EEP/IAS 118 - Introductory Applied Econometrics

Problem Set 4, Spring 2025, Villas-Boas

Due in Gradescope – Midnight, April 6

Submit materials as **one pdf** on Gradescope. After uploading the pdf to Gradescope, please **assign all and only the appropriate pages to each question**. Questions that do not have properly assigned pages on Gradescope may not be graded. Codes and outputs not properly displayed will be marked as incorrect.

For full credit, all confidence intervals/hypothesis tests must be conducted by hand - you can use functions like sd() or mean() to get values to plug into the formulas, but no credit will be given for the use of canned interval/test functions (i.e. linearHypothesis()) with no steps/calculations provided. Do not round any intermediate steps or final answers to less than four decimal digits.

Preamble

When writing R code, it's a good habit to start your notebooks or R scripts with a preamble, a section where you load all necessary packages, set paths or change the working directory, or declare other options.

Use the below code cell to load in packages you will use throughout the problem set (at least `haven`, `tidyverse` and `ggplot2`, `dplyr', ` psych', `car', ` lm.beta').

*Note: All packages that you need are already installed and can be loaded immediately using the `library()` function.

set scientific display off by typing in the cell bellow

options(scipen=999)

In [14]:
```
install.packages("pacman")
#Now load it...
# Load the 'pacman' package

library(pacman)
```

```
#packages to use load them now using the pacman "manager"
p_load(pacman, haven, tidyverse, dplyr, psych,ggplot2,car,lm.beta)

#set scientific display off, thank you Roy
options(scipen=999)
```

```
Installing package into '/srv/r'
(as 'lib' is unspecified)
```

Exercise 1.

In this problem set, we use a dataset on the annual salary of executives and the characteristics of the firm, and the firm's outcomes. If the labor market does not value a characteristic of the employer, such as an outcome in the firm that the executive is responsible for (i.e. the value of sales or change in the rate of return) or the years of tenure as an executive (proxying experience), the demand for those executives and their salary goes down and vice versa.

VARIABLE	Definition
SALARY	annual CEO salary (including bonuses) in 1990 (in thousands USD)
SALES	firm sales in 1990 (in millions USD)
ROE	average return on equity, 1988–1990 (in percent)
FINANCE	= 1 if a financial company, 0 otherwise

0. Setup *(Ungraded)*: Begin by reading in the dataset "pset4_2025.dta." Note that this dataset is in dta format so you will need use the read_dta() function from the *haven* package. Create a variable called *lsalary* that is the ln of salary and add this variable as an additional column in your dataframe. Call this variable lsalary. Explore your dataset by viewing summary statistics of key variables (salary, roe, finance) by using the summarise() function.

In [15]:
```
# Read in data
my_data <- read_dta("pset4_2025.dta")

# Create log(salary) variable
my_data$lsalary <- log(my_data$salary)

# Explore summary statistics of key variables
summarise(my_data,
          "Average Salary" = mean(salary),
          "Std Dev Salary" = sd(salary))

summarise(my_data,
          "Average ROE" = mean(roe),
          "Std ROE" = sd(roe))
```

```
summarise(my_data,
          "Average Financial Company" = mean(finance),
          "Std Dev Financial Company" = sd(finance))

# Create two separate dataframes of salaries, one for finance and one for th
financeWages<-my_data$salary[(my_data$finance==1)]
nonFinanceWages<-my_data$salary[(my_data$finance==0)]
```

A tibble: 1 × 2

Average Salary	Std Dev Salary
<dbl>	<dbl>
1272.771	1372.688

A tibble: 1 × 2

Average ROE	Std ROE
<dbl>	<dbl>
17.22244	8.590926

A tibble: 1 × 2

Average Financial Company	Std Dev Financial Company
<dbl>	<dbl>
0.2243902	0.4182014

Q1.1 Standardized Regression

a) Estimate a model of salary as a linear function of a constant, firm's sales, and average ROE, using a **standardized regression**. In other words, all variables should be expressed in terms of standard deviations. *(Hint: You can either create standardized versions of each variable manually or use the lm.beta() function from the lm.beta package. See Lecture 13 and Section 6.)*

b) Interpret the intercept and each of the estimated slope coefficients in the standardized regression using Sign, Size, and Significance (SSS). Pay special attention to units since this is a standardized regression! Interpret each coefficient in a maximum of 2 sentences.

c) In absolute terms, does average ROE or sales have a larger correlation with expected salary? Explain your answer in a maximum of 2 sentences.

In [16]:

```
# First run the usual regression
reg1<-lm(salary~sales+roe, my_data)

# Then use the lm.beta() function to standardize the coefficients
reg1_standardized <- lm.beta(reg1)
summary(reg1_standardized)
```

```
Call:
lm(formula = salary ~ sales + roe, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1346.4  -484.7  -229.1   134.1 13574.3

Coefficients:
              Estimate Standardized Std. Error t value Pr(>|t|)
(Intercept) 822.893885           NA 224.845205   3.660 0.000322 ***
sales         0.014732     0.114594   0.008932   1.649 0.100615
roe          20.257893     0.126783  11.100955   1.825 0.069496 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1360 on 202 degrees of freedom
Multiple R-squared:  0.02768,   Adjusted R-squared:  0.01805
F-statistic: 2.875 on 2 and 202 DF,  p-value: 0.05871
```

b1) In a standardized regression, there is no intercept since all variables are standardized to have a mean of 0.

b2) A 1 standard deviation increase in sales is associated with 0.1146 standard deviation increase in CEO salaries, holding ROE constant. This result is not statistically significant (p-value = 0.100615 > 0.10).

b3) A 1 standard deviation increase in ROE is associated with a 0.1268 standard deviation increase in CEO salaries, holding sales constant. This result is significant at the 10% level (p-value = 0.06950 < 0.10).

c) Since this is a standardized regression, we can compare the size of the coefficients on *sales* and *roe* directly, since both are expressed in terms of standard deviations. It appears that ROE has a larger correlation with expected salary in absolute terms since the coefficient on *roe* = 0.1268 > 0.1146 = coefficient on *sales*

Q1.2. Joint Significance Test

(a) Estimate a model of salary as a linear function of a constant, firm’s sales, average ROE, and an indicator for being in the financial sector. Then test the **joint significance** of the *ROE* and *sales* variables at the 1% significance level using the 5 steps of hypothesis testing. Conduct the hypothesis by hand, do not use any canned functions.

Hint: While you cannot answer the question using canned functions for credit, you can check your answer by comparing your manually calculated answer to the results you obtain from using the canned function linearHypothesis()

Step 1: State the null and alternative hypotheses.

H_0: $\beta_{\text{sales}} = \beta_{\text{roe}} = 0$

H_A: $\beta_{\text{sales}} \neq 0$ or $\beta_{\text{roe}} \neq 0$ or both

Step 2: Write down the two models the null hypothesis implies.

Unrestricted model: $salary = \beta_0 + \beta_1 sales + \beta_2 roe + \beta_3 finance + u$

Restricted model: $salary = \beta_0 + \beta_3 finance + u$

In [17]:

```
## Estimated the unrestricted and restricted models
# Unrestricted Model
reg2_ur<-lm(salary~sales+roe+finance, my_data)
summary(reg2_ur)

# Restricted Model
reg2_r<-lm(salary~finance, my_data)
summary(reg2_r)

# Calculate the F-statistic
r2.ur <- summary(reg2_ur)$r.squared
r2.ur
r2.r <- summary(reg2_r)$r.squared
r2.r
n <- nrow(my_data)
n
k <- 3
q <- 2
F.num <- (r2.ur-r2.r)/q
F.denom <- (1-r2.ur)/(n-k-1)
Fstat <- F.num/F.denom
Fstat

# Only for double-checking our answers (Using linearHypothesis())
linearHypothesis(reg2_ur, c("sales=0", "roe=0"))
```

```
Call:
lm(formula = salary ~ sales + roe + finance, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1341.0  -457.9  -241.8   103.7 13616.4 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept) 745.756604 243.041947   3.068  0.00245 **
sales         0.015196   0.008955   1.697  0.09127 . 
roe          22.012062  11.303814   1.947  0.05289 . 
finance     194.956419 232.191597   0.840  0.40211   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1361 on 201 degrees of freedom
Multiple R-squared:  0.03108,	Adjusted R-squared:  0.01662 
F-statistic: 2.149 on 3 and 201 DF,  p-value: 0.09528
```

```
Call:
lm(formula = salary ~ finance, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1028.9  -522.9  -250.9   116.1 13570.1 

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)    
(Intercept)  1251.90     109.09  11.476 <0.0000000000000002 ***
finance        93.01     230.28   0.404               0.687    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1376 on 203 degrees of freedom
Multiple R-squared:  0.000803,	Adjusted R-squared:  -0.004119 
F-statistic: 0.1631 on 1 and 203 DF,  p-value: 0.6867
```

0.0310783699630983

0.000803009826139923

205

3.1402681078019

A anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	203	384082868	NA	NA	NA	NA
2	201	372445276	2	11637592	3.140268	0.04540171

Step 3: Write the F-stat from the two regression outputs

$$F = \frac{(R^2_{UR} - R^2_{R})/q}{1 - R^2_{UR}/(n - k_{UR} - 1)} = 3.1403$$

In [5]:

```
# Calculate the critical value
c <- qf(0.99,2,201)
c
```

4.71231081512664

Step 4: Compare the F-stat to the correct critical value found in the F-table. Reject the null hypothesis if F-stat > critical value. Our F-stat is 3.1403 but our critical value at the 1% level is 4.7123. Since our F-stat 3.1403 < 4.7123 (critical value), we **Fail to reject** the null.

Step 5: Interpret. We fail to reject the null hypothesis that *sales* and *roe* are jointly not statistically significant in predicting CEO salaries at the 1% significance level.

Q1.3. Confidence Intervals for Prediction

(a) For this question, only use data from the finance sector *(hint, create a new filtered dataset that only includes observations where finance = 1)*. Specify and estimate a model to predict the average salary of an executive whose firm has an ROE of 9% with 4500 (meaning 4.5 billion USD) in sales. Use the change-of-variable approach demonstrated in Lecture 14 and Section 8 so that the intercept in your transformed model gives the average predicted salary for a CEO with *roe* = 9% and *sales = 4500*. Interpret your result in 1 sentence.

In [18]:
```
#create a filtered dataset only with observations from the finance sector
finance_data<-filter(my_data,my_data$finance==1)

# Generate transformed versions of our variables so that when we run the reg
# the intercept gives us the average predicted salary for a CEO with roe = .
finance_data$sales0<-finance_data$sales-4500
finance_data$roe0<-finance_data$roe-9

# Run the requested regression with the transformed versions of the variable
reg3 <- lm(salary~roe0+sales0, finance_data)
summary(reg3)
```

```
Call:
lm(formula = salary ~ roe0 + sales0, data = finance_data)

Residuals:
   Min     1Q Median     3Q    Max 
-819.1 -386.6 -207.2  -34.1 5102.4 

Coefficients:
              Estimate Std. Error t value     Pr(>|t|)    
(Intercept) 1419.24410  212.30711   6.685 0.0000000369 ***
roe0         -22.17543   26.07751  -0.850        0.400    
sales0         0.03261    0.02946   1.107        0.275    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 996.9 on 43 degrees of freedom
Multiple R-squared:  0.05562,	Adjusted R-squared:  0.01169 
F-statistic: 1.266 on 2 and 43 DF,  p-value: 0.2922
```

According to this model, the predicted salary of a CEO of a firm with *roe* = 9% and *sales* = 4.5 billion USD is 1,419.2441 (thousands USD) or $1,419,244.

(b) Construct a 95% confidence interval for the **predicted average** salary for a CEO with those characteristics. Interpret your result in 1 sentence. *(Hint: Use your results from your estimated regression in part a)*

In [7]:
```
# the constant from our estimated regression gives us the estimated test sta
sample_mean <- summary(reg3)$coefficients[1,1]
sample_mean_se <- summary(reg3)$coefficients[1,2]

# Find the critical value using the t-distribution with 43 dof and alpha = .
dof = nrow(finance_data)-2-1
```

```
c3 <-qt(0.025, 43, lower.tail=FALSE)

# Calculate the upper and lower bounds of the confidence interval
ci3_l<- sample_mean - c3*sample_mean_se
ci3_l
ci3_h<- sample_mean + c3*sample_mean_se
ci3_h
```

991.086000867626

1847.40219654031

We are 95% confident that the random interval [991.0860, 1847.4022] covers the true average predicted salary of a CEO of a firm with *roe* = 9% and *sales* = 4.5 billion USD

(c) Construct the 95% confidence interval for the **predicted salary of a specific CEO** (not the prediction for the average CEO) with those same characteristics as in parts a and b above (*roe* = 9%, *sales* = 4500)? Are you surprised that the intervals differ between questions 1.3.b and 1.3.c? How do these confidence intervals differ? Explain your answer in a maximum of 4 sentences. *(Hint: See Lecture 14 and Section 8)*

In [19]:
```
# To calcuate the CI for a specific CEO with these characteristics
# We need to calculate the variance of the prediction error
var_3c<-summary(reg3)$coefficients[1,2]^2 + sigma(reg3)^2
se_3c<-sqrt(var_3c)
ci_3c_l<-sample_mean-se_3c*c3
ci_3c_l
ci_3c_h<-sample_mean+se_3c*c3
ci_3c_h

# alternative method if students use rounding, just the output printed in th
var_3c_rounded <- (212.30711*212.30711) + (996.9*996.9)
se_3c_rounded <- sqrt(var_3c_rounded)
ci_3c_l_rounded <-sample_mean-se_3c_rounded*c3
ci_3c_l_rounded
ci_3c_h_rounded<-sample_mean+se_3c_rounded*c3
ci_3c_h_rounded
```

-636.350240148602

3474.83843755654

-636.282638886175

3474.77083629411

To calculate the CI for a specific CEO with these characteristics, we need to quantify the variance of the prediction error:

$$Var(\hat{e}^0) = Var(\hat{sales}^0) + \hat{\sigma}^2$$

We then construct our confidence interval using $se(\hat{e}^0)$ rather than $se(\hat{sales}^0)$, which is what we used in part b when we constructed a confidence interval for the average salary of CEOs with the specified characteristics. We find that there is a 95% chance that the random interval [-636.3502, 3474.8384] (or [-636.2826, 3474.7708] with rounding) covers the predicted salary for a specific CEO with *roe* = 9% and *sales* = 4500.

As discussed in Lecture 14/Section 8, a confidence interval for the average person with certain characteristics is not the same as a confidence interval for a particular person with those specific characteristics because the former only takes into the account the sampling error in our prediction which stems from the fact that we have estimated β_0 from a random sample whereas the latter also takes into account the variance in the population (unobserved) error, which captures the unobserved factors that affect y. Therefore it is not surprising that the confidence interval for the predicted salary of a specific CEO with these characteristics is much wider than the confidence interval for the predicted salary of the average CEO with these characteristics.

Q1.4. Comparing Non-Nested Models

We are selecting between model (1A.) and model (1B.) below. Estimate both models, creating any variables you need. Based on your estimated output using linear regression analysis, which model would you choose to use to most accurately predict CEO salaries? Justify your answer in a maximum of 2 sentences.

$$(Model\ 4A)\ \ salary_i = \beta_0 + \beta_1 ROE_i + \beta_2 sales_i + \beta_3 finance_i + u_i$$

$$(Model\ 4B)\ \ salary_i = \gamma_0 + \gamma_1 ROE_i + \gamma_2 log(sales_i) + \gamma_3 finance_i + v_i$$

Hint: Log() indicates the natural log ln()

In [20]:

```
## Estimate Model 4A

reg4.A <- lm(salary~sales+roe+finance, my_data)
summary(reg4.A)

## Estimate Model 4B, including creating all necessary variables
#create log sales
my_data$lsales<-log(my_data$sales)

# Estimate model 4B
reg4.B <-  lm(salary~lsales+roe+finance, my_data)
summary(reg4.B)
```

```
Call:
lm(formula = salary ~ sales + roe + finance, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1341.0  -457.9  -241.8   103.7 13616.4 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept) 745.756604 243.041947   3.068  0.00245 **
sales         0.015196   0.008955   1.697  0.09127 . 
roe          22.012062  11.303814   1.947  0.05289 . 
finance     194.956419 232.191597   0.840  0.40211   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1361 on 201 degrees of freedom
Multiple R-squared:  0.03108,	Adjusted R-squared:  0.01662 
F-statistic: 2.149 on 3 and 201 DF,  p-value: 0.09528
Call:
lm(formula = salary ~ lsales + roe + finance, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1150.7  -425.2  -200.3    63.4 13693.3 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -1399.14     835.59  -1.674  0.09560 . 
lsales        267.46      94.51   2.830  0.00513 **
roe            24.45      11.22   2.180  0.03041 * 
finance       155.94     228.95   0.681  0.49658   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1344 on 201 degrees of freedom
Multiple R-squared:  0.05486,	Adjusted R-squared:  0.04075 
F-statistic: 3.889 on 3 and 201 DF,  p-value: 0.009889
```

Since these are non-nested models, we can compare the Adjusted R^2 and choose the model with the highest Adjusted R^2. The Adjusted R^2 in model 4A is 0.01662 wherease the Adjusted R^2 in Model 4B is 0.04075, therefore I would choose model 4B since this model explains more of the variation in CEO salaries.

Ex2. Choosing between Y and log(Y)

We want to select the best model to use for future labor market analysis. We are selecting between model (2A) and model (2B) below, where log() indicates the natural log ln().

$$(Model\ 2A)\ \ salary_i = \theta_0 + \theta_1 ROE_i + \theta_2 sales_i + \theta_3 finance_i + u_i$$

$$(Model\ 2B)\ \ log(salary_i) = \eta_0 + \eta_1 ROE_i + \eta_2 sales_i + \eta_3 finance_i + v_i$$

Estimate both models in exercise 2, creating any variables you need. Which model would you choose to use henceforth? Show all code and work that you used to answer this question and explain your result in a maximum of 2 sentences.

*Hint: Remember that you **cannot** simply compare R^2 or Adjusted R^2 to choose between these two models since they use two different outcome variables with fundamentally different amounts of variation (*salary* vs *log(salary)*). Instead, you should calculate an alternative R^2 for the log model that represents how much variation in *salary* is explained by the log model (Model 2B). Then, select the model that explains the most variation in *salary*. (See Lecture 15 and Section 8)*

In [21]:

```
#Estimate both models in exercise 2, creating any variables you need.
reg2A <- lm(salary~sales+roe+finance, my_data)
summary(reg2A)
reg2B <- lm(lsalary~sales+roe+finance, my_data)
summary(reg2B)

#Predict log(y) from the log model
my_data$lsalary_pred<-reg2B$fitted.values

# Convert your predictions of log(y) to y, where yhat = e^ln(y)hat*e^(sigmah
# Note that sigma(reg) gives you the residual standard error from your regre
e2b_adj_term <- exp(0.5*sigma(reg2B)^2)
my_data$salary_pred_modelB <-exp(my_data$lsalary_pred)*e2b_adj_term

#Find the correlation and square it to calculate the alternative R^2 for the
alt_r2_modelB <- (cor(my_data$salary,my_data$salary_pred_modelB))^2
alt_r2_modelB

# You can directly compare this alternative R^2 for the log model to the R^2
#(Optional) Also manually calculate the R^2 from the linear model
my_data$salary_pred_modelA<-reg2A$fitted.values
alt_r2_modelA <-(cor(my_data$salary,my_data$salary_pred_modelA))^2
alt_r2_modelA
```

```
Call:
lm(formula = salary ~ sales + roe + finance, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1341.0  -457.9  -241.8   103.7 13616.4 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept) 745.756604 243.041947   3.068  0.00245 **
sales         0.015196   0.008955   1.697  0.09127 . 
roe          22.012062  11.303814   1.947  0.05289 . 
finance     194.956419 232.191597   0.840  0.40211   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1361 on 201 degrees of freedom
Multiple R-squared:  0.03108,   Adjusted R-squared:  0.01662 
F-statistic: 2.149 on 3 and 201 DF,  p-value: 0.09528
Call:
lm(formula = lsalary ~ sales + roe + finance, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.48056 -0.26920 -0.04222  0.24382  2.74293 

Coefficients:
               Estimate  Std. Error t value             Pr(>|t|)    
(Intercept) 6.492966878 0.092883605  69.904 < 0.0000000000000002 ***
sales       0.000015269 0.000003422   4.462            0.0000135 ***
roe         0.017268376 0.004319991   3.997            0.0000899 ***
finance     0.228014364 0.088736915   2.570               0.0109 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5202 on 201 degrees of freedom
Multiple R-squared:  0.1537,    Adjusted R-squared:  0.141 
F-statistic: 12.16 on 3 and 201 DF,  p-value: 0.0000002386
```

0.0202008443043515

0.0310783699630983

We calculate an alternative R^2 for Model 2B and find that it explains 2.0201% of the variation in salary. Comparing this alternative R^2 for Model B to the R^2 for Model 2A, which is 3.108%, we prefer model A since it explains more of the variation in salary.

Ex3: CEO Salaries by Sector (Finance/Non Finance)

Q3.1: Testing for Separate Equations

Estimate model 2A separately for executives in the finance sector and for those not in the finance sector. Formally test at the 10% significance level, using the five steps of hypothesis testing, whether the regressions should be estimated separately or whether we can pool the data like we have been doing so far. (Hint: See Lecture 17 and Section 9)

We can test whether the regressions are the same for CEOs in the finance sector versus non-finance sectors by running a **Chow test**.

Step 1: State the hypotheses

H_0: Regression should be pooled

H_A: Regressions should be run separately for finance and non-finance firms

Step 2: The Chow test statistic is given by:

$$F = \frac{(SSR_{pooled} - (SSR_{finance} + SSR_{nonfinance}))/(k+1)}{(SSR_{finance} + SSR_{nonfinance})/(n_{finance} - k - 1 + n_{nonfinance} - k - 1)}$$

To calculate the F stat, we need to obtain the SSR values for each of the three regressions. You could do this in one of three different ways

1. Calculate as $SSR = \sum(y_i - \hat{y}_i)^2$, which yields $F = 0.5819$ (Demonstrated below)
2. Rearrange the $\hat{\sigma}$ formula to solve for $SSR = \hat{\sigma}^2(n-k-1)$ using the `sigma()` function, which yields $F = 0.5819$
3. Do 2. but by manually reading off the residual standard error value from the output table, which yields $F = 0.5757$

Note that results from method 3 differs from 1 and 2, due to rounding issues and given we're multiplying by such large degrees of freedom. All approaches receive full credit when done correctly.

In [28]:
```
# First estimate Model 2A with all of the data pooled together
reg2A <- lm(salary~sales+roe, my_data)
summary(reg2A)

## Create two new datasets, one with all observations where finance = 1
# the other with all observtions where finance = 0
finance_data<-filter(my_data,my_data$finance==1)
nonfinance_data<-filter(my_data,my_data$finance==0)

# Estimate Model 2A only with nonfinance data
reg2A_nf<-lm(salary~sales+roe,nonfinance_data)
summary(reg2A_nf)

# Estimate Model 2A only with finance data
reg2A_f<-lm(salary~sales+roe,finance_data)
summary(reg2A_f)

## 3 equivalent methods for calculating the F-statistic
```

```r
# Method 1
k.e3 <- 2
ssr_pooled <- sum((fitted(reg2A) - my_data$salary)^2)
ssr_finance <- sum((fitted(reg2A_f) - finance_data$salary)^2)
ssr_nonfinance <- sum((fitted(reg2A_nf) - nonfinance_data$salary)^2)

F_num <- (ssr_pooled - (ssr_finance + ssr_nonfinance))/(k.e3 + 1)
F_den <- (ssr_finance + ssr_nonfinance)/(nrow(finance_data) - k.e3 - 1 + nrc
F <- F_num / F_den
paste("(Method 1) F Stat value is", round(F,4))

# Method 2
var.res <- (summary(reg2A)$sigma)^2
SSR.R <- var.res*(nrow(my_data) - k.e3 - 1)

var.res.finance<- (summary(reg2A_f)$sigma)^2
SSR.finance <- var.res.finance*(nrow(finance_data) - 2 - 1)

var.res.notfinance <- (summary(reg2A_nf)$sigma)^2
SSR.notfinance <- var.res.notfinance*(nrow(nonfinance_data) - 2 - 1)

F.num2 <- (SSR.R - (SSR.finance + SSR.notfinance))/(k.e3 + 1)
F.denom2 <- (SSR.finance + SSR.notfinance)/(nrow(finance_data) - k.e3 - 1 +

F.2 <- F.num2/F.denom2
paste("(Method 2) F Stat value is", round(F.2,4))

# Method 3
SSR_pooled<- 1360 * 1360 * 202
SSR_finance <- 996.9 * 996.9 * 43
SSR_nonfinance <- 1444 * 1444 * 156

top<-(SSR_pooled-(SSR_finance + SSR_nonfinance))/(k.e3 + 1)
bottom<-(SSR_finance + SSR_nonfinance)/(nrow(finance_data) - k.e3 - 1 + nrow

F.3<-top/bottom
paste("(Method 3) F Stat value is", round(F.3,4))
```

```
Call:
lm(formula = salary ~ sales + roe, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1346.4  -484.7  -229.1   134.1 13574.3 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 822.893885 224.845205   3.660 0.000322 ***
sales         0.014732   0.008932   1.649 0.100615    
roe          20.257893  11.100955   1.825 0.069496 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1360 on 202 degrees of freedom
Multiple R-squared:  0.02768,	Adjusted R-squared:  0.01805 
F-statistic: 2.875 on 2 and 202 DF,  p-value: 0.05871
```

```
Call:
lm(formula = salary ~ sales + roe, data = nonfinance_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1198.0  -450.9  -266.0   112.8 13601.9 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) 654.646974 268.241367   2.441   0.0158 *
sales         0.013799   0.009752   1.415   0.1591  
roe          27.609756  12.675902   2.178   0.0309 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1444 on 156 degrees of freedom
Multiple R-squared:  0.03976,	Adjusted R-squared:  0.02744 
F-statistic: 3.229 on 2 and 156 DF,  p-value: 0.04225
Call:
lm(formula = salary ~ sales + roe, data = finance_data)

Residuals:
   Min     1Q Median     3Q    Max 
-819.1 -386.6 -207.2  -34.1 5102.4 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1472.09130  470.30818   3.130  0.00314 **
sales          0.03261    0.02946   1.107  0.27456   
roe          -22.17543   26.07751  -0.850  0.39983   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 996.9 on 43 degrees of freedom
Multiple R-squared:  0.05562,	Adjusted R-squared:  0.01169 
F-statistic: 1.266 on 2 and 43 DF,  p-value: 0.2922
```

373751592.530085

42736751.5375795

325359140.946583

'(Method 1) F Stat value is 1.0192'

'(Method 2) F Stat value is 1.0192'

'(Method 3) F Stat value is 1.0101'

Step 3: We find the critical value from the F-distribution (or F-table) at the 10% confidence level, with $k+1=3$ and $n_{finance}-k-1+n_{nonfinance}-k-1 = 46-2-1+159-23-1 = 199$ degrees of freedom. The critical value is $c=2.1115$.

In [26]:
```
qf(0.90, 3, 199)
```

2.1115081117763

Step 4: Our decision rule is we reject the null hypothesis if our F-stat > critical value. In this case, 1.0192 < 2.1115 so we fail to reject the null hypothesis.

Step 5: Interpret: We fail to reject the null hypothesis that we can pool the data at the 10% significance level.

Q3.2 Interaction Terms

I would like to know whether the correlation between firm sales and CEO salaries differs significantly depending on whether the company is in the finance sector.

$$(Model\ 3)\ \ wage_i = \beta_0 + \beta_1 ROE_i + \beta_2 sales_i + \beta_3 finance_i + u_i$$

Estimate a model, by adjusting Model 3, that enables you to test this and please interpret your findings in a maximum of 2 sentences. Compare the p-value for the estimated coefficient of interest at the 5 percent significance level to conclude whether you reject the null hypothesis of no heterogeneity in the effect of sales on salary for finance and non-finance firms, against a two-sided alternative, holding all else equal. *(Hint: Generate an interaction term and add it to the regression, see Lecture 17 and Section 9).*

In [27]:
```
#generate an interaction term for sales X finance
my_data$salesFin<-my_data$sales*my_data$finance
reg3.2 <- lm(salary~roe+sales+finance+salesFin, my_data)
summary(reg3.2)
```

```
Call:
lm(formula = salary ~ roe + sales + finance + salesFin, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1195.7  -462.0  -250.8   100.7 13607.7 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) 747.00207  243.29626   3.070  0.00244 **
roe          22.57147   11.33884   1.991  0.04788 * 
sales         0.01361    0.00920   1.479  0.14068   
finance      14.30803  330.80757   0.043  0.96554   
salesFin      0.03092    0.04028   0.767  0.44373   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1363 on 200 degrees of freedom
Multiple R-squared:  0.03392,   Adjusted R-squared:  0.0146 
F-statistic: 1.756 on 4 and 200 DF,  p-value: 0.1393
```

We cannot reject the null hypothesis that there is no significant difference in the correlation between firm sales and CEO salaries in the finance versus non-finance sectors. We see this because the p-value for the estimated coefficient on our interaction term *salesFin* = 0.44373 which is greater than all conventional significance levels.

```
In [ ]:
```