Katherine Grantham

NRE 538 Final Exam\_Takehome

1. **Is there a significant association between gender (gender) and whether people think it’s rude to bring an unruly child on the plane (unruly\_child)? If yes, which gender tends to think that bringing an unruly child is more rude? Flying**
   1. For this test, we have two categorical variables, so we have to run an Chi Squared test to look for an association. For the Chi Squared test, we need to make sure our assumptions are met before we run the test. For the Chi Squared test, we assume that we have a random sample, and we assume that our observations are independent. We also have more than 5 expected and observed variables, therefore, we are able to run a Chi Squared test.
   2. First, we can create a contingency table, then we can run the chi squared test in R. After running the chi squared test, we see that our results are significant, as can be seen by the low p-value (0.001193). This p-value is well below our set alpha value (0.05), and so we can trust that our results are fairly accurate and we run a low risk of committing a type 1 error. This means that there is a significant association between gender and whether people think it’s rude to bring an unruly child on the plane. When viewing the contingency table, we are able to see how the results differ, based on men and women and their opinion of unruly children on planes. From this, we see that Men have a stronger association with thinking that bringing an unruly child on a plane is more rude.
2. **Is there a significant difference in tuition (tuition) by type of institution (type)? If yes, which type has a higher tuition? College**
   1. We would now like to see if there is a significant difference in tuition by type of institution. There are two types of institutions: Private nonprofit and Public. Then we have tuition as our one quantitative variable that has two different groups. Therefore, we need to run a Two-Sampled t-test.
   2. Now that we know what test to run, we need to check to see if our assumptions are met. We assume that our sample is random and that are observations are independent. We can also assume our sample size is small enough, meaning less than 10% of the population, which it is based on our samples for each.

****

* 1. Next, we have to test to see that our values are nearly normal. To do this, we can run a qqplot and a shapiro test. When we do this, we see from the graph that our values are not nearly normal. Especially after running a shapiro test, we see that we receive significant p-values (1.147\*10^-6 and 2.58\*10^-10). This means that are values are not normal, which is not what we want. However, our sample size is greater than 30, and so we can assume normality.
  2. Lastly, we need to test for equal variances. After running var.test, we see that we do not have equal variances. This is because we have a significant p-value of 2.2\*10^-16. Therefore, we can’t have equal variances and must run a Welsh’s t-test, which can be run when there are unequal variances.
  3. After running a Welsh’s t-test, we get a p-value of 2.2\*10^-16. This means our results are significant and there is a difference by type of institution. At the bottom of our R output, we see the means of each type of institution when comparing tuition. Private nonprofit institutions have a mean tuition of 28,301 and public institutions have a mean tuition of 18,885. Therefore, Private non-profit tuitions have a higher tuition rate.

**Welch Two Sample t-test**

**data: College.Priv$tuition and College.Public$tuition**

**t = 22.79, df = 1397.9, p-value < 2.2e-16**

**alternative hypothesis: true difference in means is not equal to 0**

**95 percent confidence interval:**

**8632.749 10258.878**

**sample estimates:**

**mean of x mean of y**

**28301.69 18855.88**

1. **Is there a significant difference in happiness (Hscore) by region (Region)? Happy**
   1. For this test, we have one quantitative variable (happiness Hscore) and multiple groups (regions). Because of this, we have to run an ANOVA (one-way ANOVA).
   2. First we must test the assumptions before we can run the ANOVA. We assume that the data was randomly sampled and that the observations are independent. We also assume the samples to be small enough, so smaller than 10% of the population.
   3. Next, we run the Shapiro-Wilks test to check if the data is normally distributed. The p-value was significant after running this test (0.012), meaning we do not have normal data. However, the sample size is large enough, so we can assume the variables are normal or nearly normal.
   4. Next, we run a Levene’s test to see if there is homogeneity of variances. When running this, we get a non-significant p-value (0.517), so we can assume equal variances. Lastly, we do not have the same sample size per treatment group. Therefore, we run a Kruskal-Wallis test.
   5. After running the ANOVA with the Kruskal-Wallace test, we get a significant p-value of 3.84\*10^-11. From this we can determine that there is a significant difference in Hscore or happiness based on the region.
2. **What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model. Happy**
   1. When running the pairs function and correlation function to visualize the data. When we do this, we see that corruption was somewhat correlated with Hscore (0.4) and Freedom (0.5). Because these variables are correlated, and we want to avoid multicollinearity, they can’t be used in the model. There are variables that are not highly correlated with the model, such as GDP, life, and generosity, and so they can be used in our model.
   2. First, we need to test the assumptions of our linear model. When first running all of the assumptions of the model, using the dwtest for independency, the bptest for homoscedasticity, and the shapiro-wilks test for residual normality, we do not meet the assumption of normality. When running the shapiro-wilks test, we get a p-value of 1.057\*10^-5. This is not what we want, as we want to fail to reject our null hypothesis when running this test. When the assumption of non-normal errors is not met, we must log transform the data. To do this, we square the y value of corruption. After running this, we can test our model again.
   3. After log transforming our data by squaring the y value of corruption, we are able to get accurate results and meet the assumptions. For this model (listed as model 2 in the R document), I run the Durbin-Watson test and receive a p-value of 0.07. This is no significant, which is what we want to see, meaning that we met independency. Next, we run the BP test to test for homoscedasticity. After running this test, we receive a p-value of 0.147. This means we have met the assumption of homoscedastic errors. Lastly, we run the shapiro-wilks test to check for normality, and we receive a significant p-value of 0.198, meaning our errors are not normal. Therefore, we have met all the assumptions of our model. We can now run the model and interpret the results.
   4. We see from the results that generosity, family, and Hscore are all significantly associated with corruption. Each explains significant variance in the model. We see that as generosity increases by a factor of one, corruption increases by 0.23, and it is significant because it has a p-value of 0.0018. Next, we see that as family increases by a factor of 1, corruption decreases by 0.106. This is significant because it has a p-value of 0.05. However, this is right at our alpha point, so as a scientist it may or may not be included. Lastly, we see that Hscore is significantly associated with corruption, because for every increase by a factor of 1 in Hscore, there is an increase in corruption of 0.05, and this has a p-value of 9.78\*10^-6. When we evaluate our model, however, we have an R-Squared valued of 0.213. This is not a very high model association, and so there may be other variables that better fit the model or it’s possible there is another model that would be better to use for this data. There is a significant p-value of 5.145\*10^-8, and so that is good for our model and it’s overall significance, allowing us to reject the null hypothesis.
3. **Choose one of the continuous independent variables that was significant in the model for Question 4 and interact it with region (Region) to predict corruption (Corruption). This model should only include one continuous independent variable and its interaction with region. Does the influence of your continuous variable on corruption vary by region? If yes, how do you interpret the interaction? Happy**
   1. After running the last model, we see that Generosity is associated with corruption, in that it explains some variation in the model, as shown by its significant p-value. Therefore, we can use it for our new model. For this model, we have categorical and continuous variables, so we have to run an ANCOVA.
   2. When running the ANCOVA, we see that we the interaction between region and Generosity can be used to predict corruption values. We see that for generosity, when Region is silenced, every one increase in GDP leads to a 0.188 increase in corruption values. This is has a significant p-value. However, when generosity is silenced and we are looking at regions impact of corruption, there is not a significant interaction. When we interact the terms generosity and Region, we see that see that there is a not a significant interaction. This is due to the fact that we do not have a significant p-value for this interaction, and so the coefficents are not dependent on one another. Therefore, we fail to reject our null hypothesis and we can be fairly certain in doing this, in that there is not a large risk for committing error.
4. **Which factors are significantly associated with whether a breast cancer tumor is malignant or not? Choose three continuous independent variables to include in your model. Cancer**
   1. To look at the correlation between variables, we can run the cor() function. Whether a tumor is malignant or not is correlated with the radius mean (R=0.73). It is also correlated with perimeter mean (R=0.74). We also see that it is correlated with area mean (0.70).
   2. For our model, we do not want to run into issues of multicolinearity, so we will pick variables for our model that are not highly correlated with one another. For the model, we will look at texture mean, area mean, and smoothness mean.
   3. We first attempt to run our model as a multiple linear regression. However, when running tests for independency, normality, and homoscedasticity, we are not able to meet our assumptions. However, there is a model we can run that relaxes these assumptions. We can run a generalized linear model that will allows us to look at whether a breast cancer tumor malignancy is impacted by the radius, area, and perimeter of the tumor. Through this, we can compare the log odds of malignancy and how it is impacted by these variables. After we transform our data by running a glm, we see that both area, texture, and smoothness all are important in explaining whether a breast cancer tumor is malignant or not. This is due to the fact that the p-values of are significant. Especially in regardness to smoothness, where the log odds of a tumor being malignant increases by 146 when smoothness is factored in (p-value=2.25\*10^-14).
   4. To see how good our model is, we can also do a model comparison. For this, I ran texture, radius, and area together. In this model, we get a higher AIC value (311), and se we may assume that our first model is actually good at predicting our dependent variable with the independent variables we included (AIC=190).
5. **BONUS/EXTRA CREDIT: Which independent variables are the most important in explaining whether a breast cancer tumor is malignant or not? Use the same 3 continuous independent variables you chose for question 6. cancer.**
   1. When looking at this GLM model, we see that some variables are important in explaining whether or not a breast cancer tumor is malignant. When looking at the independent variables, we see that smoothness mean from the model plays a large role. It is a significant p-value of 2.25\*10^-14 and the log odds of a tumor being malignant increases by 146 when smoothness is at it’s mean and when a cancer tumor is smooth. We also see that texture plays a role. This is because the p-value is 3.92\*10^-11, which his well below our alpha, so we can reject our null hypothesis. So for texture, the log odds of a tumor being malignant or not increases by 0.3 when texture is at it’s mean. Lastly, area is significant, having a p-value of 2.1\*10^-16, and the log odds of a tumor being malignant increases by 0.016 when area is at it’s mean. This would all be important to know in treating cancer and understanding the probability of a tumor being malignant or not.