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 [53]:
```
# Insert your code here
```

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 [54]:
```
# Insert your code here
```

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 [55]:
```
# Insert your code here
```

Type your answer here

Q1.2. Joint Significance Test

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()

In [56]:
```
# Insert your code here
```

Type your answer here

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 [57]:
```
# Insert your code here
```

Type your answer here

(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 [58]:
```
# Insert your code here
```

Type your answer here

(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 [59]:
```
# Insert your code here
```

Type your answer here

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 [60]:
```
# Insert your code here
```

Type your answer here

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 [61]:
```
# Insert your code here
```

Type your answer here

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)*

In [62]:
```
# Insert your code here
```

Type your answer here

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 [63]:
```
# Insert your code here
```

Type your answer here

In [ ]:
```

```