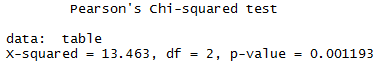
1. Is there a significant association between gender and whether people think it’s rude to bring an unruly child on the plane? If yes, which gender tends to think that bringing an unruly child is ruder?

* H0: There is no association between gender and people’s perception of bringing unruly child on the plane.
* Check Assumptions (for chi-square test as both variables are categorical)
* Independent observations—True
* No structural zeros (all cells have potential to be filled)—True
* All the cells have enough data—True
* Transformation: No
* Test and interpretation

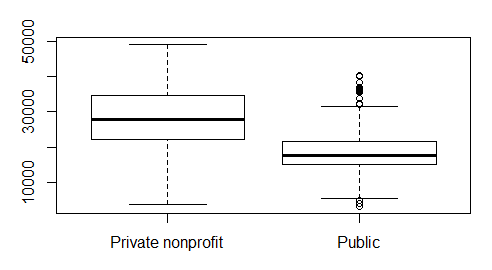


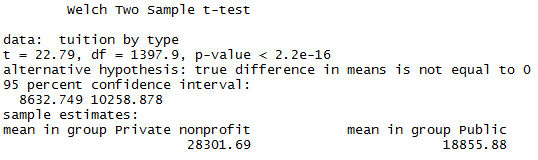


p-value<0.05. Therefore, we reject H0. There is significant association between gender and how rude people think it is to bring unruly child onto plane. From the table, it seems that males think bringing unruly child is ruder. (I have also tried assigning scores to people’s attitude and perform a t-test, and the result is the same, but I didn’t include it in the code as I’m a bit doubtful if the score can be treated as continuous)

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

* H0: There is no significant difference in tuition by type of institution.
* Check Assumption (for two-sample t-test as there are two types of institution)
* Independent observation—True
* The dependent variable is normally distributed—failed shapiro test but the histogram and qqplot looks OK, and the sample size is greater than 30, so I think the data is good enough for the test without transformation
* Equal variance—No. Fialed var.test (p-value<0.05), but I can use Welch’s test
* Transformation: No
* Test and interpretation

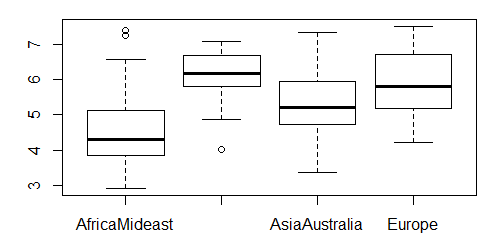




p-value<0.05. Therefore, we reject H0. There is significant difference in tuition between public institution and private non-profit institution. Private non-profit institution has higher tuition.

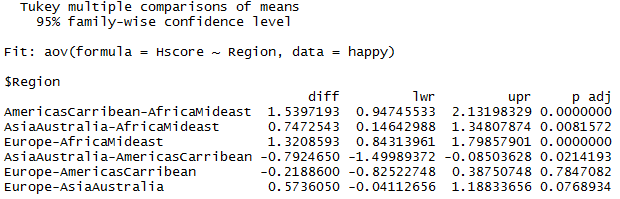
1. Is there a significant difference in happiness by region?

* H0: There is no significant difference in happiness by region.
* Check Assumptions
* Independent sample—True
* Normal Distribution—failed shapiro test (p-value<0.05) but the qqplot and histogram looks OK. Also, the sample size is greater than 30 with more than 15 samples in each group, so the data should be good for the test without transformation
* Equal variance—True (Confirmed by LeveneTest)
* Transformation: No
* Test and interpretation





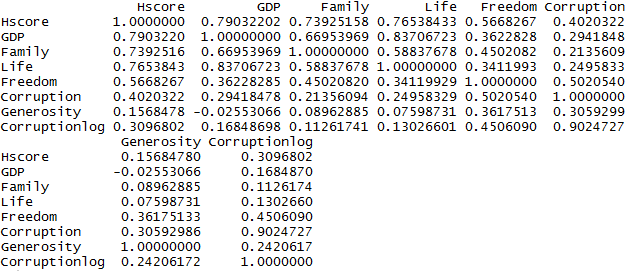
p-value<0.05. Therefore, we reject H0. There is significant difference in happiness by region.



Happiness: AmericasCarribean>Europe>AsiaAustralia>AfricaMideast. There is no significant difference in Hscore between AmericasCarribean and Europe. AmericasCarribean and Europe has significantly higher Hscore than the other two, and AsiaAustralia has significant higher Hscore than AfricaMideast.

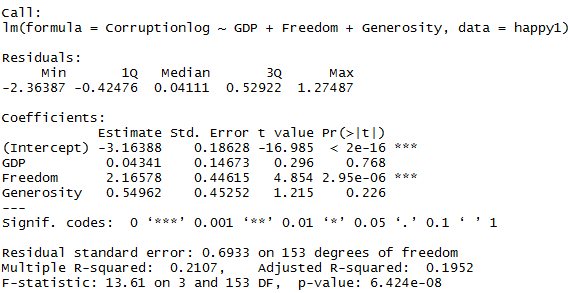
1. What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model.

* Correlation coefficient table to choose independent variables



Family and Life are removed as the correlation coefficients are smaller, and they seem highly related to GDP. Hscore is also highly related to factors GDP, family, freedom, and life, so it is also removed. The independent variable to use is GDP, Freedom, and Generosity

* Transformation: Log transformation. I plotted the histogram and qqplot of the original data, which seems a bit too skewed. After running the model, I found the residuals near-normal (qqplot) but heteroscedastic. Then I tried log transformation. The log-transformed Corruption is nearly normal. However, when running the transformed model, the GDP and Generosity are no longer significantly related to Corruption and the adjusted R2 was reduced to around 0.2. But as transformation solved the problem of heteroscedasticity, it might give me a less biased significance level, so in the end I settled for the transformation
* Model and interpretation

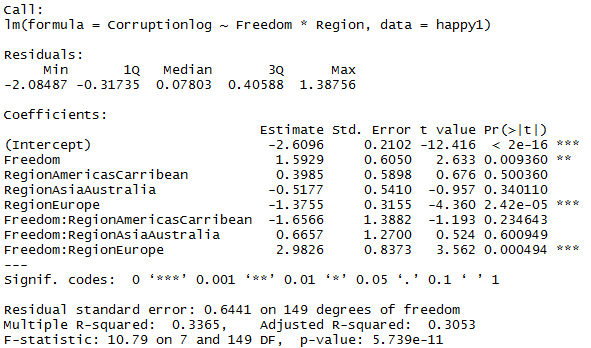


* The intercept is significantly different from 0.
* Freedom is significantly related to the log of corruption. With each unit of increase in freedom, the log of corruption increase by 2.16 units. There is no significant relationship between GDP/ Generosity to the log of Corruption. (The relationship was significant in untransformed model, but that might be biased)
* Assumption check
* Linear relationship between dependent and independent variables—assume so but cannot plot
* Homoscedastic errors—True (plot and bptest)
* Independent errors—True (plot and dwtest)
* Normally distributed errors—True. Although failed shapiro test and histogram is slightly skewed to the right, the qqplot looks OK
* Model fit

The adjusted R2 is about 0.2, meaning that the model only explains 20% of the variance in the dependent variable. But that’s best fit model with three independent variables (other factors are even less related to corruption, or might run into multicollinearity) There are probably other unincluded factors to account for corruption.

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 used log-transformed Corruption in this model as well due to the reasons stated in the previous question.
* Model and interpretation

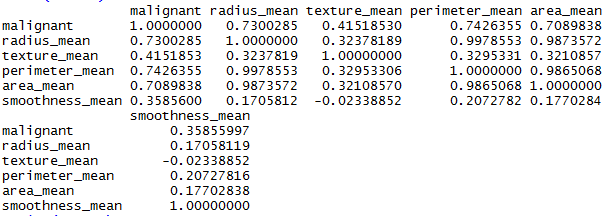


* The intercept (log of corruption in AfricaMideast) is significantly different from 0.
* Freedom is significantly related to the log of corruption in AfricaMidwest. For each unit of increase in freedom, the log of corruption in that region increases by 1.6 unit.
* The log of corruption in Europe is significantly lower than AfricaMideast. The log of corruption in the other two regions are not significantly different from AfricaMideast.
* There is significant interaction between freedom and region for Europe. In Europe, for each unit of increase in freedom, the log of corruption increase 2.98 unit more than in AfricaMideast (about 4.56 unit in total; the slope for Europe is 2.98 higher than the slope of AfricaMideast). There is no significant interaction between freedom and region in AmericasCarribean and AsiaAsutralia. (Their slope is the same with that of AfricaMideast. For each unit of increase in freedom, the log of corruption increases by 1.6 unit in those regions)
* Assumption check
* Homoscedastic errors—True (plot and bptest)
* Independent errors—True (plot and dwtest)
* Normally distributed errors—True. Although failed shapiro test and histogram is slightly skewed to the right, the qqplot looks OK
* Model fit

The adjusted R2 is about 0.3, meaning the model explains 30% of the total variance in the dependent variable, and seems to be an OK fit.

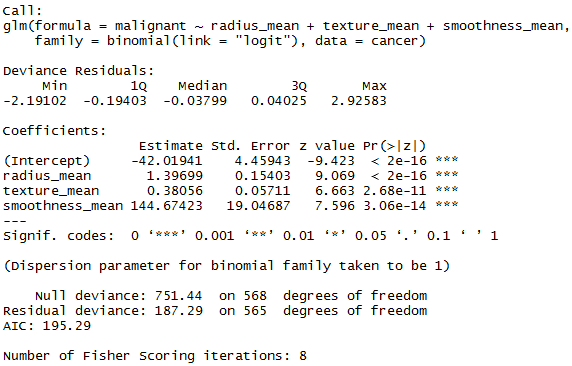
1. Which factors are significantly associated with whether a breast cancer tumor is malignant or not? Choose three continuous independent variables to include in your model.

* Whether the tumor is malignant or not is binominal distribution. We need to run a GLM
* Correlation Coefficient table to select independent variables



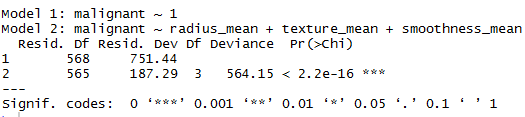
Radius\_mean, perimeter\_mean, and area\_mean are highly related to each other. Therefore, we include radius\_mean, texture\_mean and smoothness\_mean in our model

* Model and interpretation



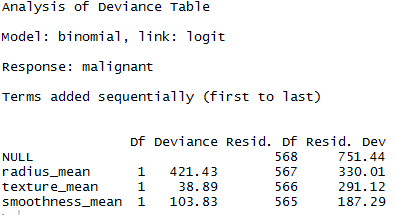
Radius\_mean (and presumably the perimeter and area calculated from it…), texture\_mean and smoothness\_mean are significantly related to the log of odds ratio of whether a tumor is malignant or not.

* When all the independent variables are 0, the odds ratio (p/(1-p)) and probability of a tumor being malignant is close to 0.
* With each unit of increase in radius\_mean, the log of odds ratio of the tumor being malignant increases by 1.39 unit (new odds ratio = 4.01\*previous one)
* With each unit of increase in texture\_mean, the log of odds ratio of the tumor being malignant increases by 0.38 unit (new odds ratio = 1.46\*previous one)
* With each unit of increase in smoothness\_mean, the log of odds ratio of the tumor being malignant increases by 144 unit. (new odds ratio = 3.4\*10^62\*previous one) \*
* \*Considering the scale of smoothness\_mean, an 0.01 unit of increase would probably make more sense. I tried using smoothness\_mean\*100 as an independent variable so one unit increase is 0.01 unit in smoothness\_mean. It did not alter the other coefficients or the significance level. With each 0.01 unit of increase in smoothness\_mean, the log of odds ratio of the tumor being malignant increases by 1.44 unit. (new odds ratio = 4.22\*previous one)
* Assumption Check
* Linear relationship between the transformed dependent variable and independent variable—assume so, unable to plot
* Independent errors—No. Failed dwtest but I did the GLM anyway…
* Model Fit



The model has significantly less deviance than the dummy model, so it should be a good 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**.**



Radius\_means took up most of the deviance, so it’s probably the most important explanatory factor. (Also, if we look at pairs(cancer), growth in radius\_mean results in the greatest separation in whether the tumor is malignant or not)