**Melissa Selva**

This exam is open book and open internet but you are NOT allowed to work with anyone else or ask anyone other than Meha or Oscar any questions about the exam. It is due at noon on Sunday, April 23.

Please answer the following questions by analyzing the associated datasets. For all tests, please:

* check whether the data meet the requirements/assumptions of the test you plan to run
* complete any transforms needed to make the data meet the required assumptions
* run the test
* interpret the results (do not include only the R output)
* check model fit in the case of linear regressions and/or glms
* if you have the option between running a linear model with a transformed y variable or a glm, choose the linear model with a transformed y variable. only run a glm when you have to.

Provide all answers in R or R markdown (similar to the take home quiz 4). Use the following scripts to load the datasets. The dataset to be used for each question is provided in bold at the end of the question.

Dataset Please use the following scripts to load in the data from GitHub

flying = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/flying.csv",header=TRUE, sep=",")

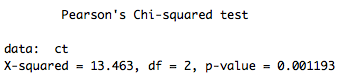
college = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/college.csv",header=TRUE, sep=",")

happy = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/happy.csv",header=TRUE, sep=",")

cancer = read.table(file="https://raw.githubusercontent.com/OscarFHC/NRE538\_2017Fall/master/Final/cancer.csv",header=TRUE, sep=",")

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**?

I’m choosing to run a chi-squared test because both the independent variable and dependent variable are categorical. All assumptions for the chi-squared test (random sample, independence of observations, sample is small enough, and enough observed and expected values) we can assume were met. Null hypothesis would be that gender and whether people think it’s rude to bring an unruly child on a plane are independent, and alternative hypothesis would be that variables are not independent.

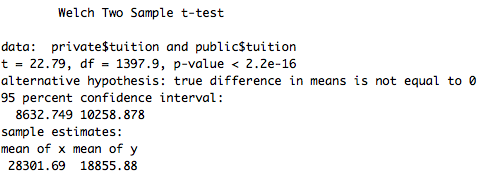


Yes, from this output of the chi-squared test, it is clear that gender and whether people think it’s rude to bring an unruly child on a plane are not independent. The p-value of 0.00119 is highly significant, meaning that we can reject the null hypothesis, and that there is a relationship between the two variables. From this contingency table that I created before running the test (see below), it seems that males are more likely to report bringing an unruly child on a plane as “VERY rude”.



1. **Is there a significant difference in tuition (tuition) by type of institution (type)? If yes, which type has a higher tuition?**

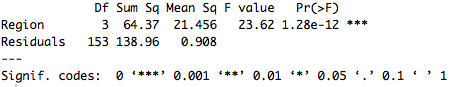
I’m going to use a two-sample t-test to answer this question. The type of institution is our categorical independent variable (with only two groups) and tuition is our continuous dependent variable. The assumptions to be tested are random sample, independent observations, sample size less than 10% of population, normal values, and equal variances. We can assume the first three, and the data failed both normality and equality of variances. However, a histogram produced looked relatively normal, and the sample size is very large so we can assume normality by CLT. I will also use a Welch’s t-test instead to deal with the unequal variances.



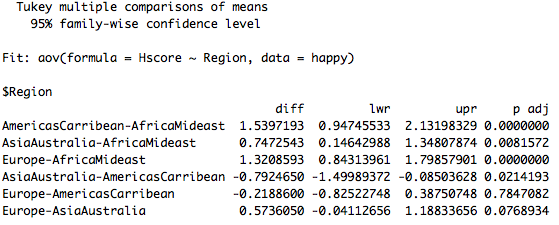
As we can see, the p-value is highly significant, meaning that there is a difference in tuition by type of institution. The mean of x, which is private nonprofit institution, has higher tuition.

1. **Is there a significant difference in happiness (Hscore) by region (Region)?**

I will run a one-way ANOVA for this question because we have an independent categorical variable with more than two “levels” and a continuous dependent variable. The assumptions to be tested are random sample, independent observations, homogeneity of variances, normal response variable, and equal sample size. We can assume a random sample, independent observations, and equal sample sizes, but need to test the other two. With Levene’s test, the data was shown to have homogeneity of variances across groups. The data failed normality by shapiro.wilk test, but with the large enough sample sizes, we can assume normality via CLT.



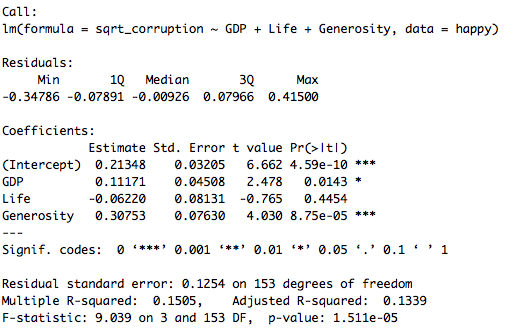
As we can see from this summary of the ANOVA, we can see there are significant differences in happiness by region. However, we need to run a Tukey HSD test to determine which regions are significantly different.



From the Tukey HSD test, it seems that there are significant differences between many of the regions except Europe - AmericasCarribean and Europe –AsiaAustralia. The largest differences in happiness are between AmericasCarribean-AfricaMideast and Europe-AfricaMideast.

1. **What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model.**

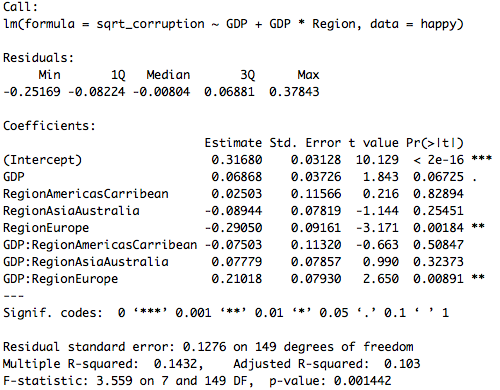
I plan to do a multiple linear regression to answer this question. I’m going to include GDP, Life Expectancy, and Generosity in my model. For a linear model, the assumptions that must be checked are independent observations, independent errors, a linear relationship between x and y, normality of errors, and homoscedasticity of errors. We can assume independent observations. For a linear relationship, we can plot the residuals of the model, which show distinctive pattern. The independence of errors and homoscedasticity of errors were both confirmed using the dwtest and bptest, respectively. The errors were not normal to begin with, so I square-root transformed the dependent variable and it then passed the shapiro.wilk test. I also checked to be sure none of the independent variables were correlated with one another, so I checked the VIFs of the model and none of them were over 5.



The y-intercept of the model is 0.213, and the slopes for GDP, Life, and Generosity are 0.111, -0.062, and 0.307, respectively. Both GDP and Generosity are significant, which means that they have a significant effect on the variability in the dependent variable, Corruption. For every increase of 1 in GDP, Corruption goes up by 0.1117 and for every increase of 1 in Generosity, Corruption goes up by 0.307. However, the adjusted R2 is only 0.1339, which means that only 13.39% of the variability in Corruption can be explained by this model, so the model fit might not be great.

1. **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?

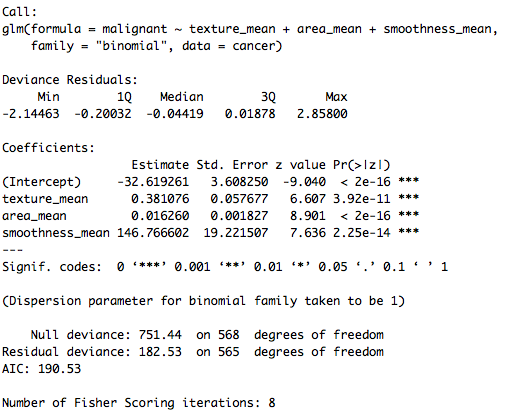
I’m choosing to run another multiple linear regression, with the one continuous independent variable and the interaction term between GDP and Region. The assumptions remain the same as with question 4. The new model passes all of the assumptions as before: homoscedasticity (bptest), independent errors (dwtest), and normal errors (shapiro.wilk test).



With this model, GDP alone becomes non-significant, and the only interaction that is significant is between GDP and Europe, as seen by the p-value of 0.00891. The slope of the effect of GDP on corruption is different for Europe than for the other regions. The slope (of the effect of GDP on Corruption) for Europe is 0.06868 minus 0.21018. The adjusted R2 of this model is pretty low at 0.103, which means that only 10.3% of the variability in Corruption can be explained by GDP and the interaction between GDP and Region.

1. **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.**

I’m choosing to use a GLM to answer this question because the dependent variable only has a yes or a no answer for whether a tumor is malignant. This easily falls into the binomial family, for which we can use a GLM to transform. The assumptions for a GLM are as follows: 1) independent observations, 2) dependent variable is from a broader exponential family, and 3) a linear relationship between x and transformed y variable. We can assume independent observations, the dependent variable falls within the binomial family, and a linear relationship between the x variable and transformed y variable.



As we can see, all of the independent variables have a significant effect on the log odds ratio of tumor malignancy. I back transformed my regression coefficients to remove the log for easier interpretation. For every increase of 1 in mean texture, the odds of the tumor being malignant increase by 1.46. For every increase of 1 in mean area, the odds of the tumor being malignant increase by 1.016. For every increase of 1 in mean smoothness, the odds of the tumor being malignant increase by 5.494e63. Because GLMs do not provide R2, I decided to compare the AIC values of other models. I ran an alternative model with mean radius of the tumor instead of mean smoothness. The AIC of the second model was 296.81, while the AIC value of my model was 190.53, suggesting that my model is a better fit.

1. 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.**