# Applied Stats I: Exam 2

Due: December 9, 2022

**Instructions**

*Please read carefully:* You have from **09:00 Wednesday December 7** until **08:59 Friday December 9** 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 collborate 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

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 coeffients 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).

Table 1: Estimated coefficients from regression predicting variation in FDI.

| | Estimate | Std. Error |
|---|---|---|
| **(Intercept)** | 52.72 | 24.84 |
| **GDP** | -3 | 0.00013 |
| **Democracy** | -2.99 | 4.36 |
| **Education** | -6.38 | 2.047 |

a) Interpret the coefficients for GDP and Democracy.

b) The author claims that she 'cannot reject the null hypothesis that GDP has no effect on FDI ($H_0 : \beta_{GDP} = 0$) . 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.

c) Calculate the difference in predicted FDI between low and high values of Education for non-democratic countries holding GDP constant at its sample mean. Use 24860.42 as the mean of GDP and use +/- one standard deviation around the mean of Education (from 10.96 to 13.02) for low and high values of Education respectively.

# Question 2

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?

Figure 1: Personal wealth and age

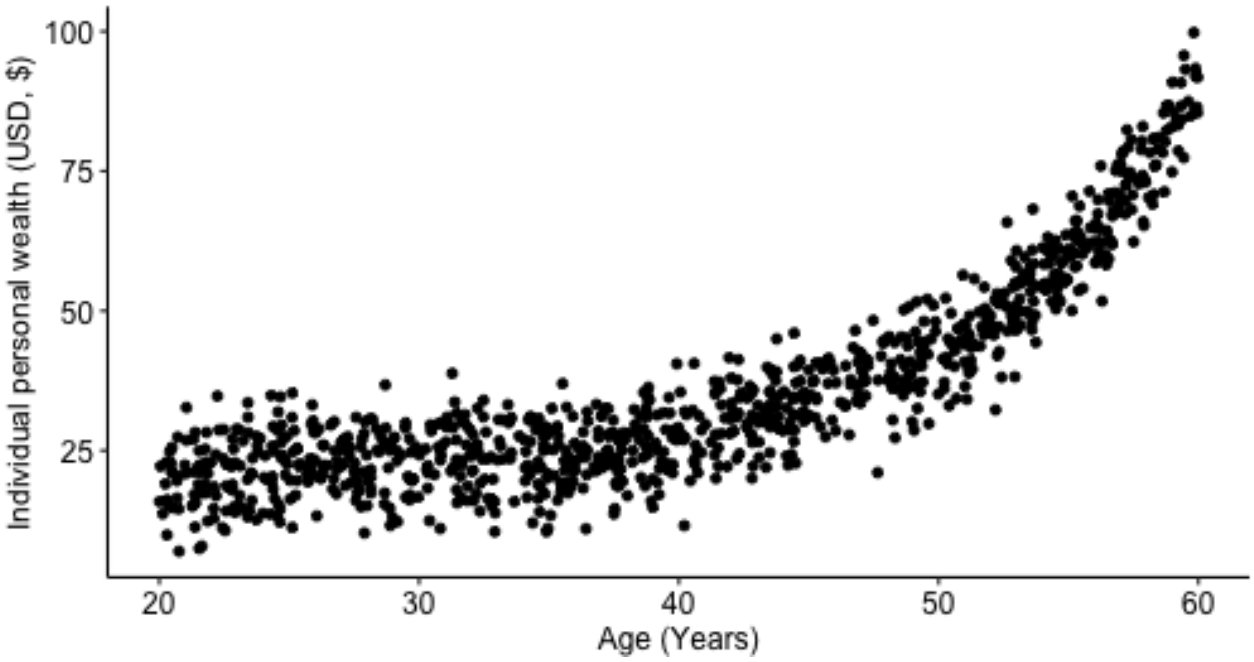

Table 2: Estimated coefficients from regression predicting arsenic levels.

| | Model 1 | Model 2 |
|---|---|---|
| (Intercept) | 0.25 (0.90) | 2.05 (1.58) |
| well_depth | 0.73 (0.86) | −1.05 (1.55) |
| dist100 | −1.01 (0.14)*** | −2.48 (1.07)* |
| well_depth:dist100 | | 1.46 (1.06) |
| $R^2$ | 0.05 | 0.05 |
| Adj. $R^2$ | 0.05 | 0.05 |
| Num. obs. | 1000 | 1000 |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

# 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.

(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. 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? What information would you need to perform that test?

(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 = 0.37) than the other (dist100 = 2.11).

# Question 4

This data set presents information on 33 lambs, of which 11 are ewe lambs, 11 are wether 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.

Table 3: Outcome and predictors for model.

| Variable | Description |
|---|---|
| Fatness | Continuous measure of leanness |
| Weight | Weight of lamb (kg) |
| Group | Factor (ewe, wether, ram) |

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:

```
> names(lambs)
 [1] "Fatness"   "Weight"   "Group"

> n=33
> Group.dummy.1=rep(0,n)
> Group.dummy.1[Group=="Wether"]=1
> Group.dummy.2=rep(0,n)
> Group.dummy.2[Group=="Ram"]=1

> lm.out=lm(Fatness ~ Weight + Group.dummy.1 + Group.dummy.2)
> summary(lm.out)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -18.1368     3.5213  -5.151 1.67e-05 ***
Weight          2.2980     0.2248  10.223 3.99e-11 ***
Group.dummy.1  -8.3622     0.9641  -8.674 1.50e-09 ***
Group.dummy.2  -4.0716     0.9045  -4.502 0.000101 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.102 on 29 degrees of freedom
Multiple R-squared: 0.8206,Adjusted R-squared: 0.8021
F-statistic: 44.23 on 3 and 29 DF,  p-value: 6.075e-11
```

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

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

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

# Question 5

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.

a) Partial F-test

b) Residuals

c) Categorical data/dummy variables

d) Constituent term

# Question 6

Please select the most appropriate option to correctly answer each question.

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

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

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

Which of the following plots is used to check for normality in the assumptions of linear regression?

1. Scatterplot between residuals and X
2. Scatterplot between residuals and Y
3. Histogram of Y
4. QQ plot of residuals