Applied Stats I: Exam 2

Due: December 8, 2023

Instructions

- Please read carefully: You have from 09:00 Wednesday December 6 until 08:59 Friday December 8 to complete the exam. Please export your answers as a single PDF file and include all code you produce in a supporting `R` file, which you will upload to Blackboard. The exam is open book; you can consult any materials you like. You must not collaborate with or seek help from other students. In case of questions or technical difficulties, you can contact Professor Ziegler via email. You should write-up your answers in R and LaTeX as you would for a problem set. Please make sure to concisely number your answers so that they can be matched with the corresponding questions.

Question 1

Define and describe why the following four (4) terms are important to hypothesis testing and/or regression. You can earn full credit with just two or three sentences, but please be specific and thorough.

`Partial F-test`:
It helps to determine if a subset of the independent variables collectively contributes significantly to explaining the variation in the dependent variable.

`Test statistics`:
The test statistic is a numerical summary of a sample that is used in hypothesis testing. It quantifies the difference between the observed data and what would be expected under the null hypothesis.

`Constituent term`:
Refers to an individual component or predictor in the regression model. Each term represents the contribution of a specific variable to the model's prediction.

`Categorical data / dummy variables`:
Is important because it helps us to include qualitative information into quantitative models to assess their impact on the dependent variable.

Question 2

We want to estimate the impact of economic, social, and political factors (GDP per capita, average years of education, and democracy/non-democracy) on foreign direct investment (FDI) into a country, which is measured in millions of dollars. We have already processed our data as well as run our regression (N= 1000), and we get the following output. Please consult the table below, which presents the estimated coefficients and standard errors from our model, to answer the following questions. Also, note that the economic variables (GDP per capita and FDI) are presented in constant-year US Dollars (2010, $), while Education equals the average number of years in school students spend and Democracy is a binary dummy variable (1=Democracy, 0=Non-democracy).

(a) Interpret the coefficients for GDP and Education.

The coefficient for `GDP` is -2, this suggest that on average a one-unit increase on GDP per capita is associated with a decrease of 2 million dollars in FDI, holding other factors constant. A higher GDP might be associated with lower foreign direct investment.

The coefficient for `Education` is -4.298, this suggest that on average a one-unit increase in the average years of education is associated with a decrease of 4.298 million dollars in FDI, holding other factors constant. Higher levels of education might be associated with lower foreign direct investment.

(b) The author claims that she 'cannot reject the null hypothesis that GDP has no effect on FDI (H0 : GDP = 0). Using the coefficient estimate and the standard error for GDP construct a 95% confidence interval for the effect of GDP on FDI. Based on the confidence interval, do you agree with the author? Explain your answer.

Assuming a two-tailed test with a standard normal distribution:

Confidence Interval for GDP = Coefficient Estimate for GDP +- (1.96 x Standard Error for GDP)

If the confidence interval includes zero, it suggests that the effect of GDP on FDI is not statistically different from zero, supporting the author's claim. If zero is not included in the interval, it would indicate a statistically significant effect.

```
#Given values
BGDP <- -2
Se_GDP <- 0.00007
critical_value <- qnorm(0.975)  # For a 95% confidence interval

# Calculate margin of error
margin_of_error <- critical_value * Se_GDP

Calculate confidence interval
confidence_interval <- c(BGDP - margin_of_error, BGDP + margin_of_error)

# Print the result
cat("95% Confidence Interval for the effect of GDP on FDI:",
    confidence_interval, "\n")

```

Confidence Interval for GDP = -2 +- (1.96 x 0.00007) Confidence Interval for GDP = -2 +- 0.0001372

The 95% confidence interval for the effect of GDP on FDI is approximately Confidence Interval for GDP = -2.0001372 - 1.9998628

The author claims she cannot reject the null hypothesis that GDP has no effect on FDI H0 : GDP = 0 Based on the values of our confidence interval, there is evidence to reject the null hypothesis. I am not agree with the author as there is evidence supporting an effect of GDP on FDI.

(c) Calculate the difference in predicted FDI between low and high values of GDP for non-democratic countries holding Education constant at its sample mean. Use 12.04 as the mean of Education and use +/- one standard deviation around the mean of GDP (from 6378.56 to 41031.78) for low and high values of GDP respectively.

```
#Option C
# Given values
mean_Education <- 12.04

# Calculate mean GDP as the midpoint of the range
mean_GDP <- (6378.56 + 41031.78) / 2

# Calculate standard deviation of GDP
std_dev_GDP <- (41031.78 - 6378.56) / 2

# Calculate high and low values of GDP (one standard deviation around the mean)
low_GDP <- mean_GDP - std_dev_GDP
high_GDP <- mean_GDP + std_dev_GDP

# Predicted FDI for low and high values of GDP
predicted_FDI_low <- BGDP * low_GDP + mean_Education
predicted_FDI_high <- BGDP * high_GDP + mean_Education

# Difference in predicted FDI
difference_in_predicted_FDI <- predicted_FDI_high - predicted_FDI_low

# Print the result
cat("Difference in predicted FDI between low and high values of GDP:", difference_in_predicted_FDI, "\n")


```

The difference in predicted FDI between low and high values of GDP for non-democratic countries, holding education constant at its sample mean, is estimated to be approximately -$69,306.44. This suggests that, on average, a one-unit increase in GDP is associated with a decrease in predicted FDI by approximately $69,306.44, when other variables are held constant.

Question 3

Many of the wells used for drinking water in Bangladesh and other South Asian countries are contaminated with natural arsenic, affecting an estimated 100 million people. Arsenic is a cumulative poison, and exposure increases the risk of cancer and other diseases, with risks estimated to be proportional to exposure.

We performed a regression analysis with the data to understand the factors that predict the arsenic level of 1000 households' drinking water. Your outcome variable arsenic is a continuous measure of household i's arsenic level in units of hundreds of micrograms per liter.

We estimated models with the following inputs:

The distance (in kilometers/100) to the closest known commercial factory

Depth of respondent's well (binary variable; deep=1, not deep=0)

(a) First, we successfully estimated an additive model with well depth and distance to the nearest factory as the two predictors of a household's arsenic level. The estimated coefficients are found in the first column of the table above. Interpret the estimated coefficients for the intercept and each predictor.

Model 1

Intercept (1.72) Represents the estimated arsenic level when both the well depth and distance to the nearest factory are zero.

Well Depth (0.67) For each unit increase in the binary variable indicating a deep well (well_depth), the estimated arsenic level increases by 0.67 units. This assumes other variables are held constant.

Distance to the Nearest Factory (-5.00) For each unit increase in the distance to the nearest factory (measured in kilometers/100), the estimated arsenic level decreases by 5.00 units. This suggests that greater distance from a factory is associated with lower arsenic levels.

Model 2

Intercept (2.32) Represents the estimated arsenic level when both the well depth and distance to the nearest factory are zero.

Well Depth (0.07) For each unit increase in the binary variable indicating a deep well (well_depth), the estimated arsenic level increases by 0.07 units. This assumes other variables are held constant.

Distance to the Nearest Factory (-5.49) For each unit increase in the distance to the nearest factory (measured in kilometers/100), the estimated arsenic level decreases by 5.49 units. This suggests that greater distance from a factory is associated with lower arsenic levels.

well_depth:dist100 (0.49) This term represents the change in the effect of distance on arsenic levels when well depth changes. In this case, for each unit increase in the interaction term, the effect of distance on arsenic levels increases by 0.49 units when the well is deep.

(b) Does the coefficient estimate for the closest known factory vary based on whether or not a house has a deep well? If so, change your interpretation of the estimated coefficients in part (a) to conform with the interactive model in column 2 of the table above.

In Model 2, the coefficient for the interaction term "well_depth:dist100" is 0.49. This suggests that the effect of the distance to the nearest factory on arsenic levels varies depending on whether or not a house has a deep well.

Model 2

Intercept (2.32) This is the estimated arsenic level when both well depth and distance are zero (assuming the reference category for well depth)

Well Depth (0.07) This is the estimated change in arsenic level associated with a one-unit increase in well depth when the distance is zero (or in the reference category for distance).

Distance to the Nearest Factory (-5.49) This is the estimated change in arsenic level associated with a one-unit increase in distance when well depth is zero (or in the reference category for well depth).

well_depth:dist100 (0.49) This represents how the effect of distance on arsenic levels changes for each one-unit increase in well depth.

What is the appropriate test to determine whether we should model the relationship between distance, well depth, and arsenic levels using an additive or interactive model?

To determine whether an interactive model is more appropriate than an additive model, we can perform a hypothesis test for the interaction term. The null hypothesis would be that the interaction term is zero, indicating no interaction effect. The alternative hypothesis would be that the interaction term is not zero, suggesting an interaction effect.

What information would you need to perform that test? The information given in table 2 is enough. The coefficient for the Interaction term and the Standard error of well_depth:dist100

If the test shows that the interaction term is statistically significant (p-value ¡ 0.05, for example), it suggests that an interactive model might be more appropriate than an additive model.

(c) Using the ‘preferred’ model from Part B, compute the average difference in arsenic levels between two households that have a deep well (=1), but one is closer to a factory (dist100 = 0.36) than the other (dist100 = 2.08).

```
#Option C

# Given coefficients
B0 <- 2.32
B_well_depth <- 0.07
B_dist100 <- -5.49
B_well_depth_dist100 <- 0.49

# Values for two households
dist100_1 <- 0.36
dist100_2 <- 2.08
well_depth <- 1  # Assuming both households have a deep well

# Calculate arsenic levels for each household
arsenic_level_1 <- B0 + (B_well_depth * well_depth) + (B_dist100 *
    dist100_1) + (B_well_depth_dist100 * well_depth * dist100_1)
arsenic_level_2 <- B0 + (B_well_depth * well_depth) + (B_dist100 *
    dist100_2) + (B_well_depth_dist100 * well_depth * dist100_2)

# Calculate average difference
average_difference <- (arsenic_level_1 - arsenic_level_2) / 2

# Print the result
cat(”Average Difference in Arsenic Levels:”, average_difference, ”\n”)

```

Average Difference in Arsenic Levels: 4.3

Question 4

This data set presents information on 33 lambs, of which 11 are ewe lambs, 11 are weather lambs, and 11 are ram lambs. These lambs grazed together in the same pasture and were treated similarly in all ways. The variables of interest are presented in the table below.

The objective is to determine whether differences in Fatness could be attributed to Group while accounting for Weight. Information on the data and the model fit in R are given below:

(a) Write out the fitted model for a ram lamb using the estimated coefficients.

$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + E \quad \text{(Multiple Linear Regression Model)}$$

$$\text{Fatness} = B_0 + B_1 \cdot \text{Weight} + B_2 \cdot \text{Group dummy.1} + B_3 \cdot \text{Group dummy.2}$$

$$\text{Fatness} = -18.1368 + 2.2980 \cdot \text{Weight} - 8.3622 \cdot \text{Group dummy.1} - 4.0716 \cdot \text{Group dummy.2}$$

(b) What is the predicted Fatness index of a weather lamb that weighs 10kg?

$$\text{Fatness} = -18.1368 + 2.2980 \cdot 10 - 8.3622 \cdot 0 - 4.0716 \cdot 1$$

Fatness index of approximately 0.7716

(c) Which lamb group has the highest Fatness index for every weight?

$$\text{Fatness} = -18.1368 + 2.2980 \cdot 10 - 8.3622 \cdot 0 - 4.0716 \cdot 0$$

Fatness index of approximately 4.8432

$$\text{Fatness} = -18.1368 + 2.2980 \cdot 10 - 8.3622 \cdot 1 - 4.0716 \cdot 0$$

Fatness index of approximately -3.519

$$\text{Fatness} = -18.1368 + 2.2980 \cdot 10 - 8.3622 \cdot 0 - 4.0716 \cdot 1$$

Fatness index of approximately 0.7716

Question 5

Please select the most appropriate option to correctly answer each question. The coefficients in an ordinary least squares regression model...

(1) are generalized additive estimates

(2) are maximum likelihood estimates

(3) **minimize the residual sum of squares**

(4) maximize the regression sum of squares

We can calculate our standard errors by taking the square root of the off-diagonal elements in our variance-covariance matrix.

(1) **True**

(2) False

For explanatory variables with multi-collinearity, the corresponding estimated slopes have ... standard errors.

(1) **Larger**

(2) Smaller

(3) The same

Suppose you are interested in knowing the different impact of age (continuous) by educational background (categorized as arts or science/engineering) on a job candidate's potential salary (continuous). Which test or technique would you use?

(1) Simple bivariate linear regression model

(2) Additive (salary = age + education) regression model

(3) **Interactive (salary = age * education) regression model**

(4) Interactive (education = age * salary) regression model

Question 6

Suppose we are interested in studying how individual personal wealth varies by age. Figure 1 plots the total amount of money an individual has in personal assets (the y-axis is in thousands of $) by their age.

What concerns might we have about using personal wealth in USD ($) 'as is' in a model that regresses 'amount of individual personal wealth' on 'age'? How could we address these concerns?

(1) **Non-linearity** The relationship between age and personal wealth might not be linear. As we observe in Figure 1 there is a curve shape, this could be because as age increases, personal wealth may not increase or decrease at a constant rate. For addressing non-linearity, a **polynomial** term such as Age squared could be incorporated into the model. This allow to capture non-linear relationships.

(2) **Scale issues** Personal wealth can vary widely, the scale of the variables might be a larege compared to other variables in the model. It can be difficult for interpretation. To address scale issues a **Log Transformation** can be applied. This transformation not only facilitates easier interpretation of the variable's effects but also stabilizes the variance, addressing scale-related challenges.

(3) **Heteroscedasticity** The presence of heteroscedasticity, where the variance of personal wealth may vary across different age groups, can violate the assumption of homoscedasticity. This violation could impact the reliability of standard errors and hypothesis tests in the regression model. In the context of heteroscedasticity, the **Log Transformation** is useful. As this transformation contributes to stabilizing the variance, addressing the violation of homoscesasticity and improving the reliability of standard erros and hypothesis tests in the regression model.