Home Assignment 1: Selvamalai Thiraviyam

Part A (Reading Assignments)

1. **Please get hold of the main book for part3:**

<u>Which part of the slides corresponds to chapter 3?</u>

Chapter 3 of the Econometrics (Hansen, B) was mainly used to prepare the slides of "Part II.1: Linear Model and OLS Regression (UEA_ecoR2PhD_CoreLectA01_OLS_stkm.pdf).

Slide number 14 explains the linear projection coefficient as the best linear predictor based on page 63 of the book (3.2, sample).

Slides 22 and 23 demonstrate least-squares residuals, which is based on 3.8 (pages 70-71). In these slides, the fitted values and the residuals of the OLS estimates were defined, and the property of the residuals was explained and proved.

Slide.24

In the book, 3.11 and 3.12 (projection matrix and annihilator matrix) were considered for preparing the 24th side (Matrix Algebra and projection matrix)

Slide-25 (Geometric interpretation of OLS)

It is based on part 3.15 of the chapter3. The slide is more informative than the geometric interpretation given in the book.

<u>Provide a sentence listing all sections you read, and explain whether you understood all, some or none of them</u>

I have fully gone through until 3.15 and all the Appendix matrixes that are listed for the reading. Part 3.2 explains the dataset or sample for estimating the parameters of the linear model, while 3.3 illustrate the moment estimator for the mean and variance of the dataset. The definition of the least estimator is given in part 3.4, and the equations to estimate the coefficient of the simple linear regression model is illustrated in 3.5 of chapter three. The 3.6 and 3.7 demonstrate the least square estimator for multiple regressors with a table example. The rest of the chapter until

3.15 illustrates the least square residuals, demanded regressors, model in matrix notation, projection matrix, annihilator matrix, estimation of error variance, analysis of variance and projections.

Appendix A provides details of matrix algebra with sample examples, for example, matrix notation (A1), matrix addition (A3), matrix multiplication (A4), rank and inverse of a matrix (A.6), orthogonal and orthonormal matrices (A7), determinant (A8), positive definite matrices (A10), idempotent matrices (singular values) and matrix calculus.

I understood all the parts of the chapter because this chapter mainly considered introductory parts, which I have already studied in my master degree. Therefore, I was able to quickly refresh when I read the contents of the book chapters.

Provide a short sentence about each of these chapters to document your reading

Appendix B1 states the different inequalities for real numbers: triangular inequality, Jensen's inequality, geometric mean inequality, Loeve's C_r inequality and norm monotonicity. Appendix B2 explains the inequalities for vectors. Triangle inequality, c2 inequality, Holder's inequality, Schwarz inequality, Minkowski's inequality and the triangle inequality.

The triangle inequality is essential in proving the vector propositions. For example, based on the concept of the triangle inequality, we can prove that open ball is an open set case and closed ball is a closed set.

Appendix B.3 shows the inequalities for matrices. In this case, Schwarz Matrix inequality, trace inequality, quadratic inequality, strong Schwarz matrix inequality, norm equivalence and Eigenvalue product inequality are used to prove matrix propositions and theorem.

2. **Revise the slides**

The slides, along with a recorded video, are posted on the backboard of the university. It is more valuable to understand the contents of the falls.

3. Outlook

Check out the slide deck on Experiments, and discuss the difference between the ATE and the ATET.

The lecture slides provide more informative information about ATE and ATET. Here, I will list very few things to differentiate the ATE and the ATET on the experimental analysis.

Average Treatment Effect (ATE) and Average Treatment Effect on the Treated (ATET) are essential concepts in the experimental analysis. A causal effect of the treatment explains the effect of a treatment on a population/sample. ATE is a difference between expected outcomes before and after treatment to a selected population/sample. The following equation can note it:

$$ATE = E[Y_i^{D=1}] - E[Y_i^{D=0}]$$

The ATE can be decomposed as follows: Average Treatment Effect on the Treated (ATET) and Average Treatment Effect on non-treated (ATENT). If treated and non-treated are random sub-samples of the population, then it implies ATE=ATET=ATENT. In the case of the non-randomized case or small sample case, ATE may not be equal to ATET or ATENT. Further, ATE does not explain the counterfactual effect but ATET or ATENT provides the counterfactual effects of the treatments on the analysis.

Check out the slide deck on IV: Give the definition of the simple IV in the univariate case

Consider the following univariate regression model:$y_i = \beta_0 + \beta_1 x_i + u_i$

If $E(u_i|x_i) = 0$, then estimates for β s are unbiased.

If $E(u_i|x_i) \neq 0$,the OLS estimates are biased.

In this case, there is a variable, z, which is correlated directly with x and indirectly with y, but it is not associated with an unobserved variable, u. The variable, z, can be a simple instrumental variable in the univariate case.

Check out the slide deck on IV **: Two Key assumptions that have to hold for an IV**

- $Cov(z_{i1}, u_{ik}) = 0$: Exogeneity
- $Cov(z_{i1}, x_{ik}) \neq 0$: Relevance

Bonus: Gove over confidence intervals and hypothesis testing

The slides of the second Stats-Prime focuses on the finite sample, small sample, or exact properties of the OLS estimates illustrating interval estimates and hypotheses testing. Further, normal t-statistics and f-statistics are applicable in this case. The Extra slide deck on asymptotics describes the interval estimates and hypothesis testing under large sample or asymptotic properties.

There are many reasons to choose the second case in the real world. In many social science researches, holding normality assumptions may not be possible. In this situation, asymptotic properties are more appropriate. In social science or economics, the large sample selection is crucial to generalize the findings. Therefore, I prefer to the second case that tests a hypothesis and interval estimates under the asymptotic assumptions.

Part B: Formal Exercises

1. **Revisit the OLS-slide deck (UEA_ecoR2PhD_CoreLectA01_OLS_stkm**

   Try to understand how we derive the OLS estimator

   Driving OLS estimates: $Q(\beta_1, \beta_2, \dots, \beta_k) = \sum_{t=1}^{T}\left(y_t - \sum_{j=1}^{k}\beta_j x_{tj}\right)^2$

$$Q(\beta) = (y - X\beta)'(y - X\beta)$$

The first order conditions for β:

$$\frac{\boldsymbol{\partial Q(\beta)}}{\boldsymbol{\partial \beta}} = -\mathbf{2}X'\left(y - X\hat{\beta}\right) = 0$$

$$X'\left(y - X\hat{\beta}\right) = 0$$

$$(X'X)\hat{\beta} - X'y = 0$$

$$(X'X)\hat{\beta} = X'y$$

$$\hat{\beta} = (X'X)^{-1}\,X'y$$

Make sure that you understand the proof of unbiasedness

An estimator $\hat{\beta}$ of $\boldsymbol{\beta}$ **is unbiased if** $\boldsymbol{E}\left(\hat{\beta}\right) = \boldsymbol{\beta}$. This can be proved as follows: $\boldsymbol{E}\left(\hat{\beta}|X\right) = \boldsymbol{\beta}$

LHS:

$\boldsymbol{E}\left(\hat{\beta}|X\right) = \boldsymbol{E}[(X'X)^{-1}\, X'y|X]$ (by $\hat{\beta} = (X'X)^{-1}\, X'y$)

$= (X'X)^{-1}\, X'E(y|X)$

$= (X'X)^{-1}\, X'E(X\beta + u)|X)$ (by $y = X\beta + u$)

$= (X'X)^{-1}(\, X'X\beta + X'E(u|X)$ (by conditional expectation

$= (X'X)^{-1}\, X'X\beta + (X'X)^{-1}X'E(u|X)$ (by linearity)

$= (X'X)^{-1}\, X'X\beta$ by assumption $E(u|X) = 0$

$= \beta$ by $((X'X)^{-1}\, X'X = I)$

2. ***Assume you are a senior economist***

If $E(\boldsymbol{u}|X) = 0$ is true, the our OLS estimates would be unbiased. I it is not the case (for example , $E(\boldsymbol{u}|X) = 2$) the estimated coefficients are biased.

We know that $\boldsymbol{E}\left(\hat{\beta}|X\right) = (X'X)^{-1}\, X'X\boldsymbol{\beta} + (X'X)^{-1}X'E(\boldsymbol{u}|X)$ by earlier proof. It can be simplified as follows: $\boldsymbol{E}\left(\hat{\beta}|X\right) = (X'X)^{-1}\, X'X\boldsymbol{\beta} + (X'X)^{-1}X'E(\boldsymbol{u}|X)$

$= \boldsymbol{\beta} + (X'X)^{-1}X'E(\boldsymbol{u}|X)$ by $(X'X)^{-1}\, X'X\boldsymbol{\beta} = \boldsymbol{\beta}$

Therefore, we can estimate the bias as follows: $\boldsymbol{Bias} = \boldsymbol{E}\left(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}\right) - \boldsymbol{\beta}$ (by definition)

$\boldsymbol{Bias} = \boldsymbol{\beta} + (X'X)^{-1}X'E(\boldsymbol{u}|X) - \boldsymbol{\beta}$ **by** $\boldsymbol{E}\left(\hat{\beta}|X\right) = \boldsymbol{\beta} + (X'X)^{-1}X'E(\boldsymbol{u}|X)$

$= (X'X)^{-1}X'E(\boldsymbol{u}|X)$

Now, we will substitute $E(\boldsymbol{u}|X) = 2$, then $Bias = (X'X)^{-1}X' * 2 = 2(X'X)^{-1}X'$ (by matrix property) Here, bias is equal to two times of the product of inverse $X'X$ matrix. The bias is not equal to zero, but it may be either positive or negative. It is depends on the sign of the elements of $(X'X)^{-1}X'$.

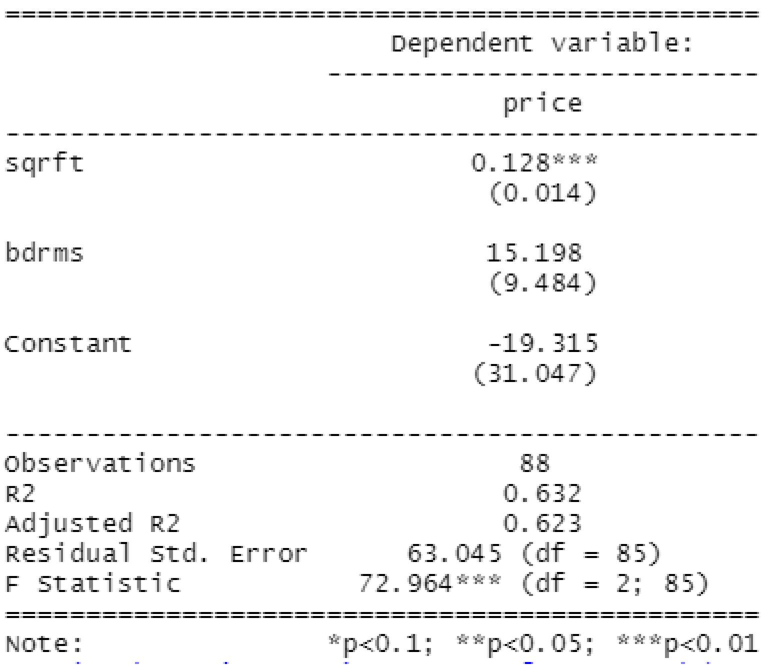
Part C: Coding

1. Do at least two of the exercise from Lab 06

Exercise 1

i. ***Write out the results in equation form***

$$\widehat{price} = -19.315 + 0.128sqrft + 15.198bdrms$$



```
===============================================
                        Dependent variable:
                    ---------------------------
                               price
-----------------------------------------------
sqrft                        0.128***
                              (0.014)

bdrms                         15.198
                              (9.484)

Constant                     -19.315
                             (31.047)

-----------------------------------------------
Observations                    88
R2                             0.632
Adjusted R2                    0.623
Residual Std. Error      63.045 (df = 85)
F Statistic           72.964*** (df = 2; 85)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

ii. ***What is the estimated increase in price for a house with one more bedroom, holding aquare footage constant?***

The coefficient of the independent variable, $bdrms$, provide the answer for this question. The price of the house will increase by 15.198 thousand dollars for one additional one bedroom

iii. ***<u>What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (ii).</u>***

In this case, when they increase the number of rooms, it also increases the square feet of the house. Therefore, the increase in the price of the house is 33.118 thousand (0.128*140 +15.198*1=33.118). The price, in this case, is higher by 17.92 thousand (33.118-15.198=17.92).

iv. ***<u>What percentage of the variation in price is explained by square footage and number of bedrooms?</u>***

The goodness of fit (R-squared) of the model explains the answer to this question. To find the percentage variation, R-squared should be multiplied by 100. In this example, 63.2% of the variation in price is explained by square footage and a number of bedrooms. It reveals that the estimated model did not capture around 36.8% of the variation of the price. Therefore, it is better to add some more relevant explanatory variables, for example, the distance from city centre, facilities (school, hospital, shopping mall), etc.

v. ***<u>The first house in the sample has sqrft=2,438 and bdrms=4. Find the predicted selling price for this house from the OLS regression line.</u>***

The predicted price of the house is 353.541 thousand.

$$\widehat{price} = -19.315 + 0.128 sqrft + 15.198 bdrms$$

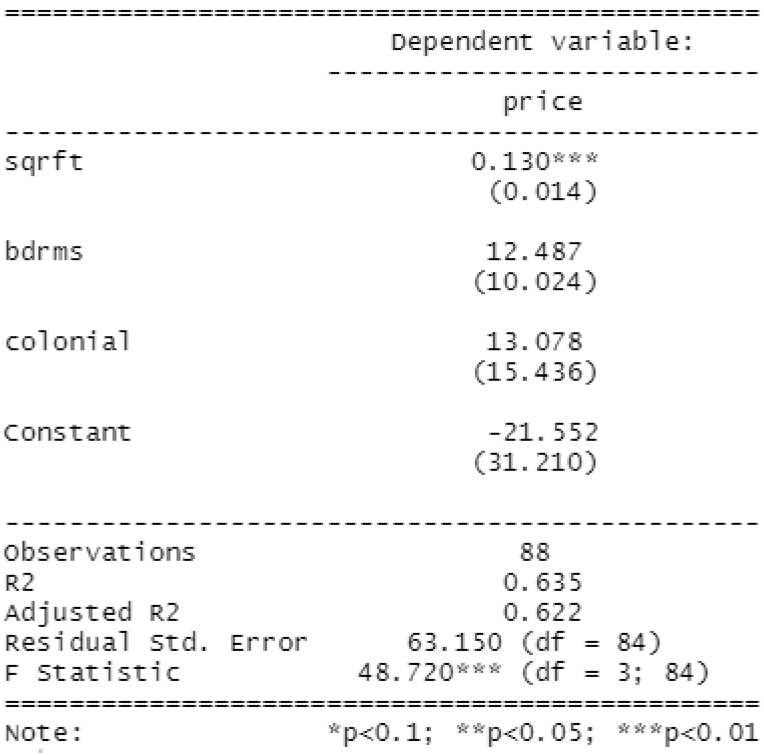
$$= -19.315 + 0.128 * 2438 + 15.198 * 4$$

$$= 353.541 \text{ thousands}$$

vi. ***<u>The actual selling price of the first house in the sample was $300,000 (so price=300). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?</u>***

The residuals can be estimated by subtracting the estimated house price from the actual price ($\hat{\mathrm{u}} = \mathrm{price} - \widehat{\mathrm{price}}$). In this example, the estimated residual is -53.541. From this, we can say that buyer underpaid by 53.541 thousand because he paid less than the estimated price of the house.

vii. ***<u>Now add the variable colonial to your model. Interpret its coefficient. Is it significant?</u>***

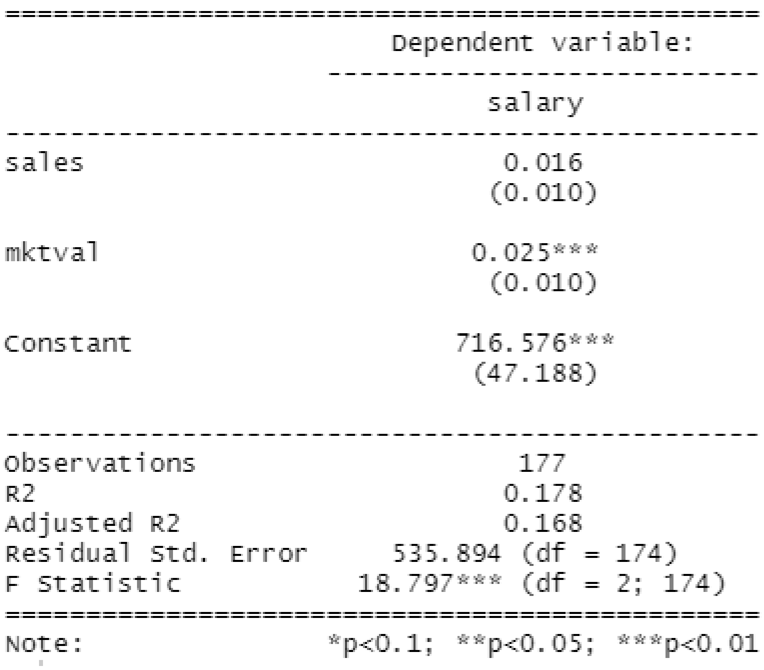


```
===============================================
                        Dependent variable:
                    ---------------------------
                               price
-----------------------------------------------
sqrft                        0.130***
                             (0.014)

bdrms                         12.487
                             (10.024)

colonial                      13.078
                             (15.436)

Constant                     -21.552
                             (31.210)

-----------------------------------------------
Observations                    88
R2                            0.635
Adjusted R2                   0.622
Residual Std. Error      63.150 (df = 84)
F Statistic          48.720*** (df = 3; 84)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

In this model, colonial is a binary variable which means that if the house is in a colonial area, it will take value one, otherwise zero. The variable coefficient is 13.078, which means that the price of a house will be higher by approximately $13 thousand for colonial places than other areas, holding other factors remaining the same. Further, this coefficient is highly significant at a one per cent level to predict the house prices because the t-statistics is more than 2.58, which is the 1% critical t-value.

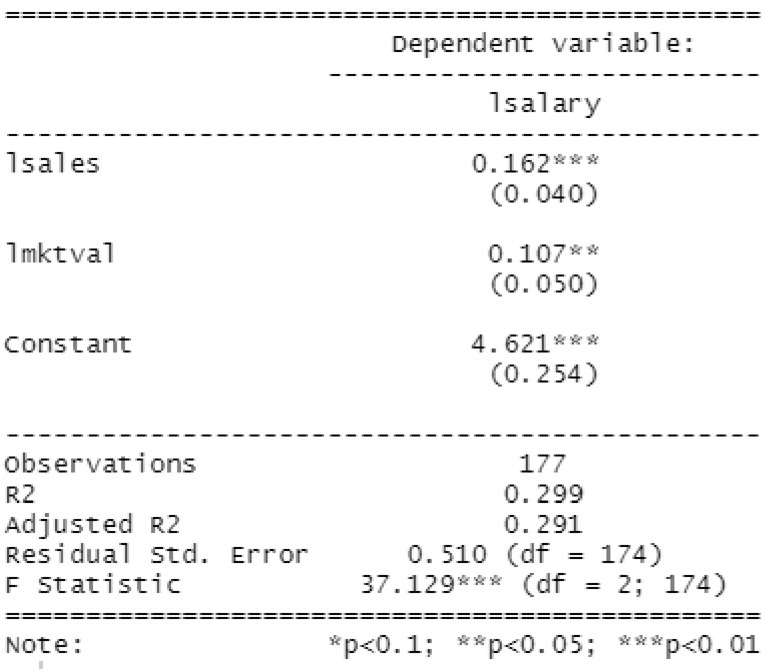
Exercise 2

i. The estimated output is given below:



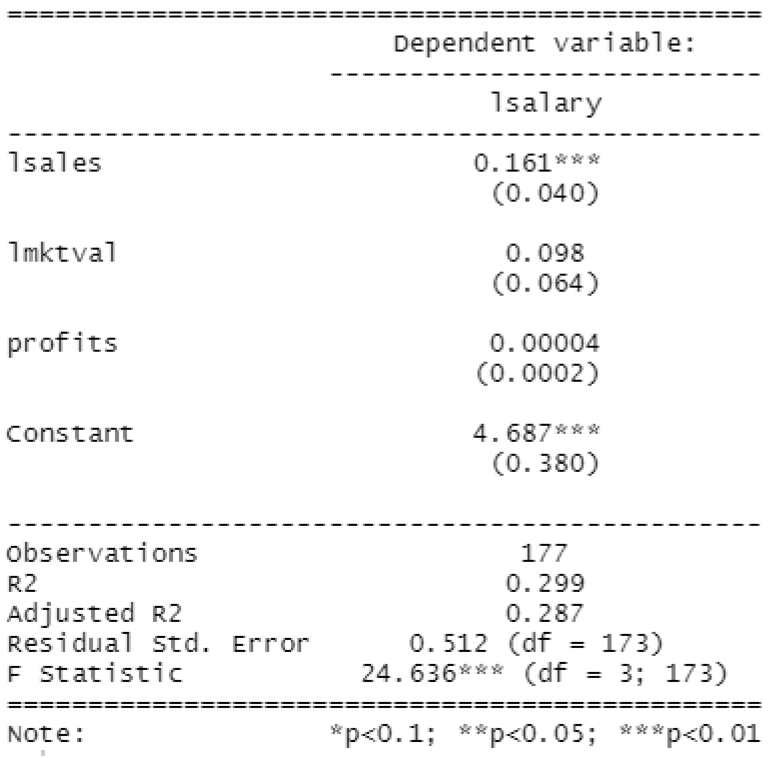
```
===============================================
                        Dependent variable:
                    ---------------------------
                              salary
-----------------------------------------------
sales                          0.016
                              (0.010)

mktval                        0.025***
                              (0.010)

Constant                     716.576***
                             (47.188)

-----------------------------------------------
Observations                    177
R2                             0.178
Adjusted R2                    0.168
Residual Std. Error      535.894 (df = 174)
F Statistic           18.797*** (df = 2; 174)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

The constant elasticity model was estimated by taking logarithms for both side of the model.



```
==============================================
                      Dependent variable:
                  ----------------------------
                           lsalary
----------------------------------------------
lsales                    0.162***
                          (0.040)

lmktval                    0.107**
                          (0.050)

Constant                  4.621***
                          (0.254)

----------------------------------------------
Observations                177
R2                         0.299
Adjusted R2                0.291
Residual Std. Error    0.510 (df = 174)
F Statistic         37.129*** (df = 2; 174)
==============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

ii. Adding variable profits with the earlier model. The logarithms cannot be defined for the negative values. In the case of profits, there may be a negative value if a business is lost. Based on the R-squared, we can answer this question. The R-squared is 0.299, which is below 0.3. It means that only 30% of the variation in the log salary was explained by the variation of the above three variables. Therefore, we can say that these performance variables did not explain most of the variation in CEO salaries (see the below table)

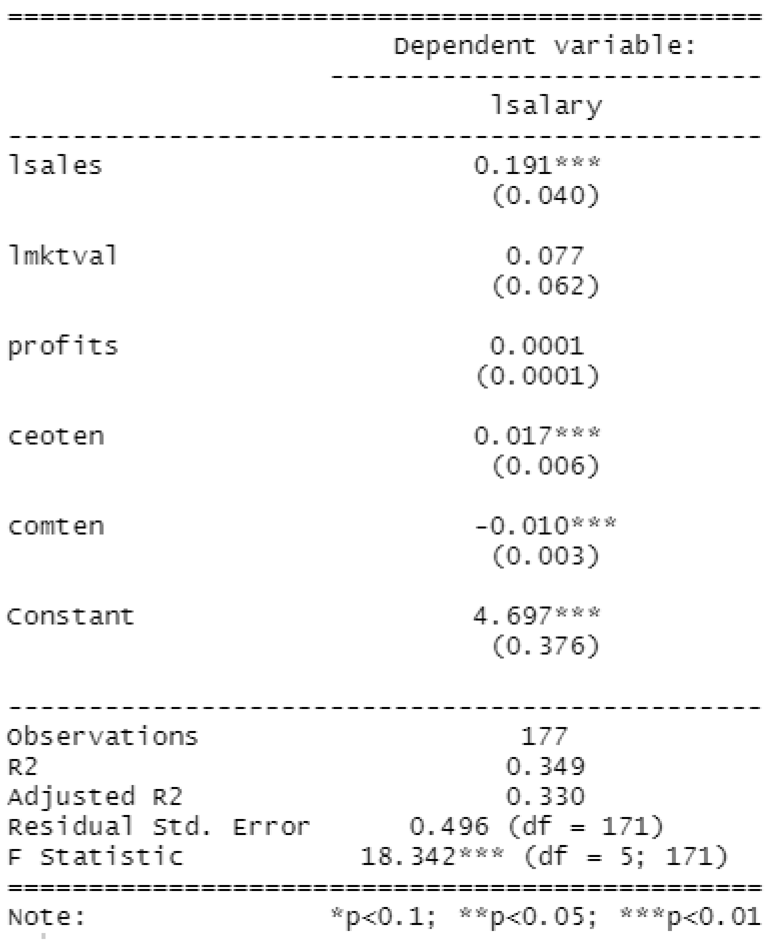


```
=================================================
                         Dependent variable:
                     ----------------------------
                               lsalary
-------------------------------------------------
lsales                         0.161***
                               (0.040)

lmktval                         0.098
                               (0.064)

profits                        0.00004
                              (0.0002)

Constant                       4.687***
                               (0.380)

-------------------------------------------------
Observations                     177
R2                              0.299
Adjusted R2                     0.287
Residual Std. Error        0.512 (df = 173)
F Statistic           24.636*** (df = 3; 173)
=================================================
Note:                 *p<0.1; **p<0.05; ***p<0.01
```

iii. The coefficient of the lmktval is 0.098, which means that 1% increase in the market values is expected to increase the salary by less than 0.1%. Further, it is not statistically significant at 5% level.

iv. The coefficient on profits is 0.00004. It has a slightly different interpretation of salary. One additional dollar increase on firm profits increases the salary by 0.004% (0.00004*100). This variable is also not statistically significant to explain the firm's salary.

v. The coefficient on ceoten is 0.017. Therefore, 1.7% is the estimated percentage return for another year of CEO tenure, holding other factors fixed. (See the output given below):



```
================================================
                       Dependent variable:
                   ----------------------------
                            lsalary
------------------------------------------------
lsales                     0.191***
                            (0.040)

lmktval                     0.077
                            (0.062)

profits                     0.0001
                           (0.0001)

ceoten                     0.017***
                            (0.006)

comten                     -0.010***
                            (0.003)

Constant                   4.697***
                            (0.376)

------------------------------------------------
Observations                  177
R2                           0.349
Adjusted R2                  0.330
Residual Std. Error     0.496 (df = 171)
F Statistic          18.342*** (df = 5; 171)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

vi. The coefficient on tenure:0.017. As I mentioned earlier, one additional tenure is expected to increase the salary by 1.7 percentage. Further, it is a highly significant variable in explaining the salary because the p-value is less than 0.01.

The coefficient on comten is -0. 01: This means that one additional unit increase on comten decrease the salary by one percent. This variable is significant at 5% since the corresponding p-value is less than 0.05.

vii. Salary does depend on several variables. In this example, only one variable (comten) affects negatively on salary. In this case, a person can gen the lowest salary even the person holds longer tenure.

viii. See the Rcripts

Part-C

See R-scripts (part-c) for this part