Applied Stats I: Exam 2

Lily Rice 16304845

Due: December 9, 2022

Instructions

- Please read carefully: You have from 09:00 Wednesday December 7 until 08:59Friday December 9 to complete the exam. Please export your answers as a singlePDF file and include all code you produce in a supporting R file, which you willupload to Blackboard. The exam is open book; you can consult any materials youlike. You must not collborate with or seek help from other students. In caseof questions or technical difficulties, you can contact Professor Ziegler via email. Youshould write-up your answers in R and LaTeX as you would for a problem set. Pleasemake sure to concisely number your answers so that they can be matched with thecorresponding questions.

Question 1

Suppose we are interested in studying whether the alignment of foreign policy goals betweencountries impacts the delivery of international disaster assistance. Figure 1 plots the totalamount of money an individual country donated or pledged to another country to aid in therecovery of a natural disaster (the y-axis is in millions of $) by the level of foreign policyagreement between the two countires (0-100).What concerns might we have about using the level of foreign policy agreement 'as is' in amodel that regresses 'amount of disaster relief provided' on 'foreign policy agreement'? Howcould we address these concerns?

Answer: First: The scatterplot shows that while both variables do seem to be related with each other, this relationship is clearly not linear. Specifically, the scatterplot of the amount of disaster relief provided against level of foreign policy agreement follows an upside-down "u" shape. This indicates that the relationship between these two variables is non-linear not monotonic (does not go in the same direction at all points). If we were to use the level of foreign policy agreement "as is" in a linear model that regresses the amount of disaster relief provided on foreign policy agreement, this model's predictive power would not be as good as it could be, as it would be treating a non-linear relationship as linear. Specifically, this model would likely underestimate the amount of disaster relief provided for levels of foreign policy agreement under 40, and overpredict the amount of relief for policy agreement scores over 40.

We could address these concerns by transforming the x values(foreign policy agreement scores). Specifically, we could potentially cut the observations into two sections - those with x<35 or so and those with x> 35 or so. Then, we could take transform the x values of the observations in the first cohort (with x<35) by taking the square root of the x values, as this would likely make these points more linear when plotted (because the initial graph has these points, when plotted, bulging in the upper left hand side). Similarly, for the second cohort of observations, we could square all of the x values (because in this section of the initial graph the plotted points are bulging toward the upper right corner, so increasing x by squaring (or maybe even cubing) could transform this section into a more linear plotted relationship). The next step would be to plot these transformed points to first see visually if these transformations actually managed to make the plotted relationship more linear. Then, we could fit linear models that utilised the transformed x variables, and because these models would better fit this data they would more accurately predict y values (amounts of disaster relief provided) for any given x values (levels of foreign policy agreement).

Please note that this is just one suggestion of how to transform the data at hand to make a linear model better fit. There are other options to achieve a

better fitting model. Y-values could also be transformed to potentially improve our model fit. Running transformations of variables where we try out cubing versus squaring them would also be beneficial to see what works better. Additionally, we could use least-squares regression to identify any skewness in this data, and if we find evidence of this could engage in further transformations (or have even more of a reason to stick with the power transformations we have already engaged in) in order to account for this skewness by making the distribution more symmetric. By simply looking at the initial graph however, it does not look drastically skewed in one direction or another. In conclusion, the main obvious concern with this data that I have discussed solutions for here is its non-linearity.

Question 2

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.

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 wether lamb using the estimated coefficients.

The fitted model for a wether lamb using the estimated coefficients is as follows:
Y_wether = -18.1368 + (2.2980)*(Weight) - (8.3622)*(1)

(b) What is the predicted Fatness index of a ram lamb that weighs 6kg?

I found this by plugging in a weight of 6kg to the fitted model for a ram lamb:

Y_ram = -18.1368 + (2.2980)*(6) - (4.0716)*(1)

Y_ram = -8.4204

So the predicted fatness index of a ram lamb that weighs 6kg is -8.4204.

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

The ewe lamb group will have the highest Fatness index for every weight. This is because when a lamb is the same weight in each group, the prediction equation for the ewe group will output the highest predicted Fatness index compared to the equations for wether and ram groups at that same weight.

Question 3

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) Constituent term

The constituent terms are the explanatory variables included in the model when they are listed individually (not multiplied with each other as part of the interaction effect). Constituent terms are especially relevant when it comes to interaction effects modeled with linear regression. For example, a linear model could be: lm(y tilda X + D + X:D). The constituent terms in this example are "X" and "D", while the interactive term is X:D. Differentiating between the constituent and interactive terms and having both in a model like this is important, because they have different potential effects on the outcome. The constituent terms must be included in this model, because they display the "main effect" that these terms have on the outcome variable y. The interactive term must also be included, as this lets us see whether there is an interaction between these two variables which would cause a different effect on the outcome. If there is not a significant interaction effect between X and D, we can return to an additive model using just the constituent terms (measuring just the constituent/"main" effects of both). If an interaction effect is suspected, running regression models initially that include both the CONSTITUENT terms plus the interactive term is necessary in order to account for the full range of effects that may be produced by the variables plus their interaction with each other. When we run regressions on multiple variables, we are finding the partial effects of each input variable on the outcome variable. These partial effects are the constituent effects, or the effect that the constituent terms have on the outcome.

(b) Test Statistic

A test statistic is used when conducting a hypothesis test, and describes how far a point estimate falls from what we would expect if the null hypothesis was true. A test statistic measures the number of standard errors that the point estimate is from the value expected if the null was true. This test statistic can be either a z-score or t-score (depending on what kind of test is being done (on proportions (z) or means (t, unless large sample size) for example). We can calculate our p-value from a test-statistic (by looking at a z- or t- table), and then compare this p-value to our significance level (denoted by alpha). If the p-value (calculated from our test statistic value) is smaller than our significance level, this provides us enough evidence to reject our null hypothesis. Therefore, the test statistic is an essential step

in the Hypothesis testing process, because it allows us to measure in a standardised way how far away from an expected value our sample statistic is, which we then use to make inferences about a population using these sample statistics. Also, in regression analysis, we use test statistics when conducting hypothesis tests in the form of t-tests, F-tests, or partial F-tests to tell us whether or not we should include a particular predictor variable (or group of predictor variables) in our model. Thus, test statistics here are also useful when identifying what variables we need to include in our model and model provides the best fit to our data/has the most predictive power.

(c) Partial F-test

Partial F-tests are a type of hypothesis test in which you can check if a subset of the predictor variables included in a linear regression model are important / influential on the outcome variable, thus should be included in the model together. This type of test is conducted following main steps of a hyothesis test. First, null and alternative hypotheses are stated. The null hypothesis is that the slopes of all variables in the subset add up to 0 (which would indicate there is no linear relationship between the outcome and this subset of predictors), and the alternative is that at least one of subsetted variables' slopes =/= 0. Next, the partial F-statistic is calculated. This can be done a few ways (e.g. with an ANOVA test or using the correlation coefficient for the subset sample). Then, the p-value is found, and if this p-value is small enough we can reject the null hypothesis for this partial F-test. Rejection of the null hypothesis implies that the subset of predictor values tested do in fact have impact on the outcome variable, and would be useful to include in the linear model. Partial-F tests are therefore a type of hypothesis test that can be used in part to determine what variables should be included in regression models to make them more accurate in predicting outcomes. This test is useful in helping us to use models that are the best fit for our data, while still are the most parsimonious (thus are not needlessly adding variables causing increases in error that are unnecessary).

(d) Residuals

Residuals are a measure of "error" between a prediction line and the observed points in a dataset. When a regression line is put through a scatterplot of observations, the residual for each observation is the distance between it and the regression line. We want our regression lines / prediction equations to minimise error, or to be a "best fit line" to this data, thus we want the sum of the distances between all observations and the line to be smallest (the sum of the absolute values of all residuals to be as small as possible).

When trying to identify what prediction equation / linear regression model reduces this error the most we often transform residuals. This is because residuals can be both positive or negative, so a line that does not actually match the data very well, but is drawn with the same number of residuals above it as are below it, could have a sum of residuals that equals out to 0 (indicating it is the "best fit" when it is actually not). To account for this, we transform residual values in a number of ways (for instance, by looking at the sum of squared residuals to determine what prediction equation best minimises error / produces a line of best fit) that overcomes the problem that arises when dealing with residuals in their initial form.

Thus, residuals (especially when transformed) are useful in determining what prediction equations / linear regression models best fit the data at hand. Futher transformations of residuals like the residual standard error are often used to assess how accurate a regression model is; and transformations like studentised residuals can be used in identifying outliers in data that may have influence on regression outcomes.

Question 4

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

   Answer = 4. QQ-plot of residuals

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

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

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

   Answer = True

4. The coefficients in an ordinary least squares regression model "fill in the blank" .

   Answer = minimizes the residual sum of squares

Question 5

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 processedour data as well as run our regression (N= 1000), and we get the following output. Pleaseconsult the table below, which presents the estimated coeffients and standard errors from ourmodel, to answer the following questions. Also, note that the economic variables (GDP percapita 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 dummyvariable (1=Democracy, 0=Non-democracy).

(a) Interpret the coefficients for GDP and Democracy.

Interpreting the coefficient for GDP = Holding education and democracy variables constant, a one unit increase in GDP leads to a decrease of 2 in FDI on average.

Interpreting coefficient for Democracy = Holding education and GDP variables constant, a country changing from a non-democracy (0) to democracy (1) leads to an on average increase in FDI of 4.389.

(b) Based on the confidence interval, do you agree with the author? Explain your answer.

First I constructed a 95 percent confidence interval for the effect of Demcoracy on FDI. I did this with the following equation: 95 per cent CI = DemocracyBeta (+ and -) t value * DemocracySE.
I calculated my t value, and then calculated my lower and upper levels of the 95 percent confidence interval with the following code:

```
t_value <- abs(qt((1-.975), 1000-2))
CI_lower <- 4.389 - (t_value*.4)
CI_higher <- 4.389 + (t_value*.4)
```

Which resulted in a 95 percent confidence interval for the effect of Democracy on FDI of (3.604, 5.174). The null hypothesis value of the Democracy coefficient was 0. Because this 95 percent confidence interval does not contain this null hypothesis value, this is evidence to reject the null hypothesis that Democracy has no effect on FDI.
I then calculated my Test Statistic and p-value with the following code:

```
TS_dem <- 4.389/0.4
p_value_dem <- 2*pt(abs(TS_dem), 1000-3, lower.tail = F)
```

Which resulted in a p-value that was very small (1.6 to the e-26 power). This is further indication that the results of this hypothesis test are significant, and I have grounds to reject the null hypothesis.

In conclusion, I disagree with the author's claim that she cannot reject the null hypothesis. Instead, I reject the null hypothesis that Democracy has no effect on FDI, because the null hypothesis's coefficient value for Democracy of 0 does not fall within the 95 percent confidence interval I have constructed, and further, the chance of getting the Democracy coefficient estimate of 4.389 if the null hypothesis was true is extremely low (as illustrated by the very small p-value), which is yet another reason to reject the null hypothesis. Instead, we have evidence that Democracy has an effect on FDI.

(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 23936.45 asthe mean of GDP and use +/- one standard deviation around the mean of Education(from 10.99 to 12.89) for low and high values of Education respectively.

The difference in predicted FDI between "low" and "high" values of Education for non-democratic countries holding GDP constant at its sample mean is 9.2701 million dollars. This means that holding GDP per capita constant (at 23936.45 dollars/year), a non-democracy going from an average education rate of 10.99 years in school to 12.89 years in school, results in an on average increase in FDI of 9.2701 million dollars.

This difference was calculated by plugging in these values and relevant estimated coefficients into the prediction equation for this model. The prediction equation is: predicted FDI = (intercept) - 2*(GDP) + 4.389*(Democracy) - 4.879*(Education).
The predicted FDI at the "low" education level was calculated with the following equation:

```
FDI_Low_education <- -1.426 - (2*23936.45) - (4.879*10.99)
```

which resulted in the predicted FDI at the low education rate = -47927.94621.
The predicted FDI at the "high" education level was calculated with the following equation:

```
FDI_High_education <- -1.426 - (2*23936.45) - (4.879*12.89)
```

Which resulted in the predicted FDI at the high education rate = -47837.21631.

The difference in predicted FDI between low and high education rates in this case was then calculated by subtracting FDI_High_education from FDI_Low_education, which equals 9.2701.

Question 6

Many of the wells used for drinking water in Bangladesh and other South Asian countries arecontaminated with natural arsenic, affecting an estimated 100 million people. Arsenic is acumulative poison, and exposure increases the risk of cancer and other diseases, with risksestimated to be proportional to exposure.We performed a regression analysis with the data to understand the factors that predictthe arsenic level of 1000 households' drinking water. Your outcome variable arsenic is acontinuous 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 tothe nearest factory as the two predictors of a household's arsenic level. The estimatedcoefficients are found in the first column of the table above. Interpret the estimatedcoefficients for the intercept and each predictor.

Interpreting Model 1's "intercept" coefficient of 0.42: This intercept of 0.42 in Model 1 implies that if the distance of a household to the closest known commercial factory was 0 kilometers/100, and the respondent's (household's) well was "not deep", the household's arsenic level is predicted to be 0.42 hundred micrograms per litre on average.

Interpreting Model 1's "well_depth" coefficient of 2.49: This coefficient of 2.49 implies that holding the distance of the household to the closest known commercial factory constant, a household's well depth changing from "not deep" (0) to "deep" (1) will lead to an average increase in that household's arsenic level of 2.49 hundred micrograms per litre.

Interpreting Model 1's "dist100" coefficient of -3.99: This coefficient of -3.99 implies that holding a household's well depth constant, a one kilometer/100 increase in their distance away from the closest known commercial factory (moving 1 kilometer/100 further away fom a factory) will lead to an on average decrease in household drinking water arsenic level of 3.99 hundred micrograms per litre.

(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 coefficientsin part (a) to conform with the interactive model in column 2 of the table above. Whatis the appropriate test to determine whether we should model the relationship betweendistance, well depth, and arsenic levels using an additive or interactive model? Whatinformation would you need to perform that test?

The coefficient estimate for the closest known factory does not seem to vary statistically significantly based on whether or not a house has a steep well. While model 2 indicates (with the interaction effect coefficient) that the impact that distance to a factory has on arsenic levels is 0.06 times less when they have a deep well, versus if they had a non-deep well; this interaction effect coefficient is not cited as statistically significant.

The appropriate test to determine if we should model the relationship between distance, well-depth, and arsenic levels using an additive or interactive model would be a partial F-test. This test compares the sum of squared errors in the "reduced" model (e.g. Model 1) and the "full" model (e.g. Model 2), to determine if the full model is actually a better fit to the data. To complete a partial-F test I would need to state my null and alternative hypotheses, would need the Residual Sum of Squares for each model, and the mean of squared errors of the "full" model (aka the interaction model). If I had access to this dataset, I could complete this partial F- test with the ANOVA command in R, which would produce an F-statistic and p-value. I would then use this p-value to either reject or fail to reject the null hypothesis (that there is no significant difference in the SSE of the Full and Reduced models). I would decide to go with the interactive model (Model 2) if I rejected my null hypothesis; and would decide to stay with the additive model if I failed to reject the null hypothesis (if the p-value was too large to reject).

(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 = 0.4) than the other (dist100 = 2.06).

I am sticking with the additive model because the predictive power of Model 1 is the same as Model 2 (R squared is .54 for both). I will therefore compute the average difference in arsenic levels asked for in this section using the additive model as my "preferred model".

The prediction equation for this additive model is:(arsenic) = (0.42) + 2.49*(well_depth) - (3.99)*(dist100)

Then, I found the predicted arsenic levels for the two households, one with dist100 = 0.4, and the other with dist100 = 2.06. This was calculated in R with the following code:

```
arsenic_1 <- (0.42) + (2.49*1) - (3.99*0.4)
arsenic_2 <- (0.42) + (2.49*1) - (3.99*2.06)

diff_in_arsenic <- arsenic_2 - arsenic_1
```

diff_in_arsenic = -6.6234

The average difference in arsenic levels between two households with deep wells, but one is closer to a factory (dist100 = .4) than the other (dist100 = 2.06) is 6.6234. This means that based on our additive model, in this example the house with a deep well that was farther away from the factory (2.06 km/100 away) had on average 6.6234 hundred micrograms per litre less arsenic in their drinking water than the house with the same "deep" well-depth ("1"), but located a closer distance to the factory (0.4 km/100 away).