NRE538 Take Home Final

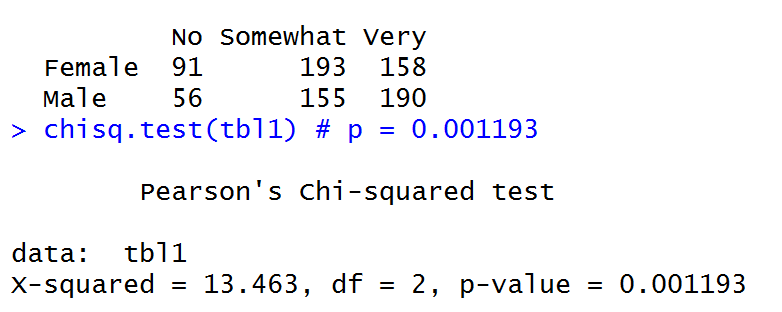
Anne Gidley (adirkes)

Question 1

Two categorical variables: Use Chi Square

Check assumptions:

* Random sample: would need to understand how data was collected
* Independent observations: would need to understand how data was collected
* Sample size small enough: < 10% of the population surveyed
* Enough expected and observed values: no cells with value <5



p = 0.001193 : there **is** a significant association between gender and rudeness rating.

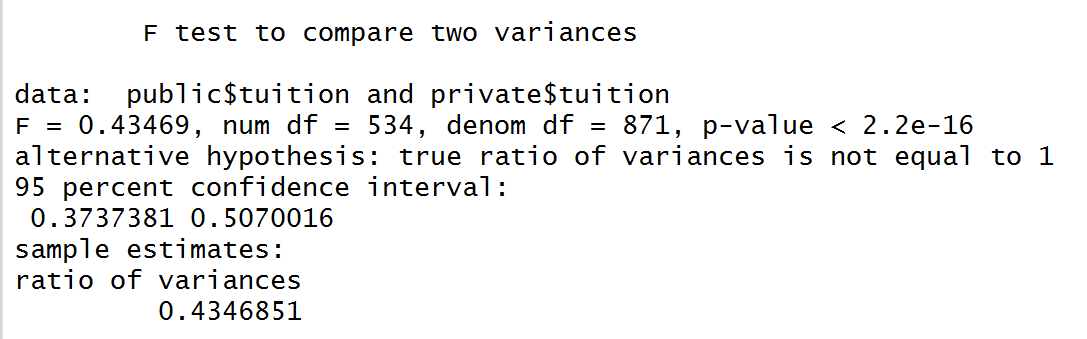
Men tend to think it is more rude to have an unruly child on a flight.

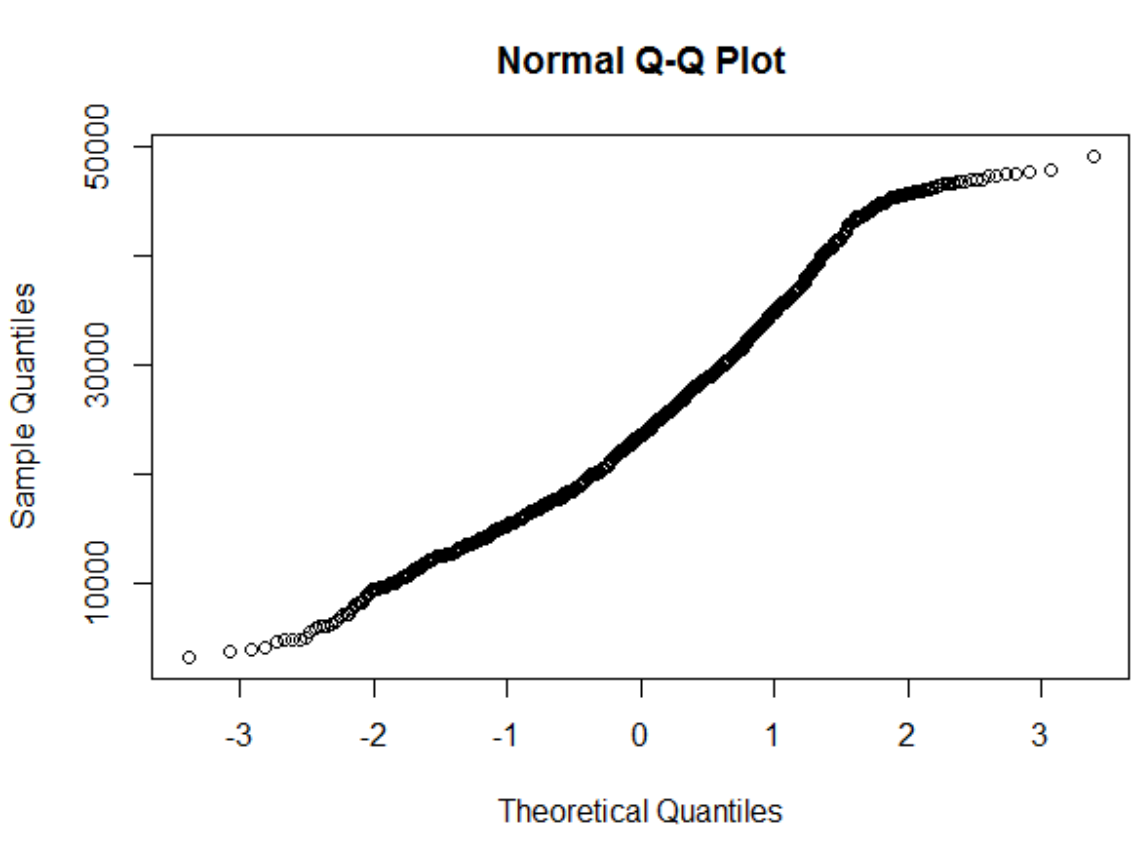
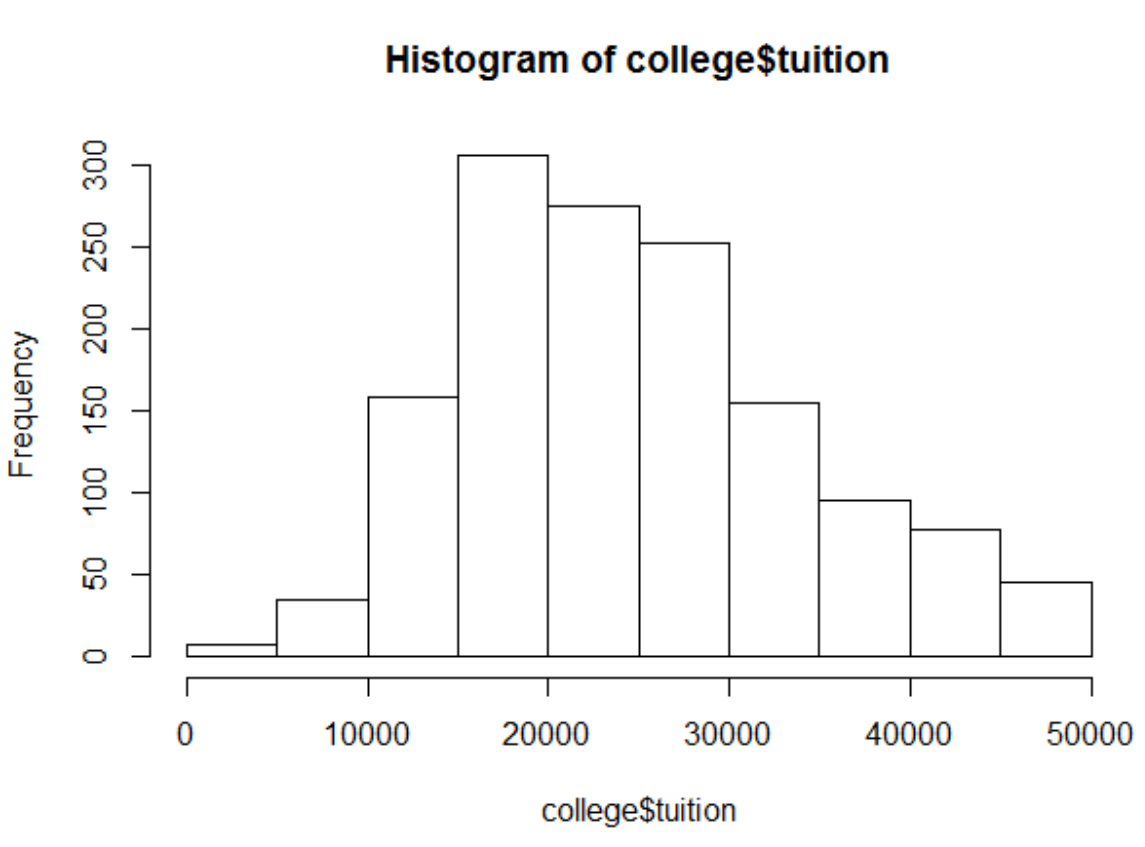
Question 2

One categorical (type) and one continuous (tuition): two-sample t-test

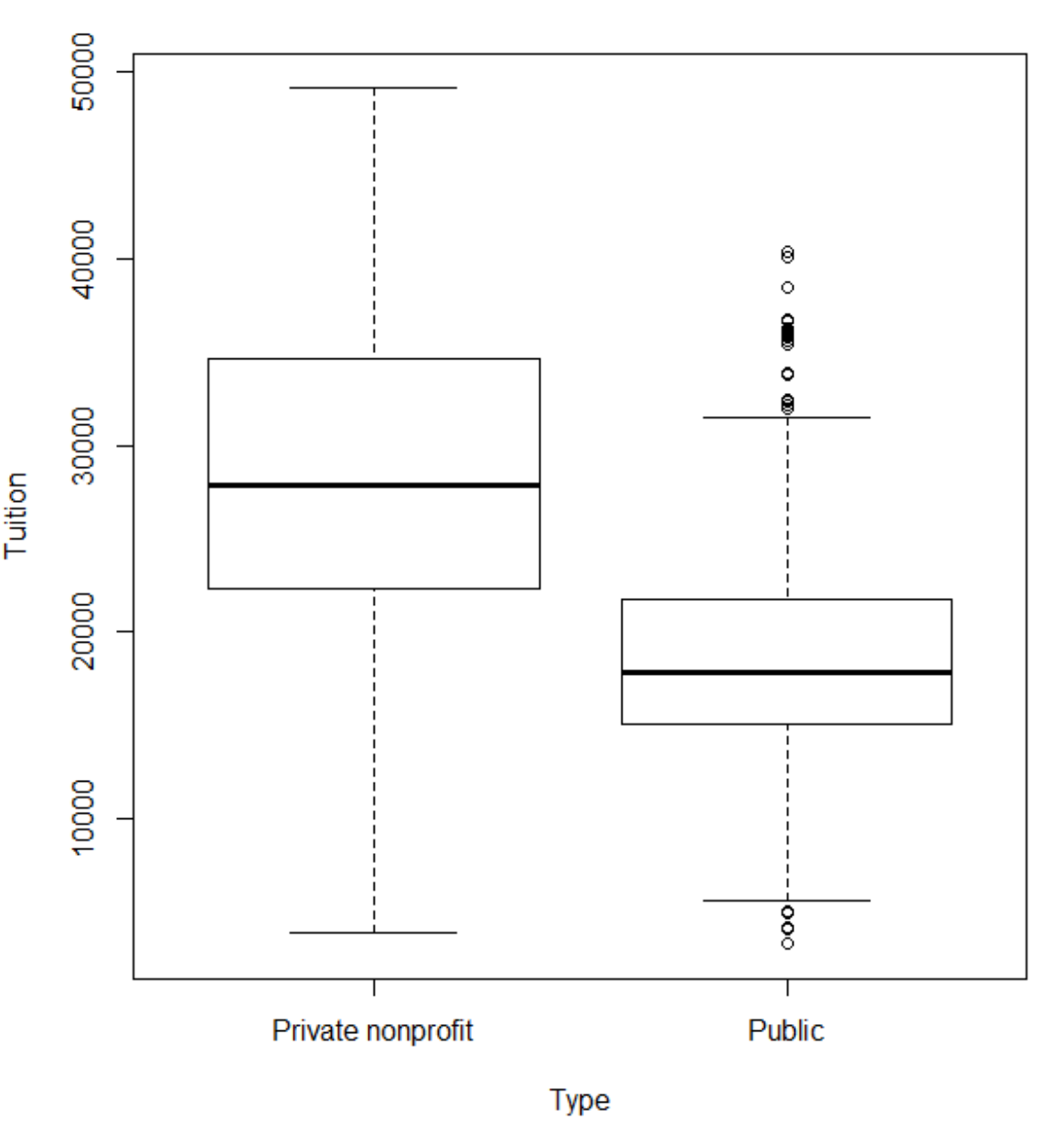
Check assumptions:

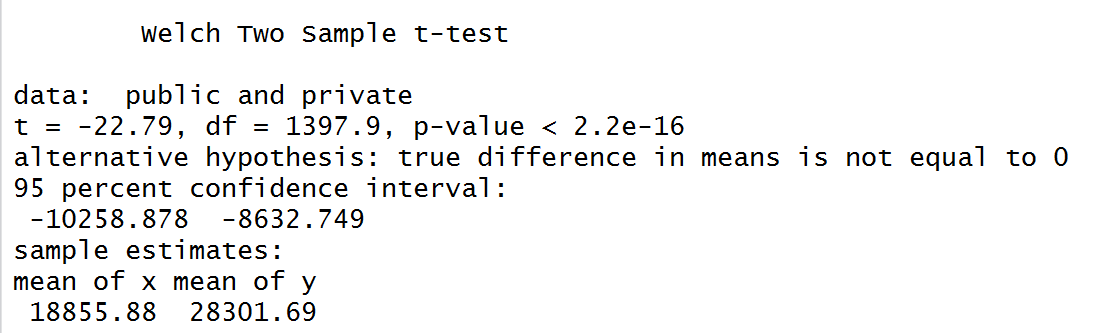
* Are variances equal? No (use Welch’s t-test)



* Are observations independent, random? Would need to understand how data was collected
* Small enough sample: 1407 observations < 10% of total number of colleges? (probably)
* Are values normal, nearly normal, or large enough sample size? Yes, sample size is much larger than 30. Double check with histogram, qqplot:

(appears roughly normal, although skewed)





P-value < 0.05, so there is a significant difference in tuition rates. As seen in the boxplot, tuition is significantly higher at private nonprofit universities.

Question 3

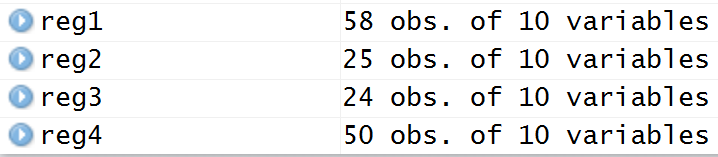
Indep variable categorical (region). Dependent variable continuous (happiness score): use one way ANOVA

Check assumptions:

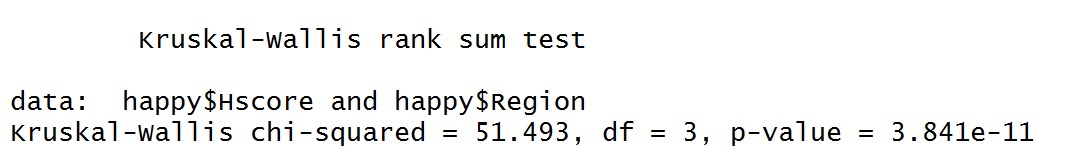
* Are observations independent, random? Would need to understand how data was collected
* Is the response variable nearly normal? Yes, histogram appears roughly normal



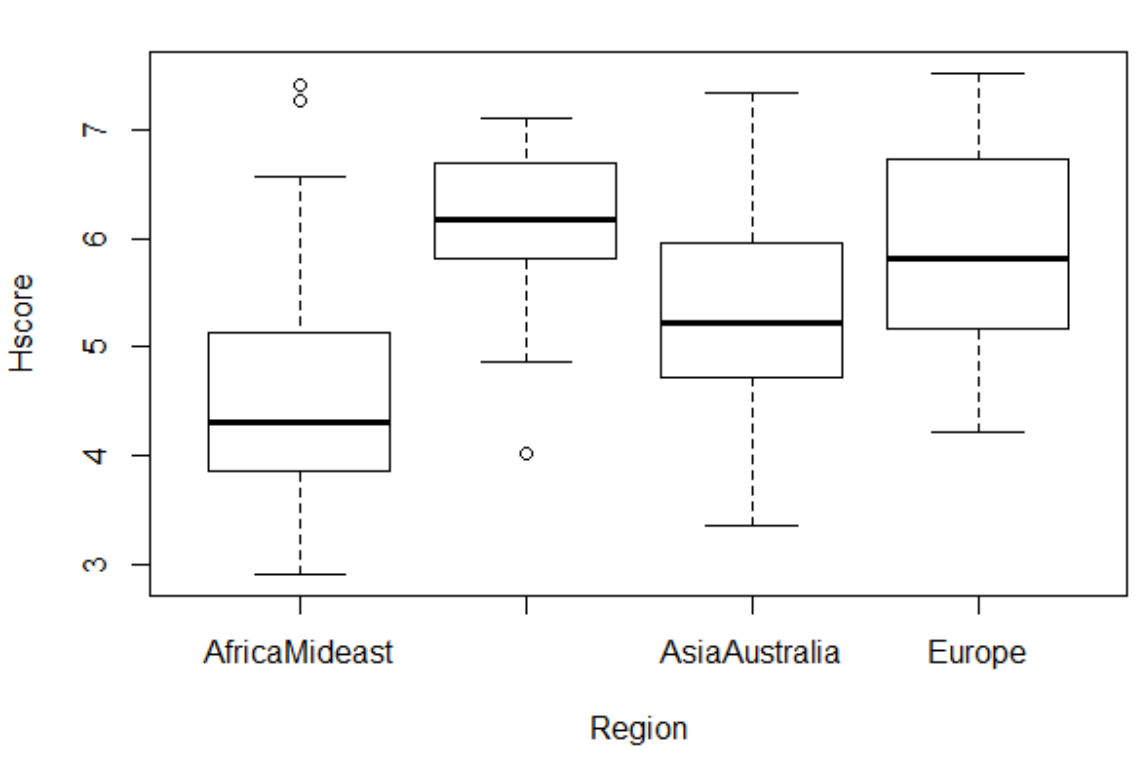
* Do the treatment groups have the same sample size? No



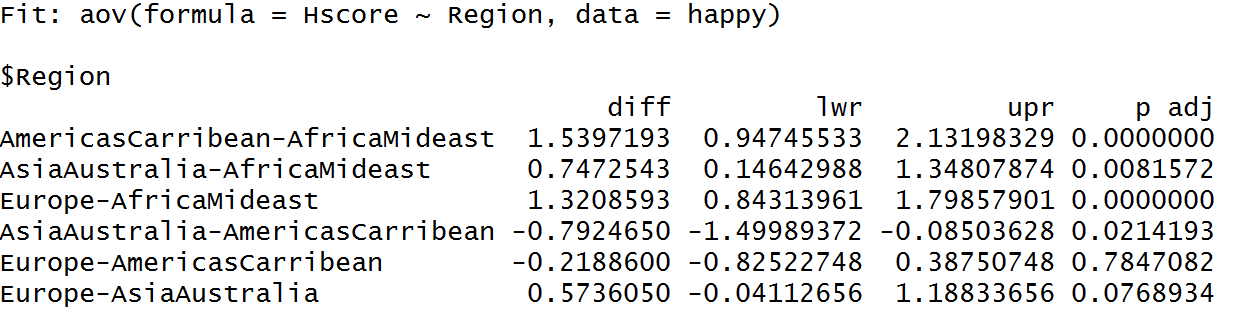
Use kruskal-wallis



P-value << 0.05: there is a significant association between region and happiness score.



Happiness scores are highest in the Americas and Carribean, followed by Europe, then Asia and Australia, with lowest scores in Africa and the Middle East. All of these differences are significant *except* for Europe vs. Americas and Carribean, and Europe vs. Asia and Australia, as shown based on the Tukey’s HSD test.



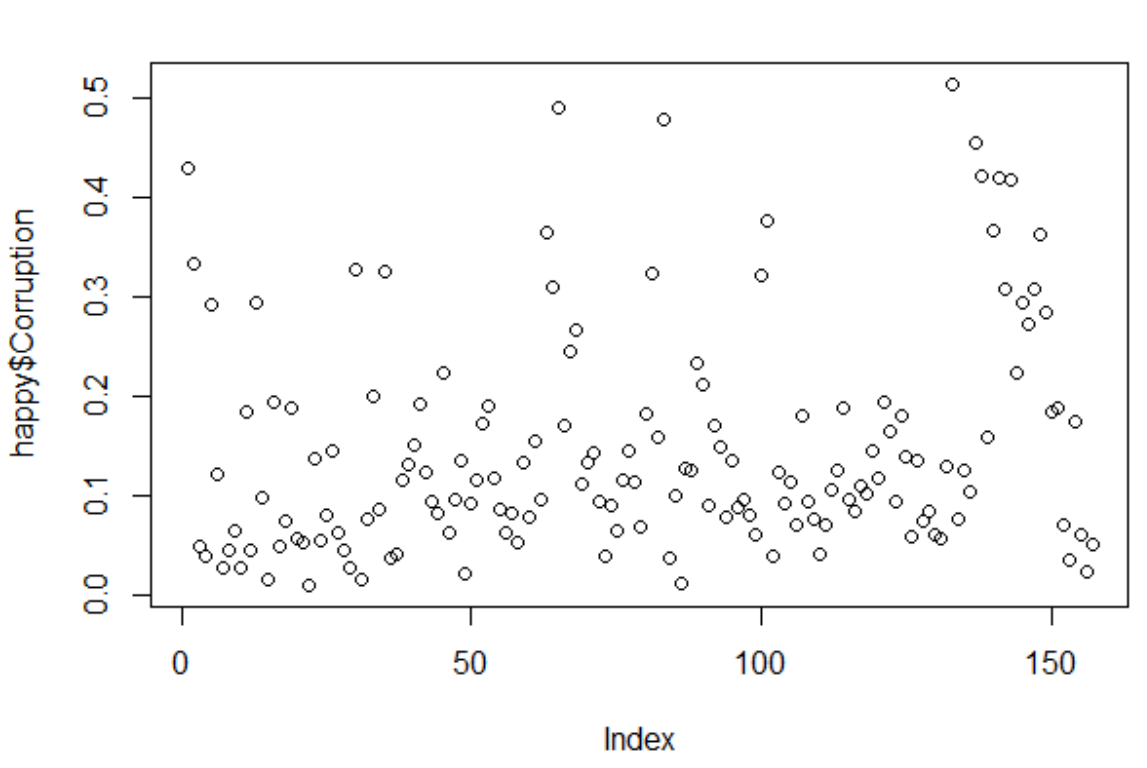
Question 4

Dependent variable (corruption) is continuous

3 continuous independent variables: use multiple linear regression

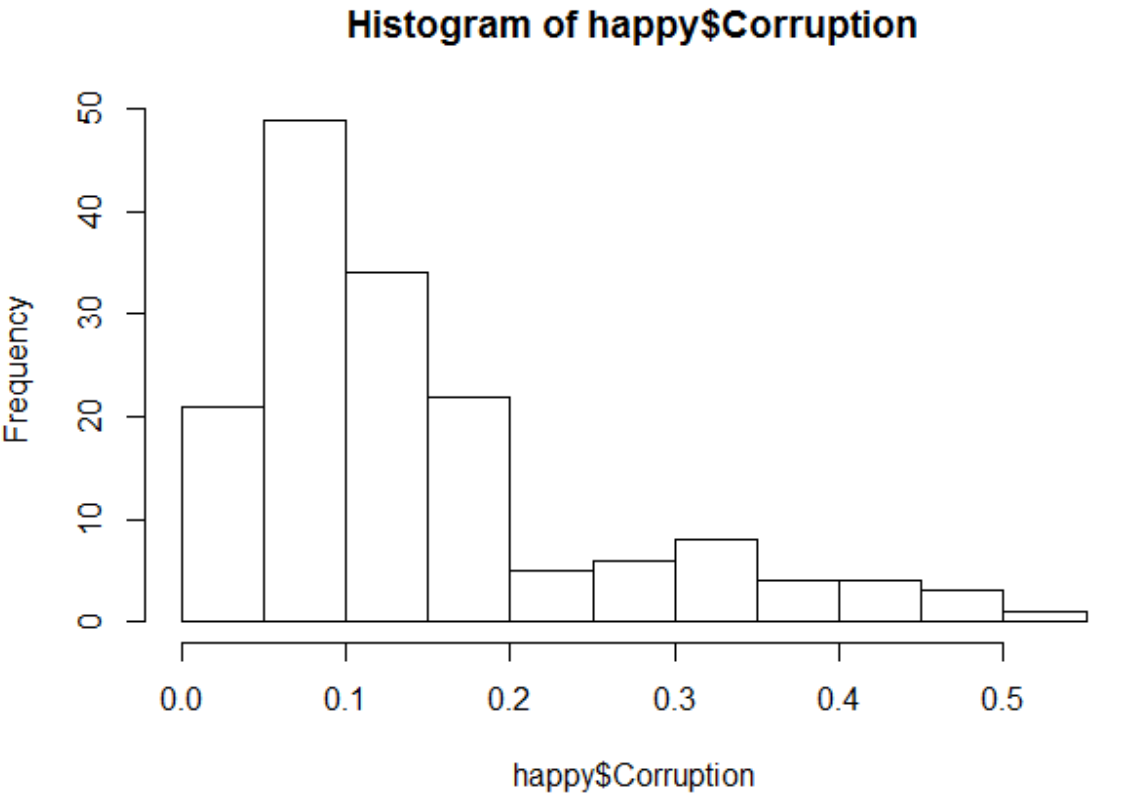
Check assumptions

* Are observations independent, random? Would need to understand how data was collected
* Homoscedasticity?



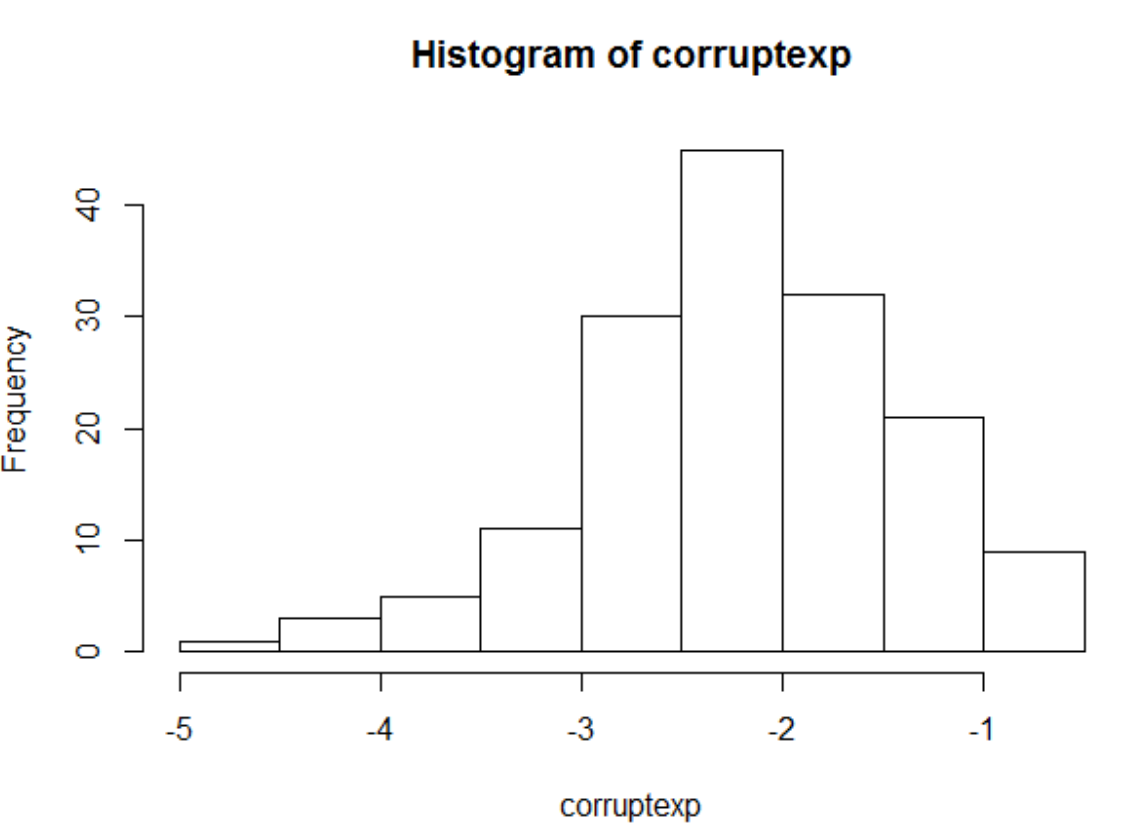
Based on the plot of the dependent variable, there is no sign of heteroscedasticity, but residuals need to be checked after the model is complete.

* Normal errors? Plot dependent variable to check



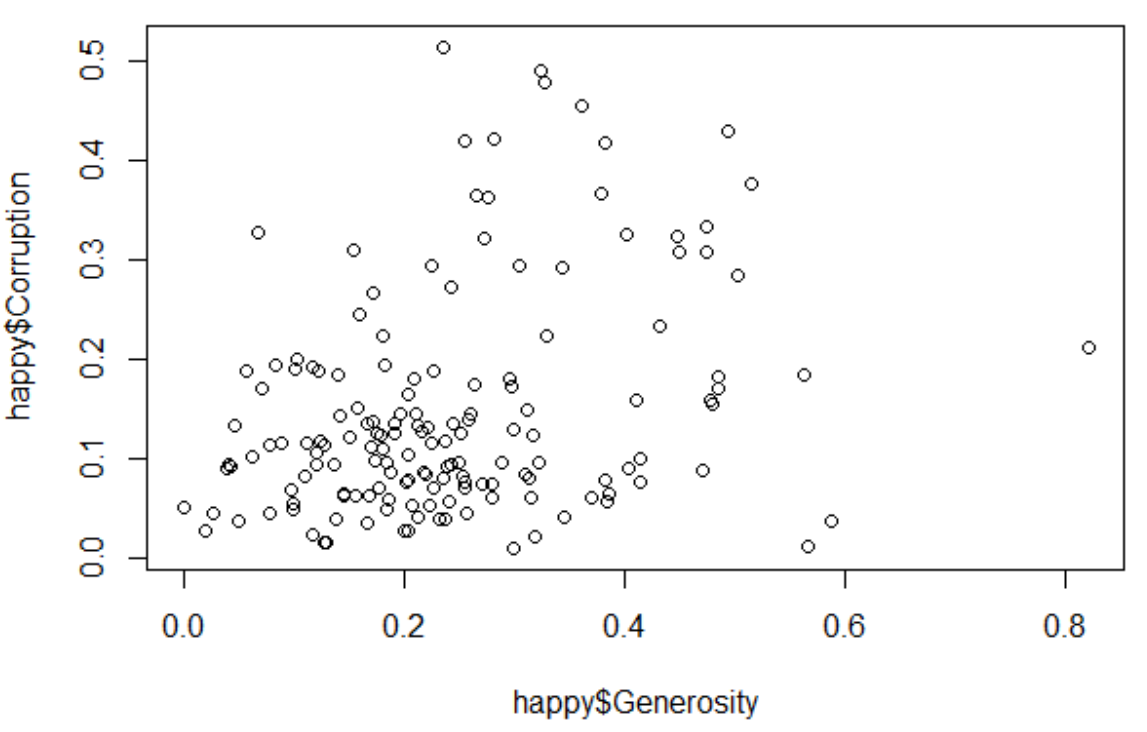
Not normal! Need to transform dependent variable

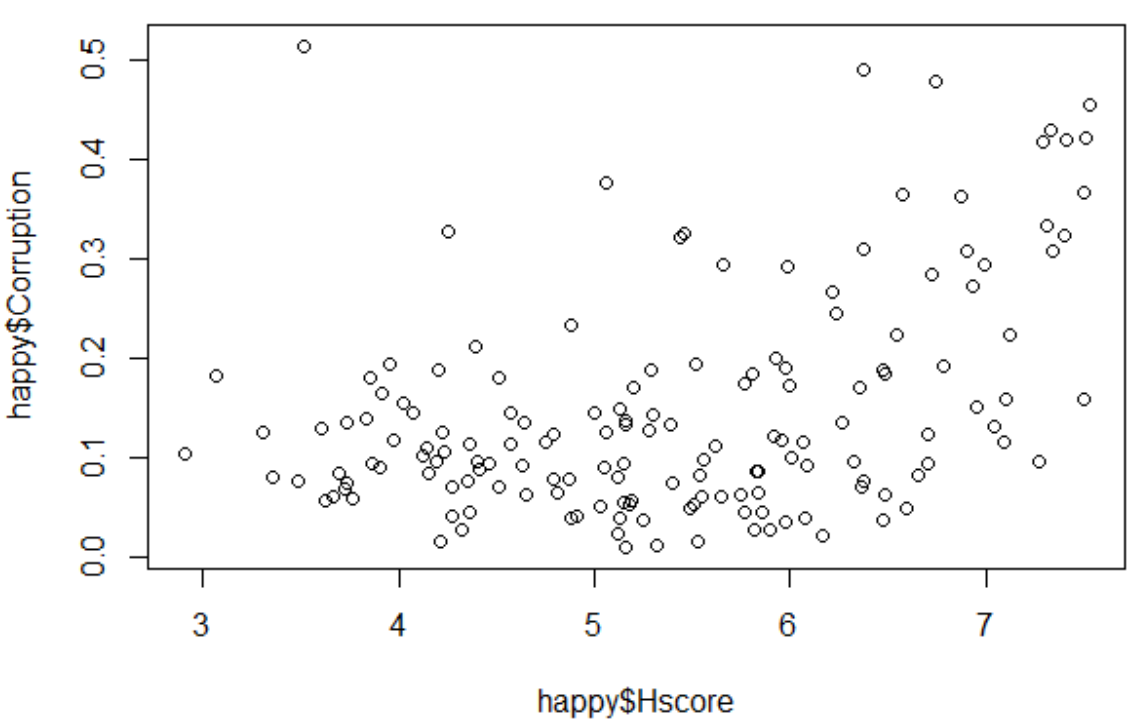
Use log(corruption)

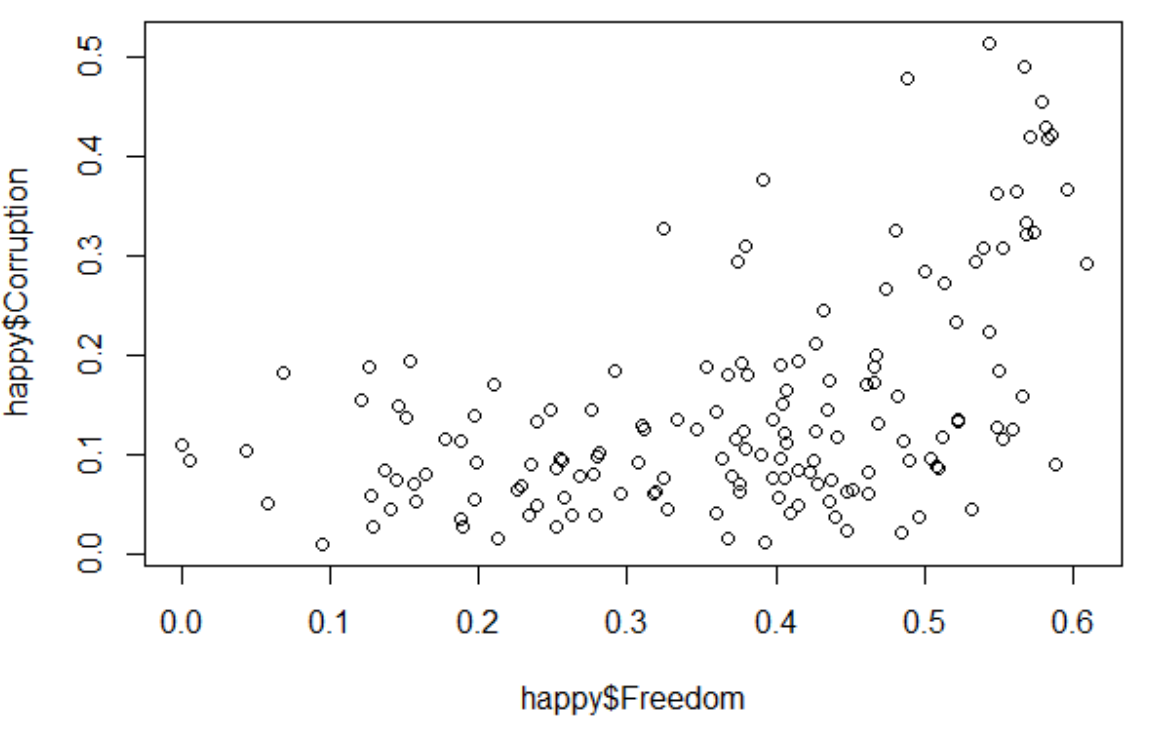


Still skewed, but looks generally normal.

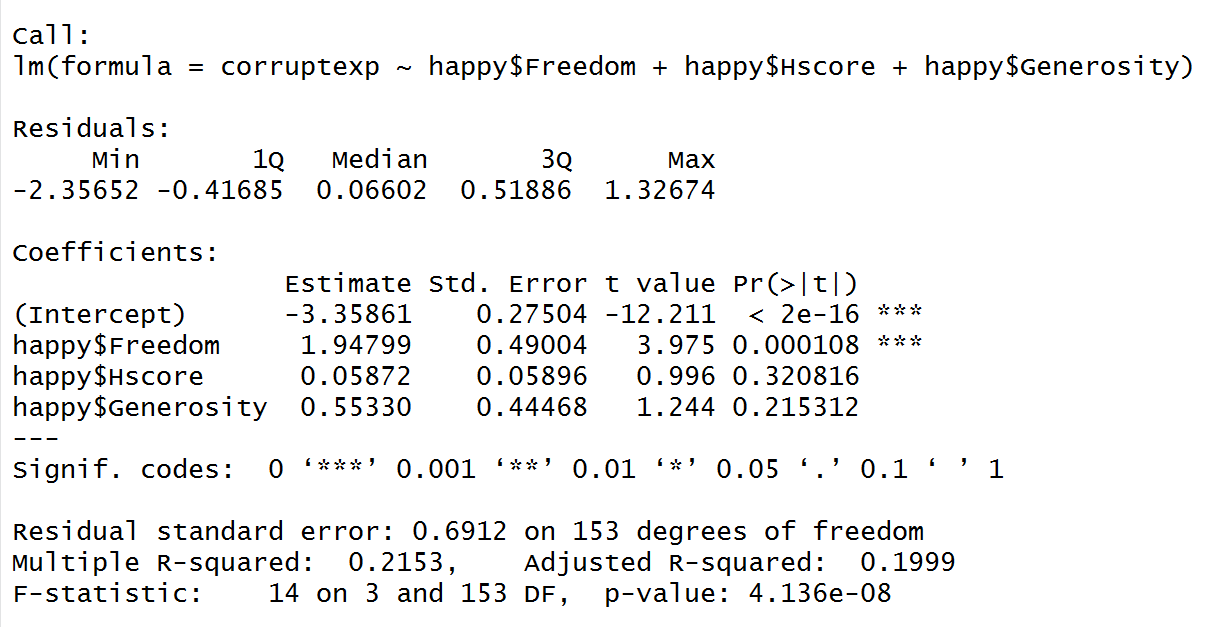
* Linear relationship between variables? Chosen independent variables: Generosity, Happiness Score, Freedom







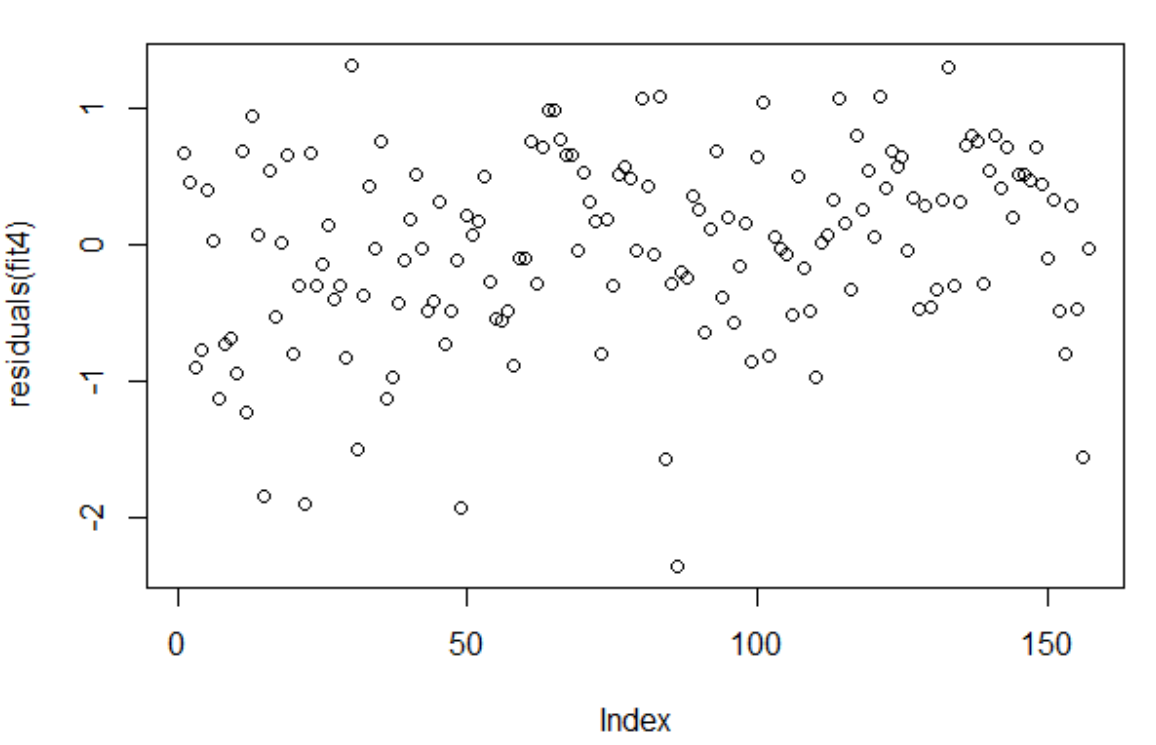
All chosen variables appear to have a linear relationship; however, they are correlated with each other. Based on the variables available, they’ll be used in the model anyway.



Only Freedom has a significant effect on Corruption, and an increase of 1 in the Freedom score predicts 1.948 increase in log(Corruption score), or e^1.948 increase in Corruption score.

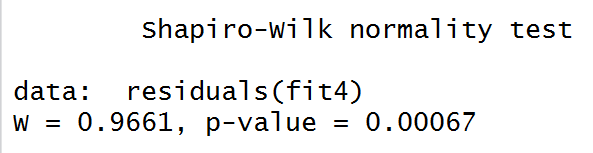
Model checking:

Adjusted R squared = 0.2: not a great fit, but not insignificant either. The low R squared value may be related to the correlation of all the variables in the dataset.



Residuals appear homoscedastic!

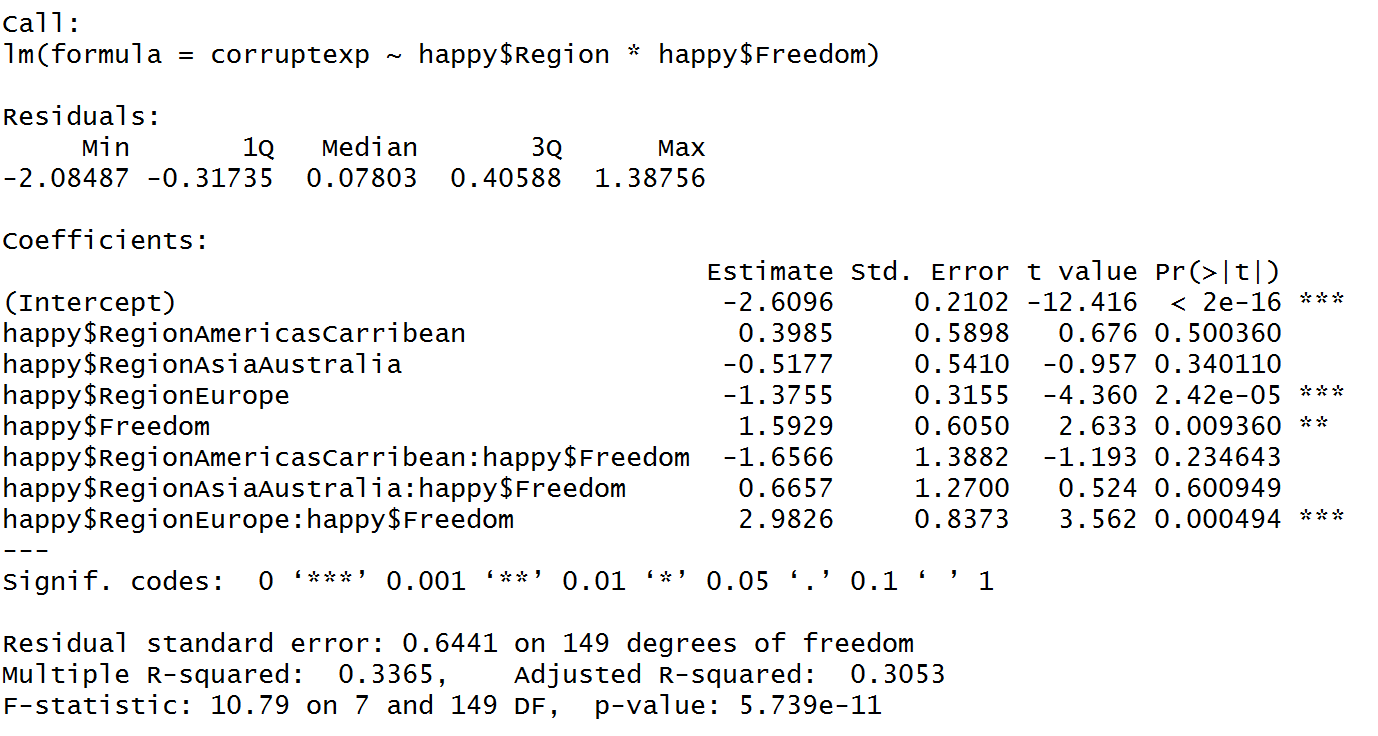
Confirmed normal errors with Shapiro-Wilk test (p < 0.05)



Question 5

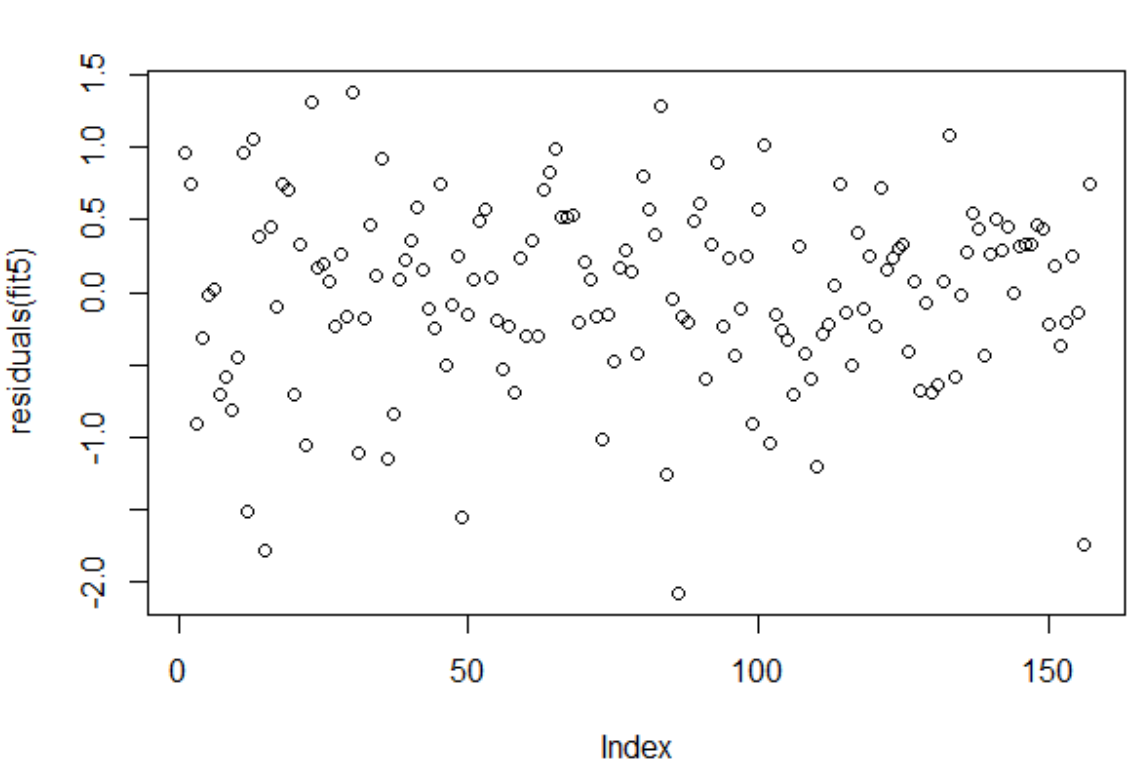
Check assumptions (same as Question 4, see above)

Continuing to use log transform of corruption



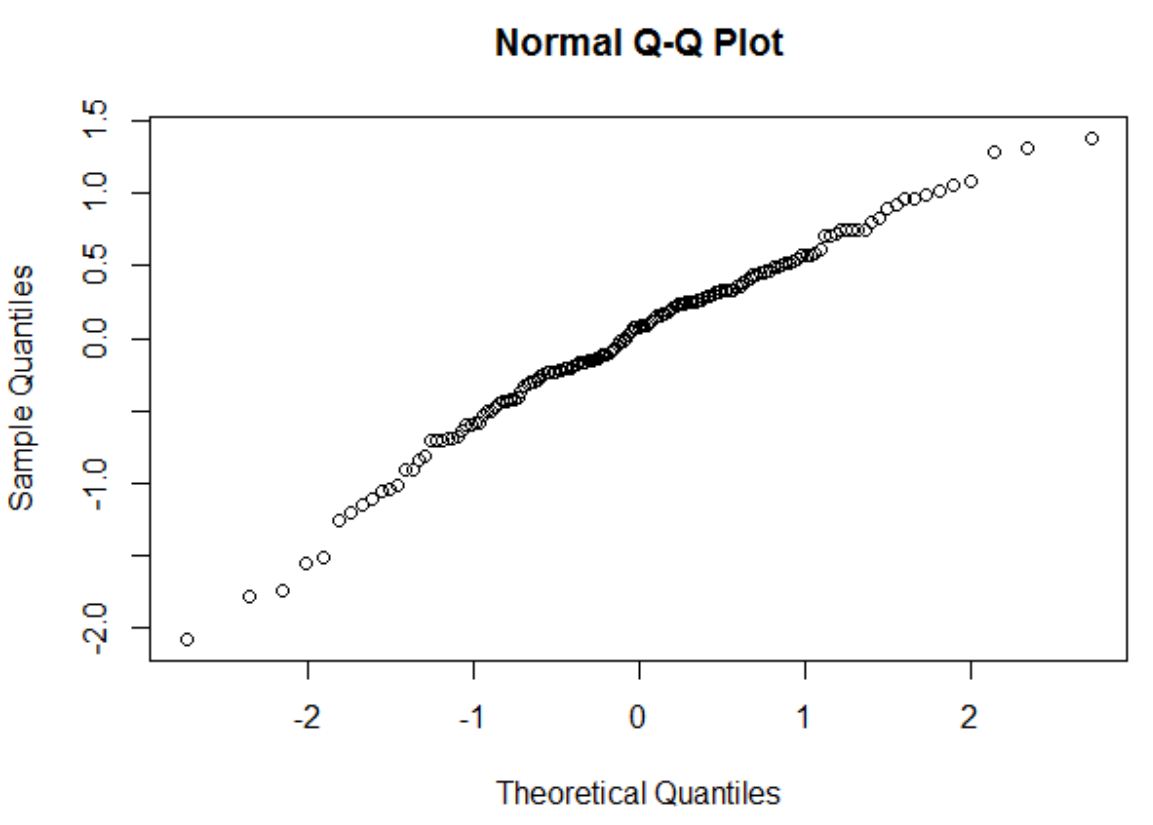
The interaction between freedom and region is only significant in Europe (p = 0.000494). The positive effect (2.9826) suggests that an increase in freedom by 1 point in Europe increases corruption by log(2.9826) points.

Model Checking:

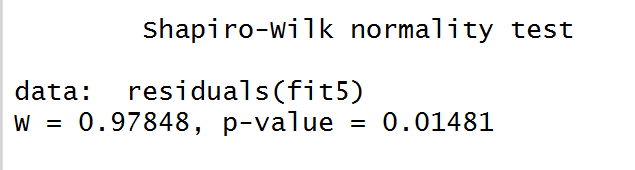


Errors appear homoscedastic

Verify normal errors with qq plot:

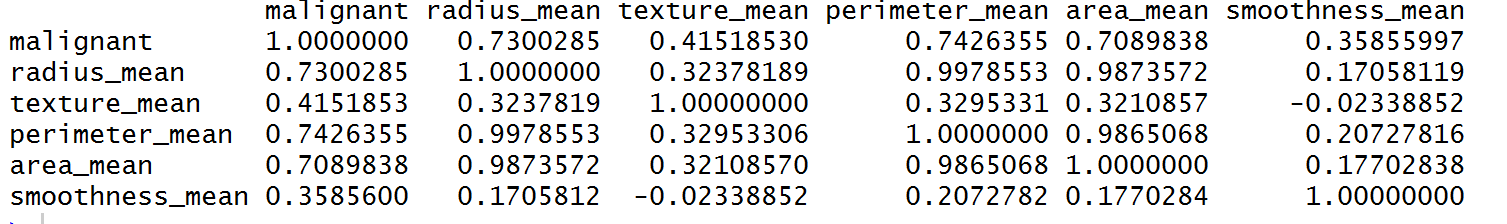


Shapiro-Wilk confirms normality



Question 6

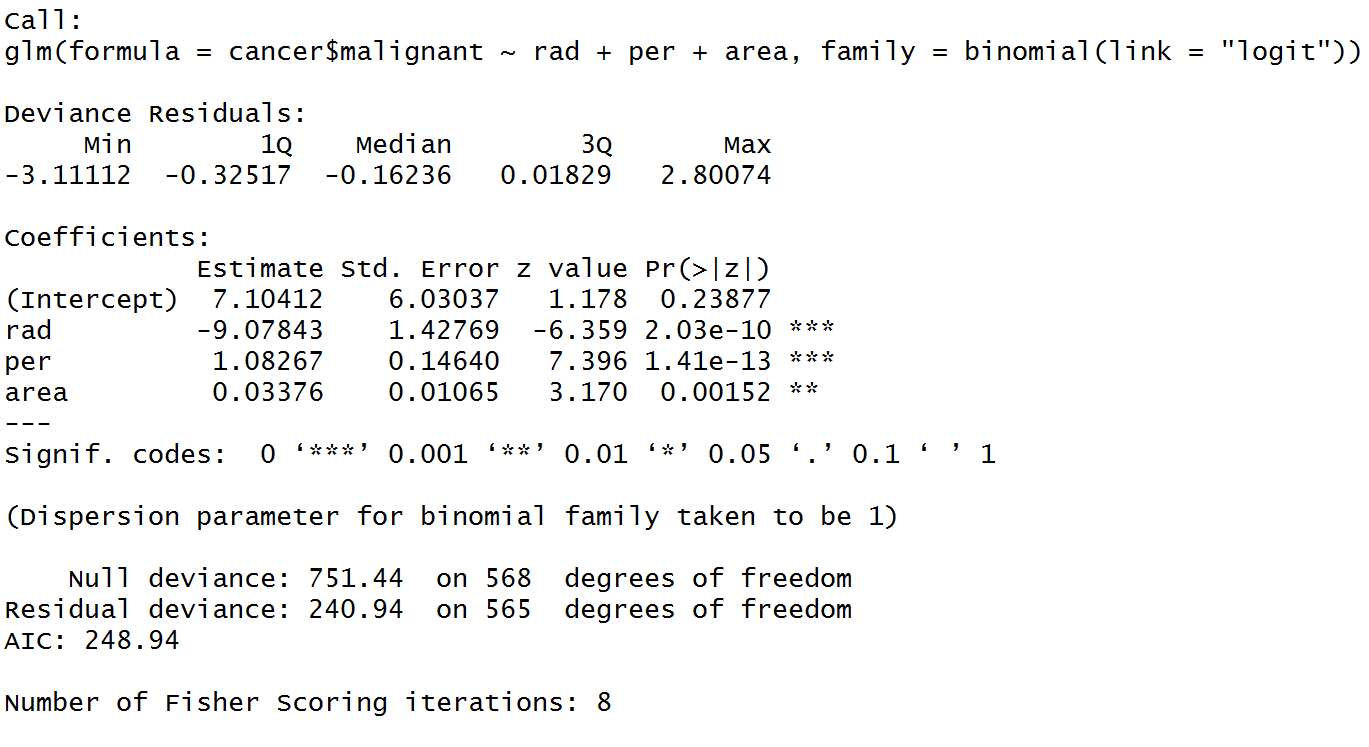
Check correlations



Three continuous independent variables: radius, perimeter, area are most correlated with malignancy.

Binary dependent variable: use logistic regression!

Need to use GLM

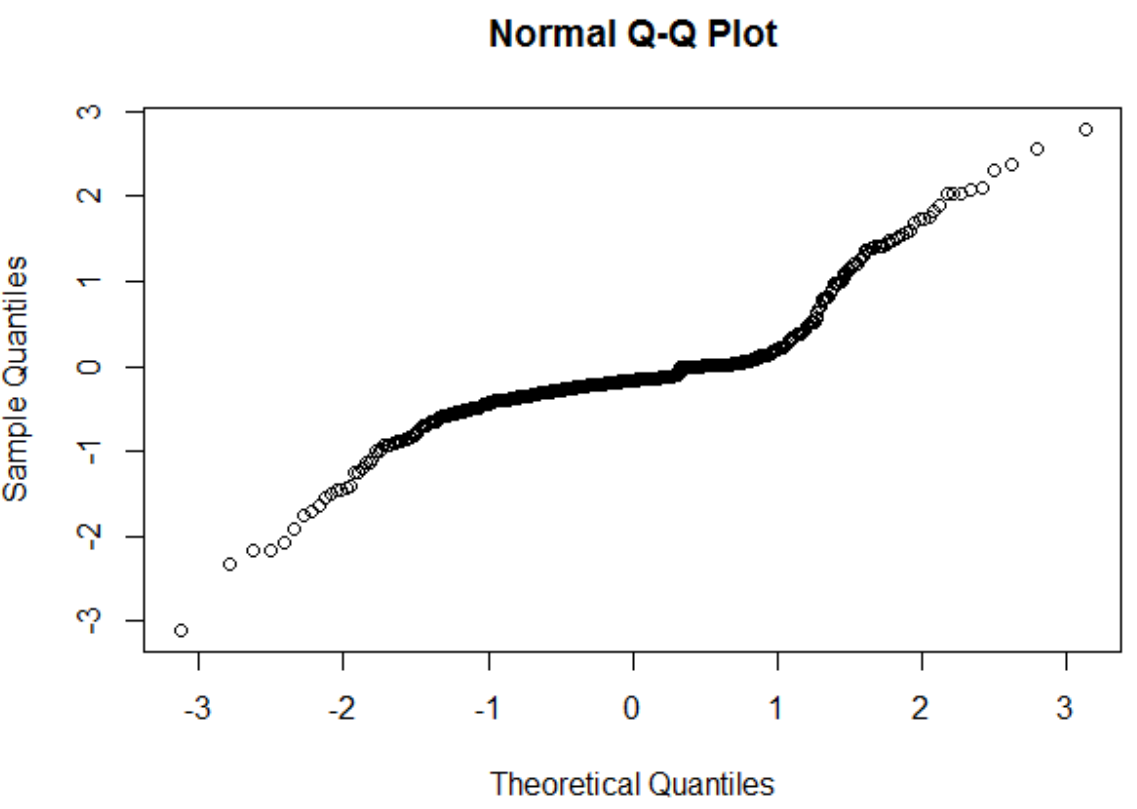


All dependent variables are significant (p < 0.05)

For each increase of 1 in radius, the log odds of the tumor being malignant decreases by 9.07843. For each increase of 1 in perimeter, the log odds of malignant increases by 1.08, and for each increase of 1 in area, the log odds of malignant increases by 0.03376.

AIC 248.94: need another model to compare this to in order to understand goodness of fit

Model checking:



Residuals don’t necesssarily appear normal based on qqplot, but Shapiro-Wilk confirms normal errors

