Stats 141XP Discussion 1: HW 1 (Charles’ Portion)

**Question one. Relevant articles can be found in ethics folder on week one.**

a) Read the article entitled: “Data Science and Covid-19- Harvard Science initiative”.

Summarize the major points you learned from this article as a data scientist.

b) Read the article entitled: “Evidence Vs. Truth” in the corona virus epidemic by Andrew

Gelman. Summarize the major points you learned from this article as a data scientist.

Based on the knowledge you gained from the above two articles, write a case explaining…

1) A study involving research on COVID-19,

2) The problem that the data scientist working on the project is supposed to solve,

3) One or more ethical dilemmas that the data scientist is facing, and

4) The strategy/strategies that you think the data scientist should follow to solve the

ethical dilemma or dilemmas involved.

**Answer One:**

1. Data science is important for understanding Covid-19 through methods such as development and testing of the Covid-19 vaccine, factors that’ll slow the infection rate, understanding how airborne transmission works, reopening schools and when, finding socioeconomic and environmental factors, and tracking mobility to contain the virus. Covid-19 has been filled with lots of uncertainties making it difficult to decide when to reopen, but it was easily called when to close and commence shut down. Data scientists choose to acknowledge the uncertainties and embrace them as compared to leaders, who may choose to reject uncertainty. Data scientists use a range of numbers to determine their findings (such as “Confidence Intervals”), and a true statistician shouldn’t just trust a single number but a range of numbers. Sometimes, there are problems that not even data scientists can solve instantly because data scientists do not have an emergency response team. It can be difficult task for data scientists to have to collect the data and then analyze it. The collection process of data is difficult in itself because there are many details about the data that is left out, and they’ll receive only the results to analyze. Dr. Meng tried to emphasis the point that data science is meant to be targeted for everyone to know and learn about to some degree. Due to Covid-19 pandemic, data science is currently focusing on Covid-19 and racial discrimination data and analysis. It is also important for data scientists to remain ethical when analyzing the data.
2. One problem that was being dealt with during the early stages of the Covid-19 pandemic is that people knew the number of confirmed Covid-19 cases were lower than the number of exposed cases, but they were unsure how much lower. This problem was further exacerbated due to the fact that early stages of Covid-19 testing were slower. A Stanford study showed that 50-85-fold the amount of people were confirmed Covid-19 cases in Santa Clara County. This event had people questioning the study. The rate went from 1.5% to 2.5% confirmed rate. This rate actually came from a sample size rather than the actual population. Sadly, the adjustments made do not account for several factors such as age, ZIP code difficulty, and others. After a while, the Stanford study was shown to have incorrect Confidence Intervals, and it was not the best choice to use classical statistical methods for so many uncertainties. As a statistician, they are meant to focus on the data at hand and whatever assumptions are being made on the data. This is called a “statistical error” if a quantitative claim is not supported by the data and assumptions made. This is not a matter of ethics because statisticians have a chance of testing errors in their research. Once they find a mistake/error, data scientists should do their best to correct their errors, acknowledge their errors, and find out what they did wrong. It is considered an ethical violation if you do not acknowledge someone pointing out a possible error in your research, even if they are an outsider. Data science is difficult to do, and it can be hard to distinguish truth vs. evidence. Overall, as long as data scientists are able to acknowledge their errors and correct their mistakes, data scientists are always able to continue working and improving their research.
3. One study on Covid-19 would be to track which states have the most infections (or deaths by Covid-19) in the U.S., along with their political affiliation with their respective state. This might help statisticians understand if a certain political state and their policies on Covid-19 response might be a possible causation in increased infections.
4. One possible problem they may face is that some states refuse to give accurate data on their Covid-19 infection and death rates. Without an accurate data for statisticians to study, we will be going back to the article “Evidence vs. Truth”, where we question if the actual data is correct or the statisticians study. Another problem is to determine what states are politically affiliated with prior and during the pandemic.
5. One such ethical dilemma that statisticians may face is that the state might attempt to have the statistician altercate the data/study to make it seem like their state’s infection/death rates are not that high. As stated in “Evidence vs. Truth” article, statisticians could also be incorrect in their study of the data and the analysis that follows.
6. The first dilemma is quite simple. Statisticians should NOT altercate the data/study in any way. If the state asks such a request, they’ll either refuse the request or continue to publish the study and analysis no matter how good or bad it is reflected on the state. As for the second dilemma, if a statistician’s work is deemed incorrect in some way by an outsider, they should take the time to reevaluate and ensure that their findings are accurate. They might ask other statisticians to do the study to ensure that multiple statisticians’ findings are accurate (or even relatively close) to the findings of the original. If they are incorrect, the statistician should immediately remedy their mistake by stating they made a mistake and should recall their findings.

**Question Two**

Given the following information, answer questions a to d.

SIMS (second international mathematics study) data set was used to carry out the analyses given

below.

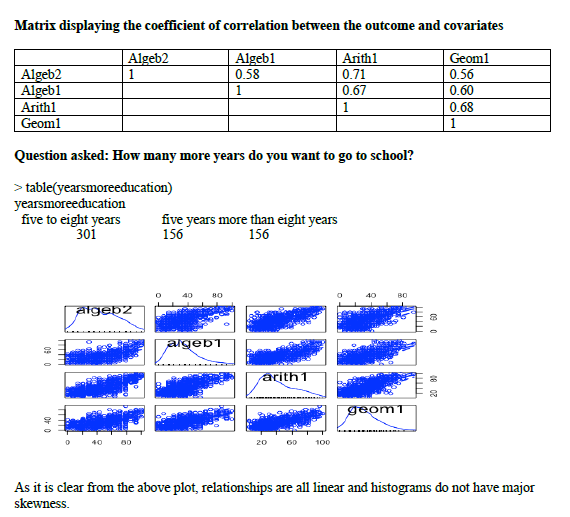
Outcome = algeb2 (scroe on algebra after 8th grade)

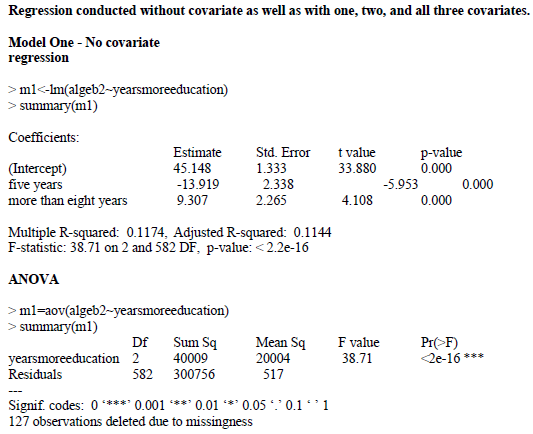
Covariates = algeb1, arith1, geom1 (pretest on algebra, arithmetic, and geometry)

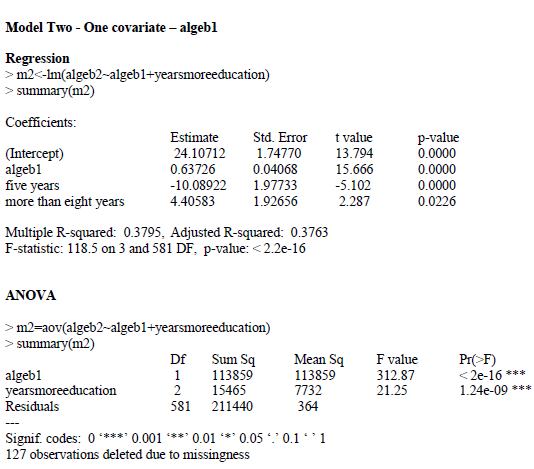
Predictor we are interested in = How many more years of education the student plans to have

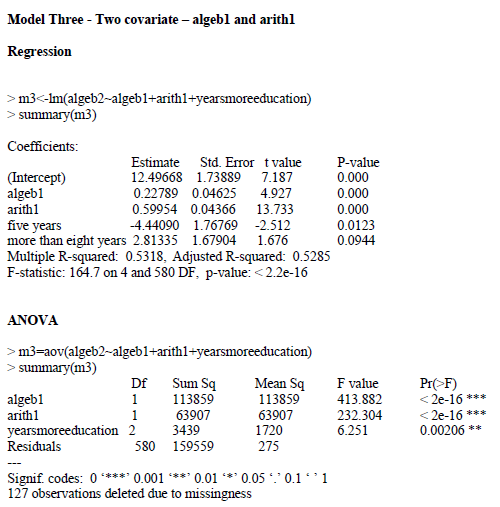
after high school (two years, five years, more than eight years).

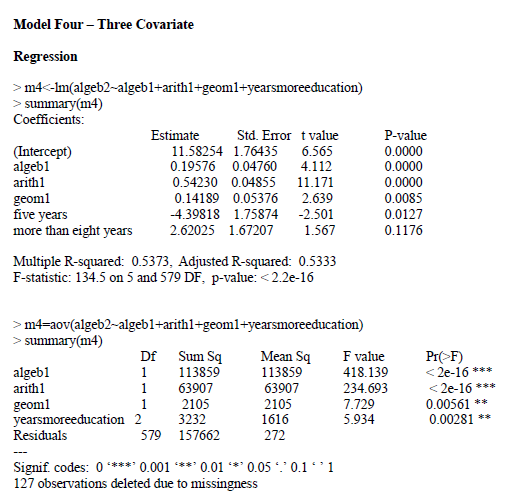
Given the analysis below, answer the following questions:











**Using the above information…**

a) Complete the following table.

b) How would you explain the concept of covariate conceptually?

c) Elaborate how the covariates of algeb1, arith1, and geom1 affect the coefficients of years more education.

d) Elaborate how the covariates of algeb1, arith1, and geom1 affect the practical

significance of years more education.

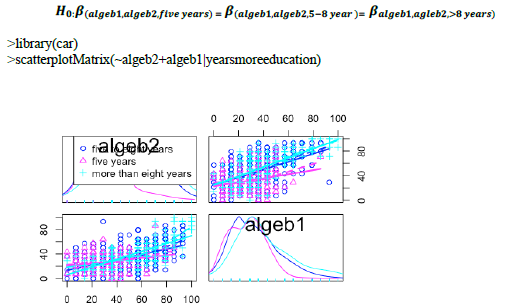
e) Compare SS TOTAL for different models and explain what it shows mathematically and conceptually.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients | Model 1 | Model 2 | Model 3 | Model 4 |
| Intercept | 45.148 | 24.10712 | 12.49668 | 11.58254 |
| Algeb1 | N/A | 0.63726 | 0.22789 | 0.19576 |
| Arith1 | N/A | N/A | 0.59954 | 0.54230 |
| Geom1 | N/A | N/A | N/A | 0.14189 |
| 5 Years vs. 2 Years | -13.919 | -10.08922 | -4.44090 | -4.39818 |
| > 8 Years vs. 2 Years | 9.307 | 4.40583 | 2.81335 | 2.62025 |
| SST | 340765 | 340764 | 340764 | 340765 |
| SS due to Covariates (SSR) | 300756 | 211440 | 159559 | 157662 |
| SS due to Years More than Covariates (SSE) | 40009 | 129324 | 181205 | 183103 |
| R-Squared due to Covariates | 0.1174 | 0.3795 | 0.5318 | 0.5373 |
| R-Squared due to Years More than Covariates | 0.1144 | 0.3763 | 0.5285 | 0.5333 |

f) If you were the TA for this class, how would you use the following plot to explain the

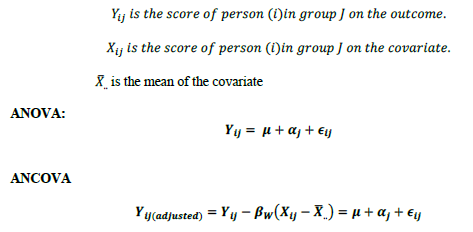
following assumption to your students? To answer this question, refer to the chapter on

ANCOVA posted in the ANCOVA folder.



g) Using the following linear models elaborate the major difference between one way

ANOVA and one-way ANCOVA with fixed effects.



**Answer Two**

1. “completed in the table above”
2. The covariates are continuous predictor variables that are complementary to the response variable. They are related in some way to the dependent variable. As for what it is, the covariate is a possible explanatory variable of the dependent variable. The covariates help with reducing error variance and reducing bias.
3. We can see that as we add more covariates into our model, the coefficient for “5 years vs. 2 years” increases and the coefficient for “> 8 years vs. 2 years” decreases. The intercept coefficient (“5 to 8 years”) decreases as more covariates are added. The reason the coefficients decrease is because as we add more covariates to the model, this will reduce the variance of the residuals.
4. As we increase the number of covariates we use, the statistical significance of “years more education” decreases. The p-value of “yearsmoreeducation” can be seen increasing, but it isn’t enough to change the fact that it is still a statistically significant variable to our model.
5. Mathematically, the SST is pretty much the same. Model 1 and Model 4 have an SST of 340765, and Model 2 and Model 3 have an SST of 340764. The SST is the dispersion of observed dependent variables around the mean. Conceptually, this tells us the total amount of variation in our model. We can see that there is slightly more variation for Model 1 and Model 4.
6. One assumption that is being made is that the adjustment of the within-groups Sum of Squares is that the within-groups regression coefficients are all estimates of the same common population regression coefficient. The size of the error variance is determined by our dispersion of the “Conditional Distributions”. The conditionals being “algeb1”, “algeb2”, and the different type of “yearsmoreeducation”. The higher they are correlated with each other, the narrower our scattered points become, and the higher our reduction in the error variance becomes due to ANCOVA. **[unsure if the points needs to be more scattered for good or less scattered for good (?)].**
7. The main difference between one-way ANOVA and one-way ANCOVA is that the Yij(adjusted) takes into account the variation of dependent scores that is *not* associated with the linear regression of Y on X. The adjustment made is made to remove the linear effect on the covariate(s). This will reduce the Sum of Squares for Y if the slope of the regression does not equal zero. The covariates and their differences affect the observed difference between Yij and Yij(adjusted). The terms from Yij can be rearranged to get the “adjusted score”. The thing about the “adjusted score” is that it is free from the effects of the covariates, and it also provides an estimate of the ANOVA terms of the model equation for a completely randomized design. One difference is that the ANCOVA error term will be usually smaller than the error term in ANOVA.