Alex Truelove

NRE 538

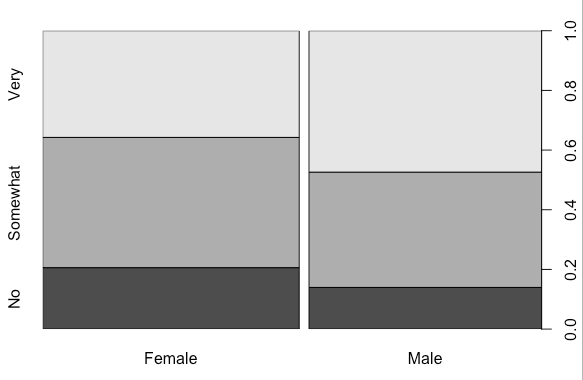
**Take Home final**

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**

**Both variables are categorical (gender = female/male, unruly child = no/somewhat/very), so we use a Chi-square test to answer this question. The assumptions are met: 1) observations are independent and 2) no “structural zeros.”**

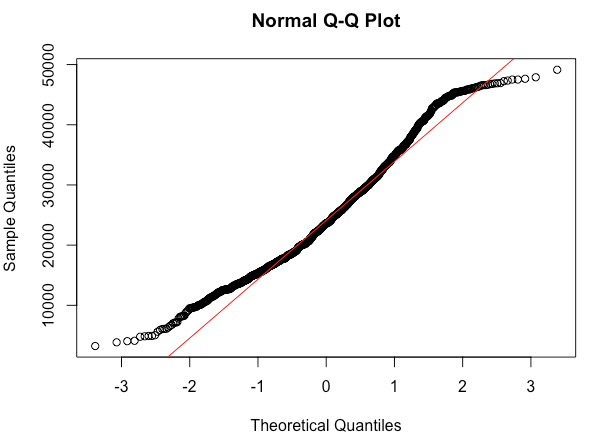
* **Pearson’s chi-square test result: p-value = 0.0009995**

**So, yes there is a significant association between gender and whether people think it’s rude to bring an unruly child on the plane. To find out which gender, we can plot the results in a table (shown below). Clearly, more men think it’s very rude (and fewer men think it’s not rude) to bring an unruly child onto an airplane.**

****

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

**Type of institution is categorical (public/private) and independent, while tuition is continuous, so we use an unpaired t-test to compare the difference in tuition between public and private universities. First, we check normality:**

****

**The qqplot (and histogram, in R script) show that tuition is somewhat normally distributed, enough that no transformation is necessary. Regarding equal variance, R defaults to a Welch’s t-test, so unequal variance is accounted for.**

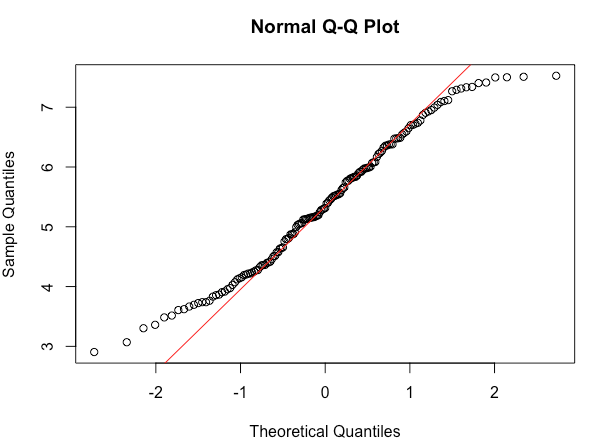
* **Welch’s t-test results: p-value < 2.2e-16, so the difference is very significant**

**We can see from the output that private nonprofit universities have higher tuition, as one might expect:**

* **Mean public university tuition: $18855.88**
* **Mean private nonprofit university tuition: $ 28301.69**

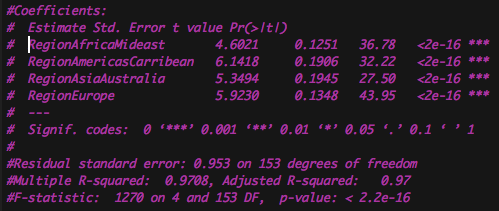
1. Is there a significant difference in happiness (Hscore) by region (Region)?

**Region is categorical (public/private) while Hscore is continuous, so we use an ANOVA to compare the difference in tuition between regions. First, we check assumptions. The samples are independent. The normality:**

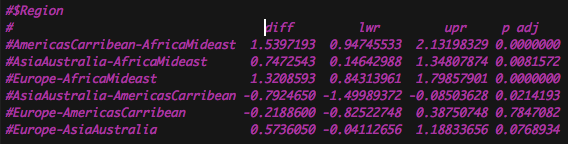


**The qqplot show that tuition is Hscore somewhat normally distributed, and the histogram confirms that shape, enough that no transformation is necessary. The variance is close enough to equal that the outcome should not be affected too greatly.**

**The ANOVA results are as follows:**



**p-value < 2.2e-16, so there is a significant difference. However, we need a post hoc test to tell us which are significantly different, so we use the TukeyHSD:**

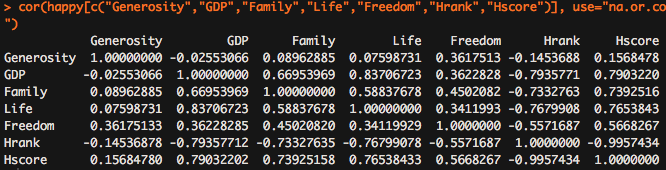
****

**Happiness is significantly different between 4 of 6 possible pairwise comparisons. Only Europe-AmericasCarribean and Europe-AsiaAustralia are not significant at the 0.05 level.**

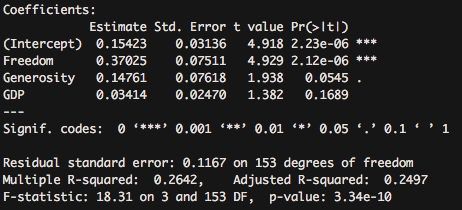
1. What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model.

**Firstly, the variable “Corruption” must be checked for linear regression assumptions. The qqplot and histogram reveal that Corruption is not normally distributed at all and is in fact very right-tailed. A square root transformation appears to make the data (now “sqrtCorruption”) somewhat normally distributed.**

**When individually regressed against “sqrtCorruption,” variables “Generosity", "GDP", "Family", "Life", "Freedom", "Hrank", "Hscore" were all significant at the 0.05 level. However, some of those variables were highly correlated with each other (GDP**~**Life, Hscore**~**Hrank) and thus should not be included together in the model. Correlation results:**

****

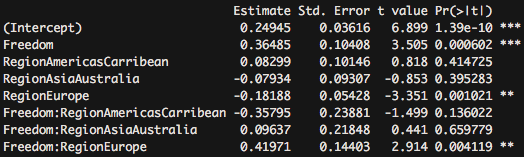
**For the model, I chose three significant variables which are relatively uncorrelated: Generosity, GDP, and Freedom. In this model, all three are positively correlated with corruption score (it’s unclear whether a higher score means more or less corruption). Freedom accounts for the largest coefficient and highest level of significance. Regression results:**



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?

**Variable to interact with Region = Freedom.**

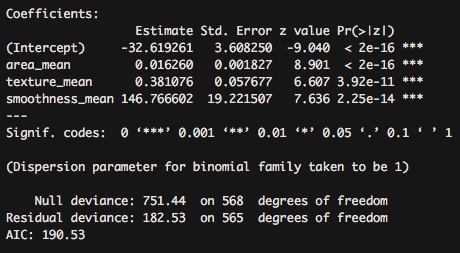
**The** interplot **function reveals some strong interactions. The output describes them numerically:**



**AfricaMidEast is the omitted/unlabeled Region, see label for interaction “Freedom.” It has the most significant interaction (p = 0.0006), meaning that the AfricaMidEast region strongly and positively interacts with Freedom. This makes sense, given that countries in Africa and the Middle East appear to have lower Freedom scores. The interaction between Europe and Freedom is also significant (0.001), however it’s negative.**

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.

**The variable “malignant” is categorical, which created a non-normal (binomial) distribution, so we use a GLM. Because independent variables “radius”, “perimeter”, and “area” are strongly correlated, I’ve chosen to include are, texture, and smoothness in the model, since they are not strongly correlated with one another. The results:**



**All three variables are statistically significant and positively correlated with likelihood of cancer being malignant.**

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.

**Between area, texture, and smoothness, the coefficient for smoothness is extremely high, though the units to measure smoothness are unclear. If the units are similar somehow, than the explanatory power of smoothness is very high.**