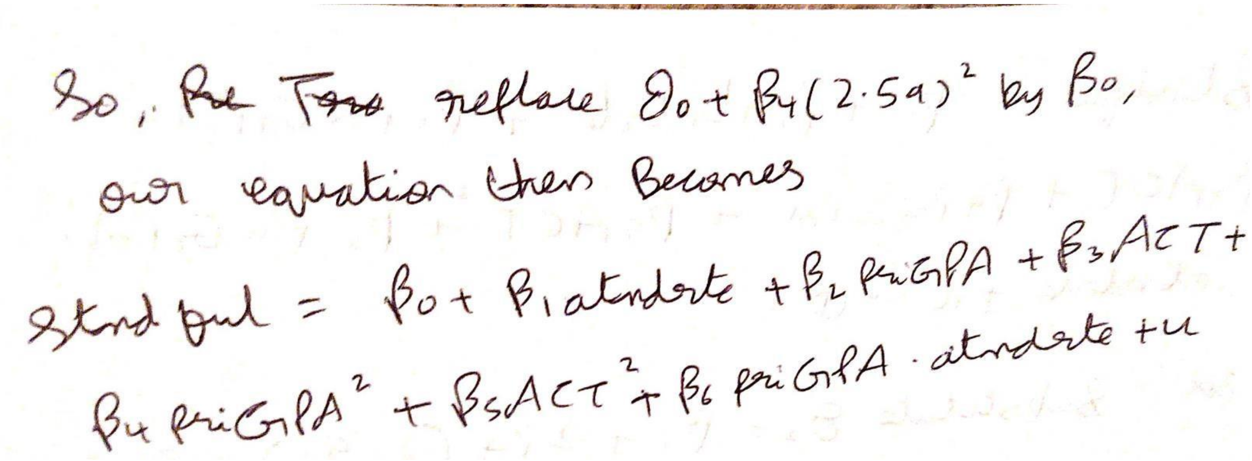
$= \theta_0 + \beta_1 atndrte + (\beta_2 + 2\beta_4(2.59) + \beta_6(82)) priGPA + \beta_3 ACT + \beta_4 PriGPA^2 + 2\beta_4(2.59)^2 - 2\beta_4 priGPA \cdot 2.59 + \beta_5 ACT^2 + \beta_6 PriGPA \cdot atndrte - \beta_6 \cdot PriGPA \cdot 82 + u$

$= \theta_0 + \beta_1 atndrte + \beta_2 PriGPA + 2\beta_4 \cancel{Pri}GPA \cdot (2.59) + \beta_6(82)\cancel{\cdot priGPA} + \beta_3 ACT + \beta_4 PriGPA^2 + \beta_4(2.59)^2 - 2\beta_4 \cancel{pri}GPA \cdot 2.59 + \beta_5 ACT^2 + \beta_6 priGPA \cdot atndrte - \beta_6 \cancel{pri}GPA \cdot 82 + u$

$= \theta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 PriGPA^2 + \beta_4(2.59)^2 + \beta_5 ACT^2 + \beta_6 priGPA \cdot atndrte + u$

$\therefore$ Since $\theta_0 + \beta_4(2.59)^2$ is constant, this serves as an Intercept β_0

$\therefore \theta_0 + \beta_4(2.59)^2 = \beta_0$



So, replace $\theta_0 + \beta_4(2.59)^2$ by β_0, our equation then Becomes

$$stnd\,fnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA^2 + \beta_5 ACT^2 + \beta_6 priGPA \cdot atndrte + u$$

Interpretation of θ_2 : θ_2 can be interpreted as the partial effect of priceGPA on stndfnll if variables atndrte and priGPA^2 take on their mean values.

iv.)(4 points) Following (iii), suppose that, in place of $pri GPA \cdot (atndrte - 82)$, you put $(priGPA - 2.59) \cdot (atndrte - 82)$. Now how do you interpret the coefficients on *atndrte* and *priGPA*?

Ans : Following (iii), if we replace $priGPA \cdot (atndrte - 82)$ by $(priGPA - 2.59) \cdot (atndrte - 82)$,the equation gets transformed as

$$stndfnl = \theta_0 + \theta_1 atndrte + \theta_2 priGPA + \beta_3 ACT + \beta_4 (priGPA - 2.59)^2 + \beta_5 ACT^2 + \beta_6 \cdot (priGPA - 2.59) \cdot (atndrte - 82) + u$$

The coefficient of atndrte = θ_1
The coefficient of priGPA = θ_2

Interpretation of atndrte coefficient = The partial effect of atndrte on students final exam performance(stndfnl) at the mean value of priGPA (2.59).

Interpretation of priGPA coefficient = The partial effect of priGPA on students final exam performance(stndfnl) at the mean value of atndrte (82).

Reference : Lecture-5 slide-11.

R-CODE USED

```
#Qno-1 library(wooldridge)data("charity") mean(charity$gift)
100*(sum(charity$gift == 0) / nrow(charity))mean(charity$mailsyear) max(charity$mailsyear) min(charity$mailsyear)
model <- lm(gift ~ mailsyear, data = charity)summary(model)
2.0141 + (2.6495) * min(charity$mailsyear)#Qno-2
data("ceosal2")
model <- lm(lsalary ~ lsales + lmktval, data = ceosal2)summary(model)
model <- lm(lsalary ~ lsales + lmktval + profits, data = ceosal2)summary(model)
model <- lm(lsalary ~ lsales + lmktval + profits + ceoten, data = ceosal2)summary(model)
percentage_return <- 100 * (exp(1.168e-02) - 1)percentage_return
```

```
cor(ceosal2$lmktval,ceosal2$profits, method = "pearson")#Qno-3
data("twoyear")
model <- lm(lwage ~ jc + totcoll + exper, data = twoyear)summary(model)
min(twoyear$phsrank) max(twoyear$phsrank) mean(twoyear$phsrank)
model <- lm(lwage ~ jc + totcoll + exper + phsrank, data = twoyear)summary(model)
model <- lm(lwage ~ jc + totcoll + exper + phsrank + id, data = twoyear)summary(model)
#Qno-4 data("gpa1")
model <- lm(colGPA ~ PC + hsGPA + ACT, data = gpa1)summary(model)
0.157309 + (1.977431*0.057287)
0.157309 - (1.977431*0.057287)
model1 <- lm(colGPA ~ PC + hsGPA + ACT, data = gpa1)
model2 <- lm(colGPA ~ PC + hsGPA + ACT + fathcoll + mothcoll, data = gpa1)summary(model2)
anova(model1,model2)#Qno-5
data("wage1")
model <- lm(wage ~ educ + exper + tenure,data=wage1)summary(model)
residuals <- resid(model)
hist(residuals, main = "Histogram of Residuals", xlab = "Residuals")model <- lm(lwage ~ educ + exper +
tenure,data=wage1)
```

```
summary(model) residuals <- resid(model)
hist(residuals, main = "Histogram of Residuals", xlab = "Residuals")#Qno-6
-1.63 + (2*(0.296)*(2.59)) + (0.0056*82)
data("attend")
model <- lm(stndfnl ~ atndrte+priGPA+ACT+priGPA^2+ACT^2+priGPA*atndrte,data=attend)summary(model)
mean(attend$atndrte)
attend["priGPA"] = attend["priGPA"] - 2.59attend["priGPA"]
model1 <- lm(stndfnl ~ atndrte+priGPA+ACT+(priGPA^2)+(ACT^2)+(priGPA*atndrte),data=attend)summary(model1)
attend["atndrte"] = attend["atndrte"] - 82
model2 <- lm(stndfnl ~ atndrte+priGPA+ACT+(priGPA^2)+(ACT^2)+(priGPA*atndrte),data=attend)summary(model2)
model2 <- lm(stndfnl ~ atndrte+priGPA+ACT+(priGPA^2)+(ACT^2)+(priGPA*atndrte),data=attend)summary(model2)
```