Exam 2-Darragh Kane O Toole-18431115

Applied Stats/Quant Methods 1

Due: December 9th, 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

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) Residuals- Residuals can be used to determine if variables are related by measuing the difference etween observed and fitted values.

(b) Categorical data- This a type of data that is representing a set of characteristics or traits,examples of this could be nationality or gender where there is a defined amout of groups which you cna be assigned to. It has no order or rank.

Dummy variables- These are when you use a categroical variable as a predictor in a regression to seperate a group to see if the effect varies within the groups characteristics. An example fo this could be that without rnak at a job considered better work leads to mre income but if a dummy variable is implemented it can show if this effect is true for managers and not managers adding insight to data.

(c) Test statistic- This is a statistic that sumarrises how much data is different from what would be expected to observe if the null hypothesis were true

(d) Constituent term-This is at times also called "main effects" as in a multiple Regression are the variables which represent a portion of the effect.

Table 1: Estimated coefficients from regression predicting arsenic levels.

	Model 1	Model 2
(Intercept)	−1.83 (1.21)	1.11 (2.08)
well_depth	6.51 (1.15)***	3.60 (2.04)
dist100	−2.86 (0.19)***	−5.17 (1.35)***
well_depth:dist100		2.30 (1.33)
R^2	0.20	0.20
Adj. R^2	0.20	0.20
Num. obs.	1000	1000

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

Figure 1:

Question 2

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.

The intercept indicates there is rarely arsenic in the water overall as its a negative number.
Based of the table well depth in the additive model leads to(6.51) increased arsenice in the water while distance to the well (-2.86) decreased the arsenic in the water.

(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?

A partial F test is an apropriate test to determine if an interactive model is apporriate to use in this data,If variance is constant the estimate is better with an interaction model. To do an a partial F test you need mean squares(treatement)/mean squares (error)

(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 = 0.42) than the other (dist100 = 2.12).

```
#general model 1 <- -1.83+ (6.51*well depth) - (2.86*distance)
n_h <- -1.83+ (6.51*1) - (2.86*.42)
f_h <- -1.83+ (6.51*1) - (2.86*2.12)
n_h - f_h # = 4.862
```

```
> 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 6kg?

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

Figure 2:

Question 3

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.

```
#gerneral for a ram <- -18.1386 + (2.2980*weight) - (4.0716*group.dummy.2=1)
Ram <- -18.1386 + (2.2980*weight) - (4.0716*1)
```

(b) ewe is 0 is neither Ram or wether so if boths dummy variable is set to 0 all theats left is ewe

```
#Ewe lamb general -18.1386 + (2.2980*weight) -(8.3622*group.dummy.1) - (4.0716*group.dummy.2)
-18.1386 + (2.2980*6) -(8.3622*0) - (4.0716*0)
#4.3506
```

(c) For all weights I set to 0 to remove bring the varibale as low as possible and comapare.

```
#C
#Ewe
ewe <- -18.1386 + (2.2980*0)-(8.3622*0) - (4.0716*0)
ewe
Ram <- -18.1386 + (2.2980*0)-(8.3622*0) - (4.0716*1)
Ram
wether <- -18.1386 + (2.2980*0)-(8.3622*1) - (4.0716*0)
wether
#ewe has the highest fat index for every weight
```

Question 4

Suppose we are interested in studying whether the alignment of foreign policy goals between countries impacts the delivery of international disaster assistance. Figure 1 plots the total amount of money an individual country donated or pledged to another country to aid in the recovery of a natural disaster (the y-axis is in millions of $) by the level of foreign policy agreement between the two countires (0-100).

What concerns might we have about using the level of foreign policy agreement 'as is' in a model that regresses 'amount of disaster relief provided' on 'foreign policy agreement'? How could we address these concerns?

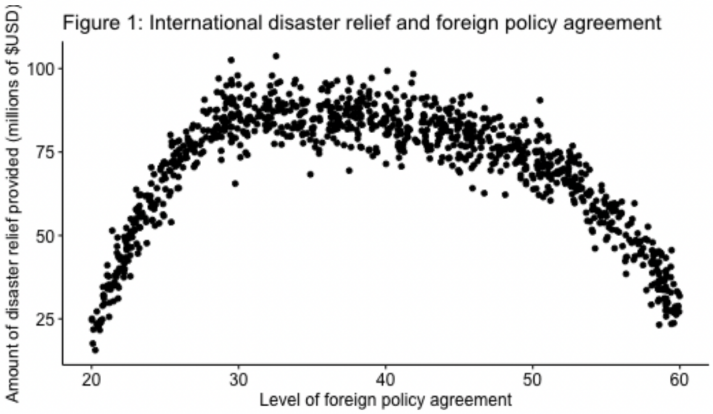

Figure 3:

(a) A concern about ths model is the shape of the relationship which is in this case not linear. To resolve this you would square route the variable and the graph would become linear. This

Question 5

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

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-ANSWER

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-ANSWER

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-ANSWER
2. False

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-ANSWER
3. Interactive (salary = age * education) regression model
4. Interactive (education = age * salary) regression model

Question 6

Question 6

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 3: Estimated coefficients from regression predicting variation in FDI.

	Estimate	Std. Error
(Intercept)	-61.03	43.45
GDP	-3	0.00021
Democracy	7.609	7.285
Education	4.433	3.561

a) Interpret the coefficients for GDP and Democracy.

b) The author claims that she 'cannot reject the null hypothesis that Education has no effect on FDI ($H_0 : \beta_{Education} = 0$) . Using the coefficient estimate and the standard error for Education construct a 95% confidence interval for the effect of Education 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 25491.1 as the mean of GDP and use +/- one standard deviation around the mean of Education (from 11.06 to 13.08) for low and high values of Education respectively.

Figure 4:

(a) Interpret the coefficients for GDP and Democracy.
GDP is related to a decrease in FDI(-3 coeffecient)
Democracy is related to an increase in FDI(7.609)

(b) The author claims that she 'cannot reject the null hypothesis that Education has no effect on FDI (H0 : BEducation = 0) . Using the coefficient estimate and the standard error for Education construct a 95F DI. Based on the confidence interval, do you agree with the author? Explain your answer.
Given how wide the range is,it is unlikely in my view that the education has no effect on FDI so I disagree with the Author and would rejec tthe Null hypothesis educations is a non zero effect on FDI

```
#B
4.433 + (1.96*3.561)
4.433 - (1.96*3.561)
##range is -2.54656 to 11.41256
```

(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 25491.1 as the mean of GDP and use +/- one standard deviation around the mean of Education (from 11.06 to 13.08) for low and high values of Education respectively.

```
general -61.03 - (3*gdp) + (7.609*deomcracy) + (4.433*education)
l_ed <- -61.03 - (3*25491.1) + (7.609*0) + (4.433*11.06)
h_ed <- -61.03 - (3*25491.1) + (7.609*0) + (4.433*13.08)
h_ed - l_ed
#8.95466
```