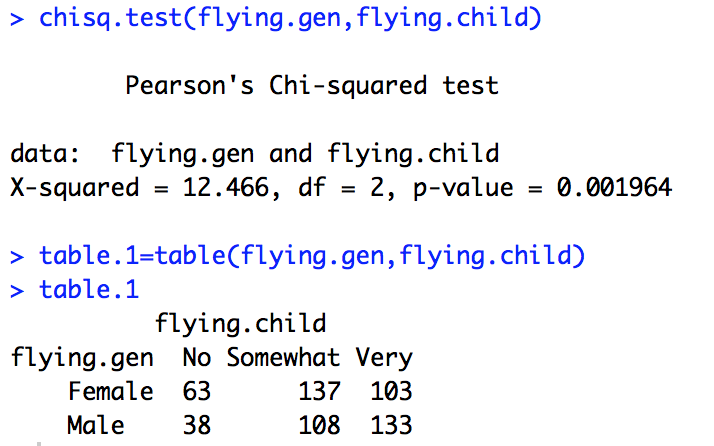
Katie Williamson

NRE 538

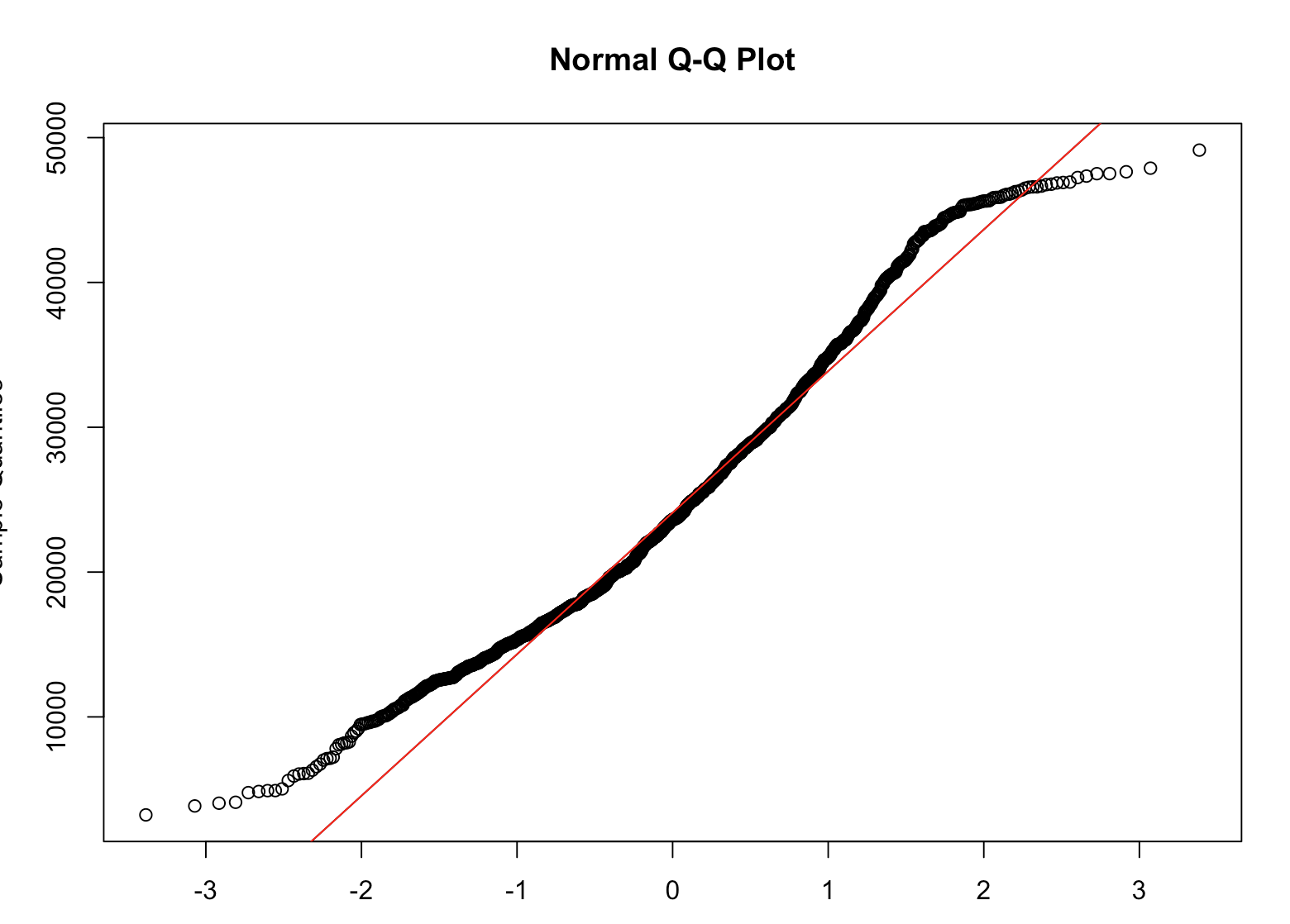
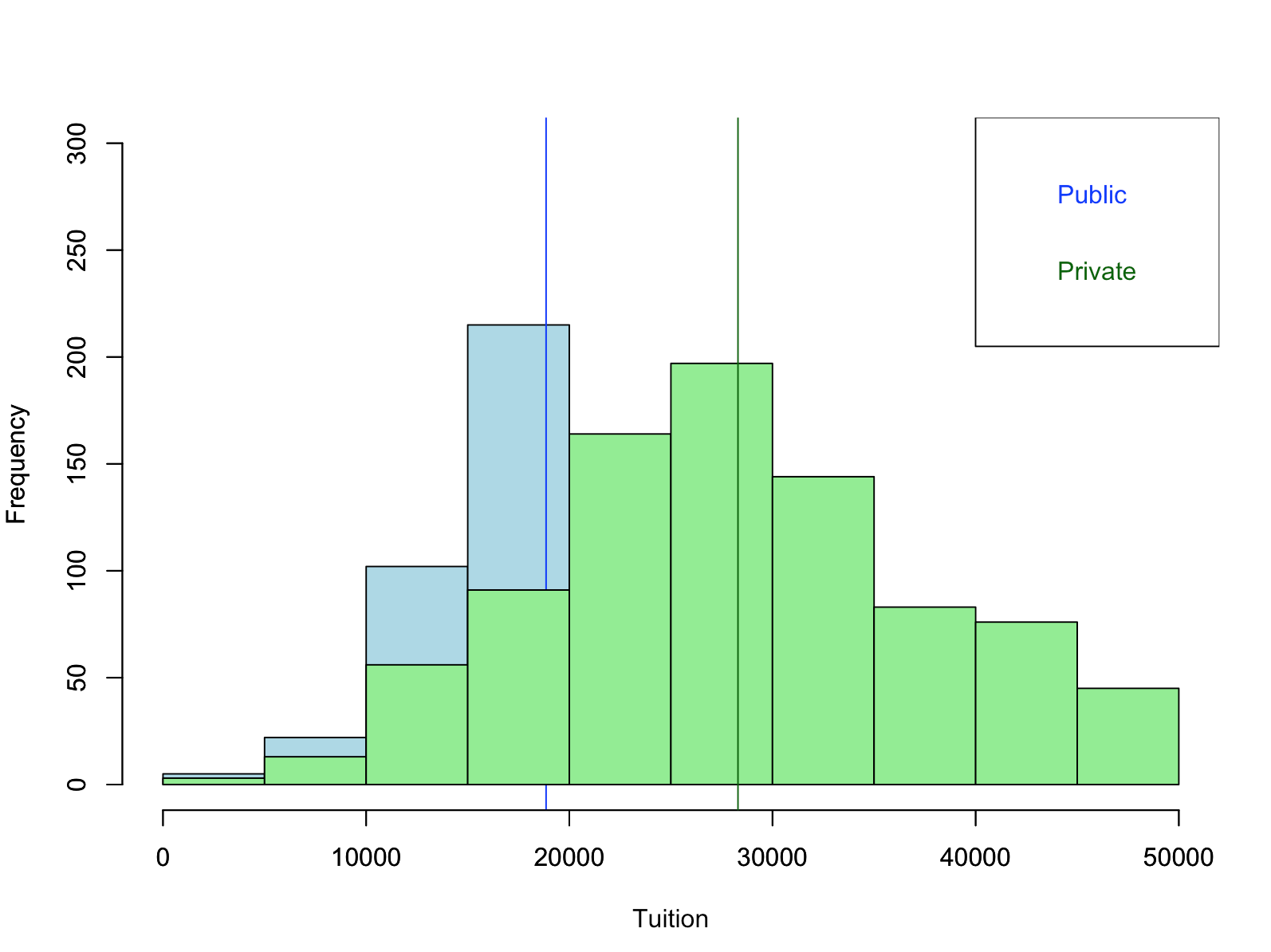
Final Exam – R

\*My answers in blue\*

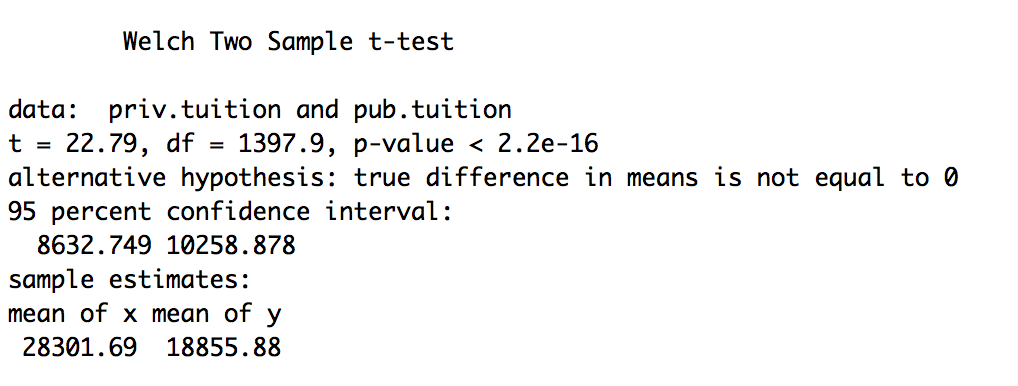
1. **Is there a significant association between gender (gender) and whether people think it’s rude to bring an unruly child on the plane (unruly\_child)? If yes, which gender tends to think that bringing an unruly child is more rude?** 
   1. **Chi-squared test**
   2. **Assumptions**
      1. Enough values for contingency table = yes, there is a large number of values of 587 after cleaning for NA values, and all cells are filled
      2. Random sample = yes, assume from the dataset
      3. Independent observations = yes, assume from the dataset
   3. **Results** 
      1. Chisq.test produces p-value of .001964, therefore yes, there is a significant association between gender and thinking it’s rude to bring an unruly child on the plane. Therefore we would reject the null hypothesis that these variables are independent.
      2. According to the contingency table, males in this sample more often say it is “very” rude for an unruly child to be brought on the plane and women more often think it is not rude.



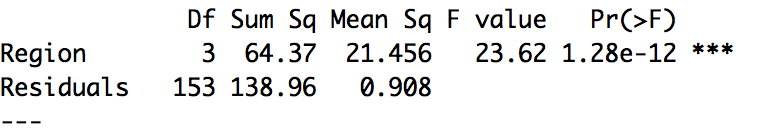
1. **Is there a significant difference in tuition (tuition) by type of institution (type)? If yes, which type has a higher tuition?** 
   1. **Unpaired two-sample T-test**
   2. **Assumptions**
      1. Continuous data = yes can assume from data
      2. Y variable (tuition) is normally distributed = yes passes, fails the Shapiro test, but through the qqplot and central limit theorem, we can assume normality for tuition variable as well as for both private and public tuition subsets

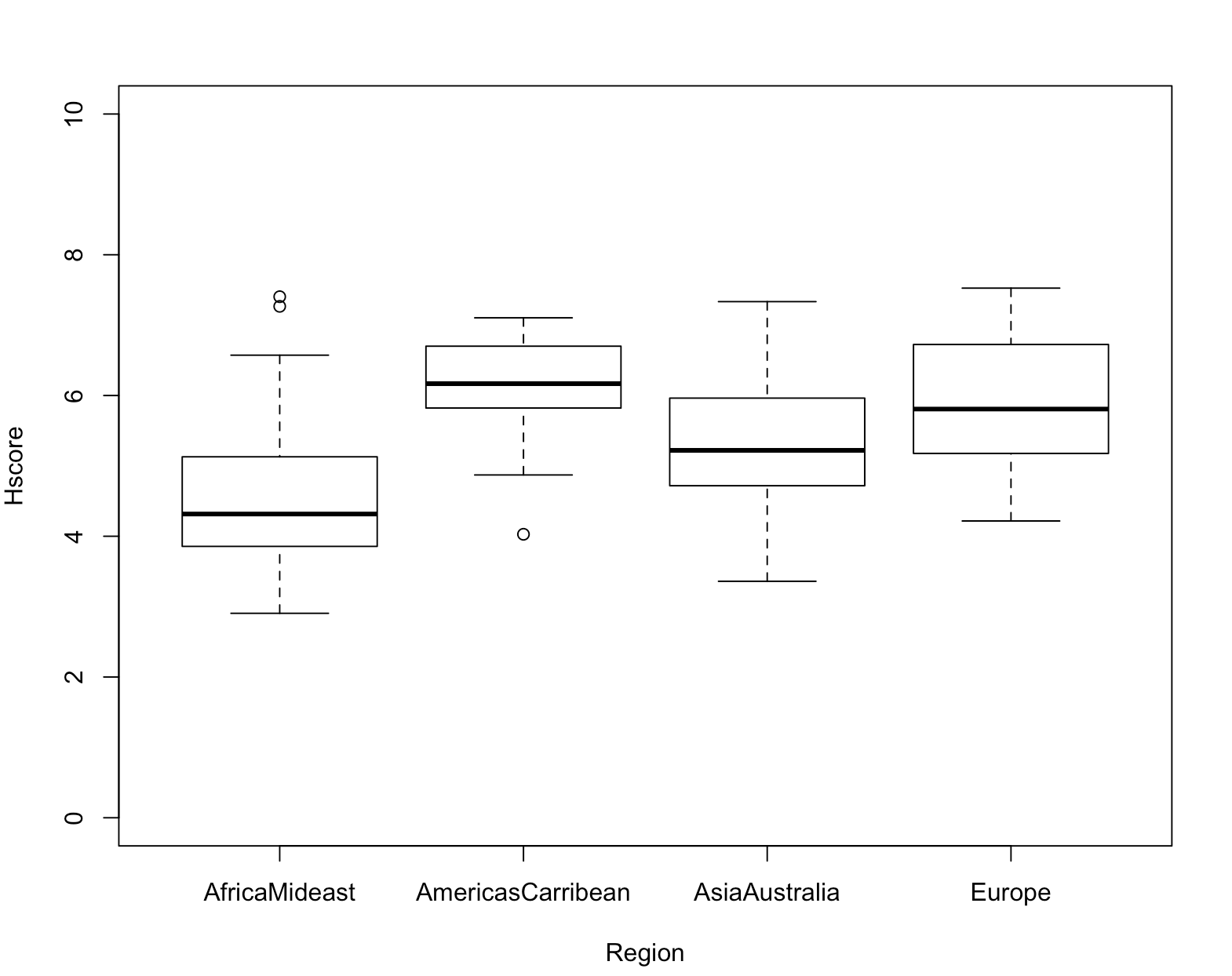


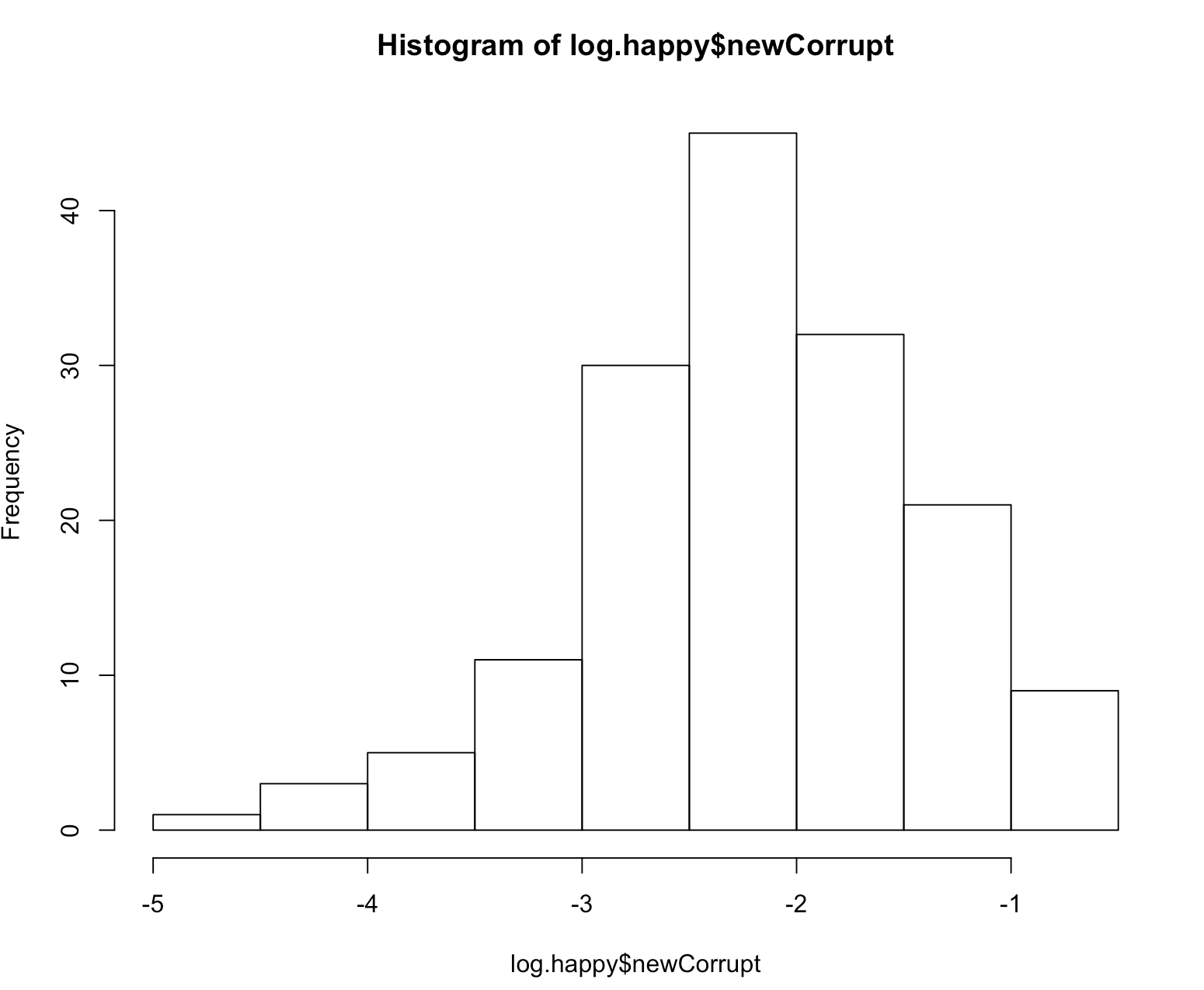
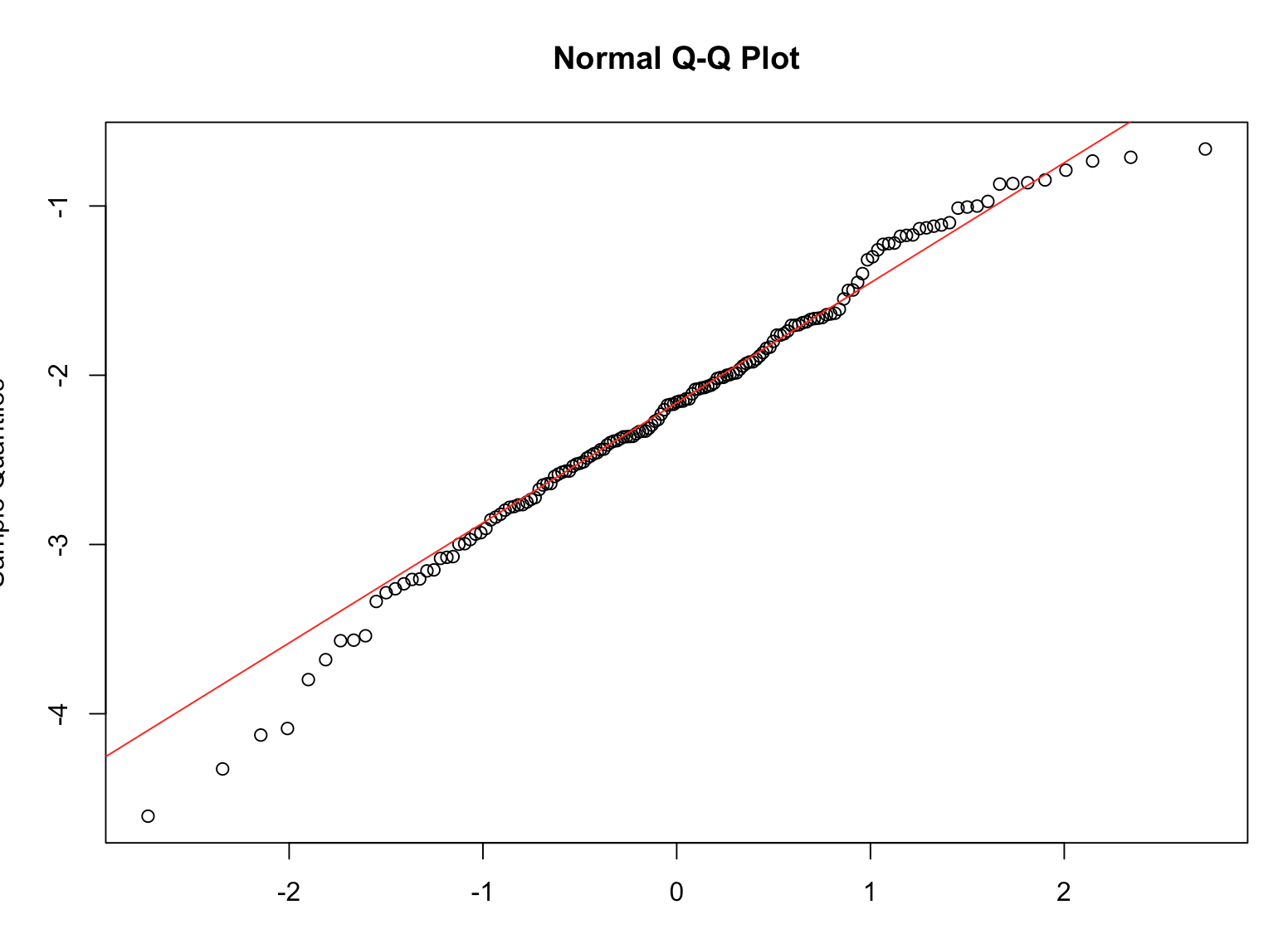
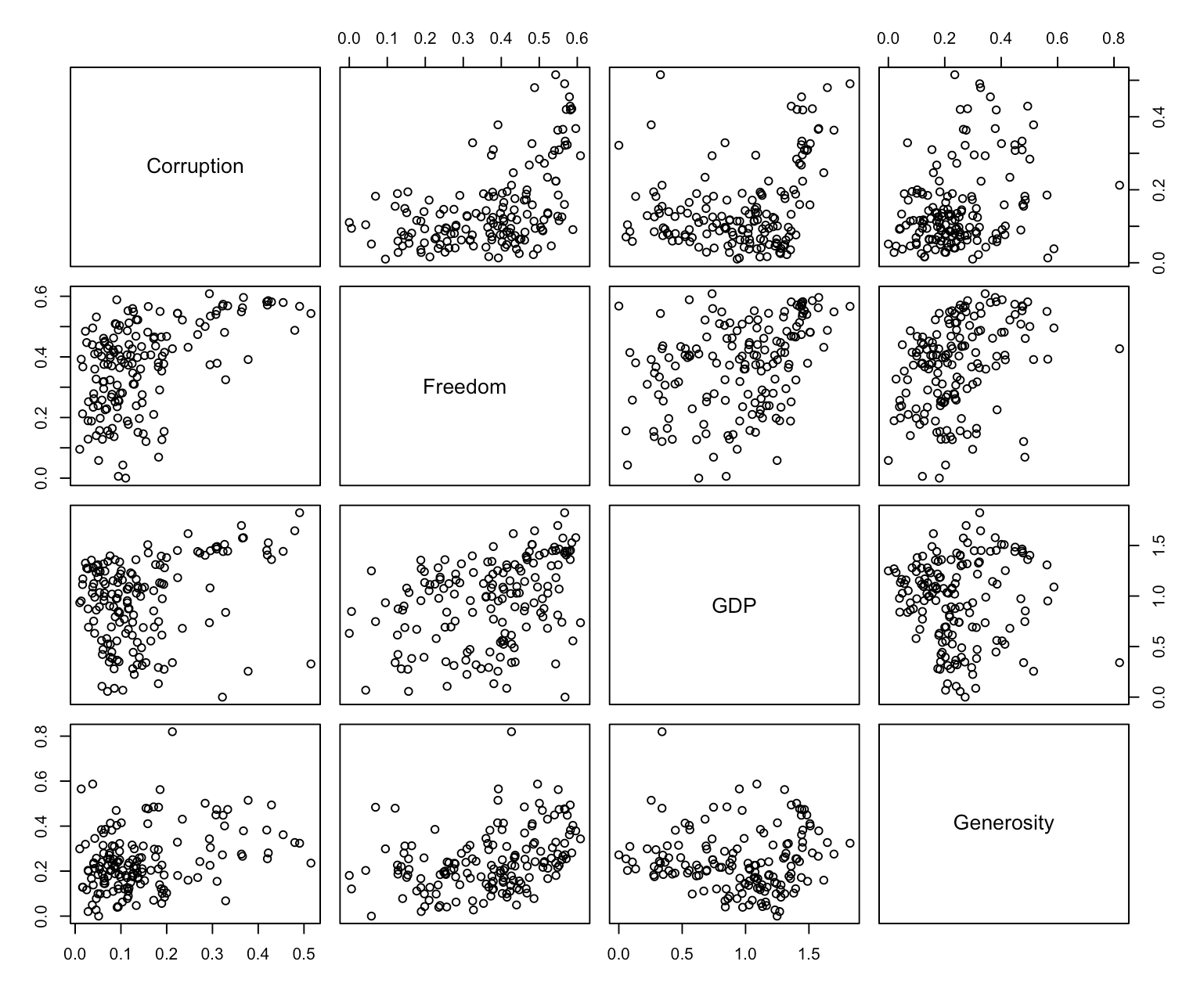
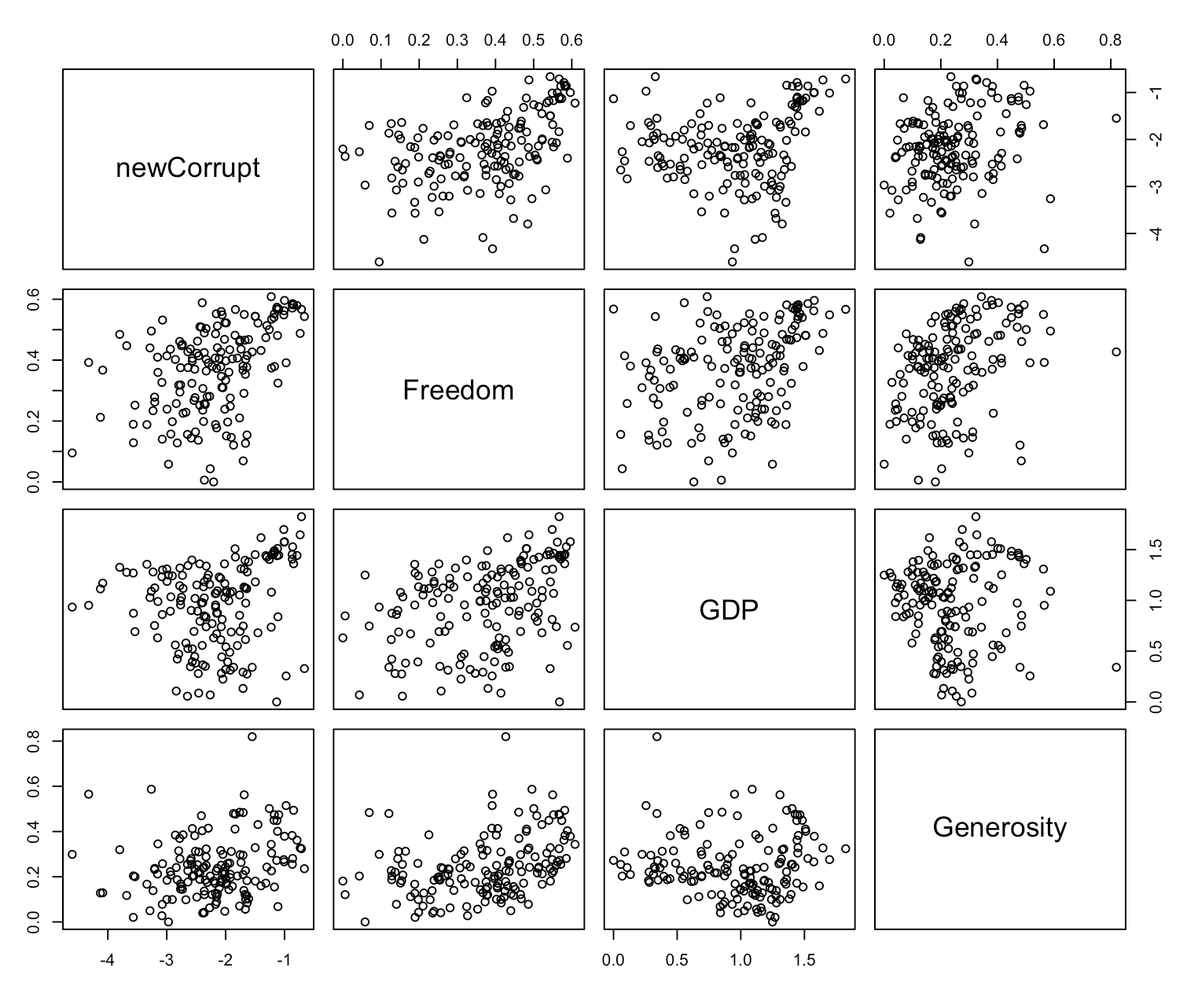
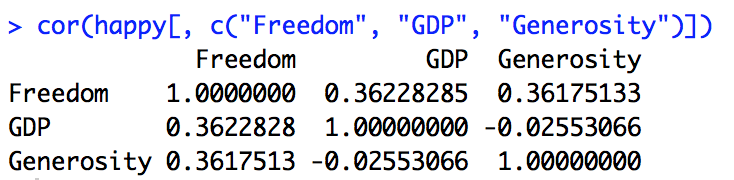
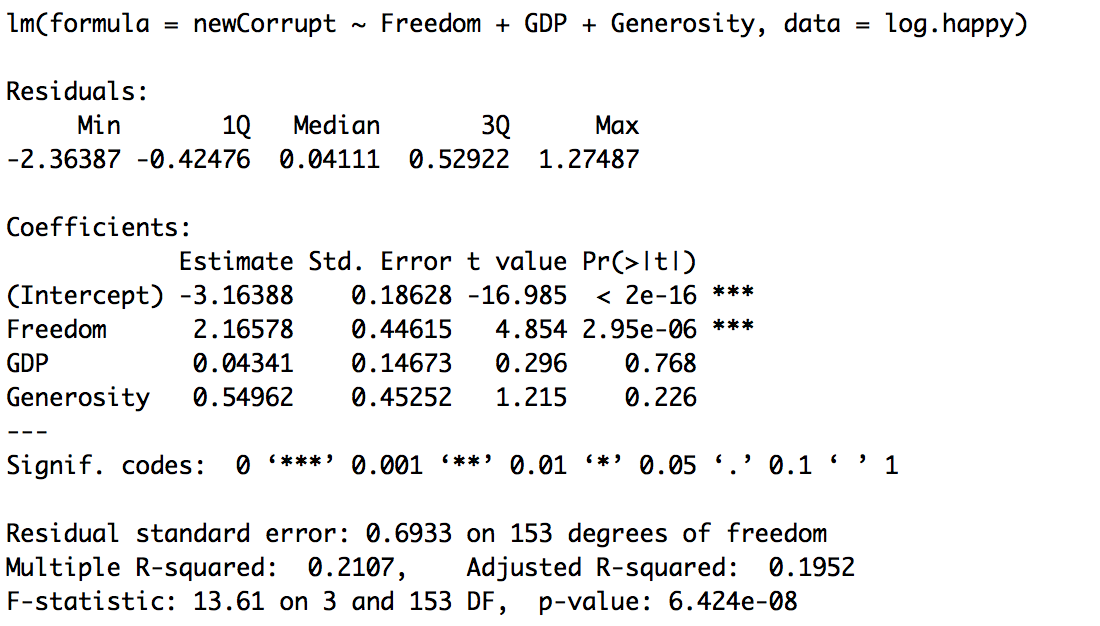
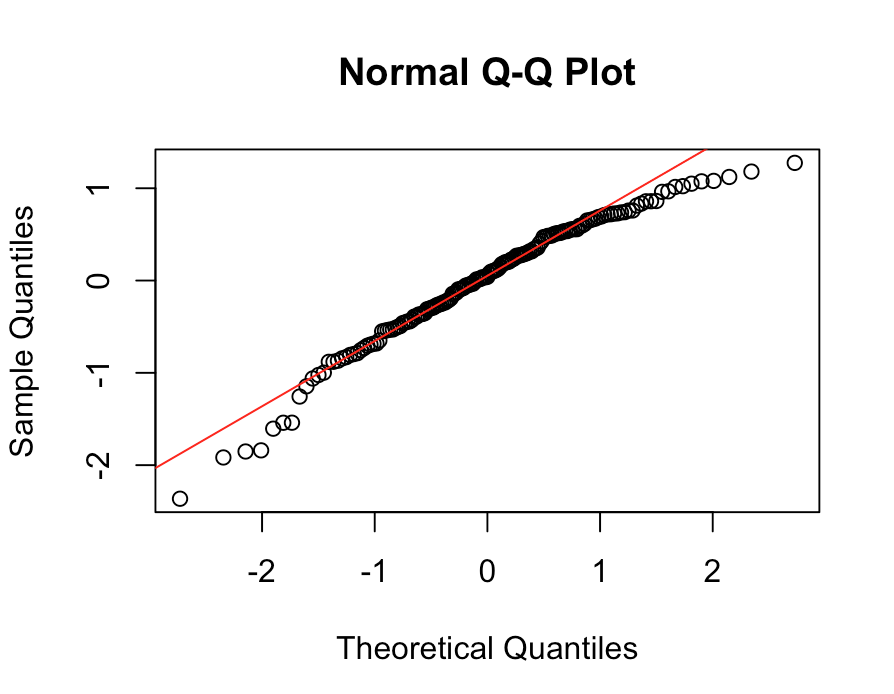
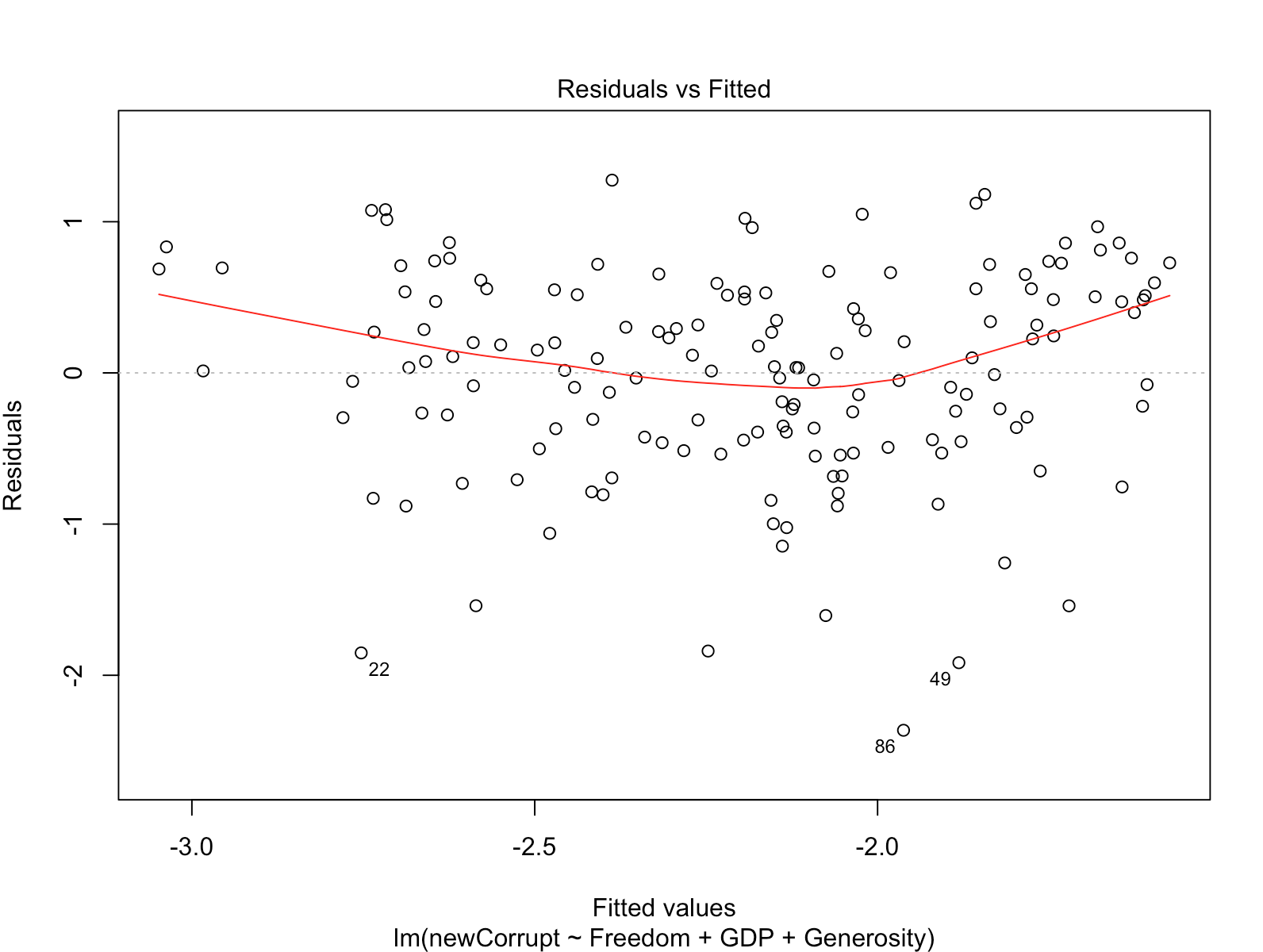
* + 1. Random sample = yes, assume from dataset
    2. Data is independent = yes, assume from dataset
    3. Equal variances = no violates this assumption where var.test produces p~0 🡪 so we use Welch’s t-test
  1. **Results**
     1. Welch’s t-test gives p-value~0, therefore there is a significant difference between private and public university tuition, and we can reject the null hypothesis that the difference between the tuition of these university types is 0.
     2. Private tuition is higher than public tuition with means of $28,301.69 and $18855.88 respectively.

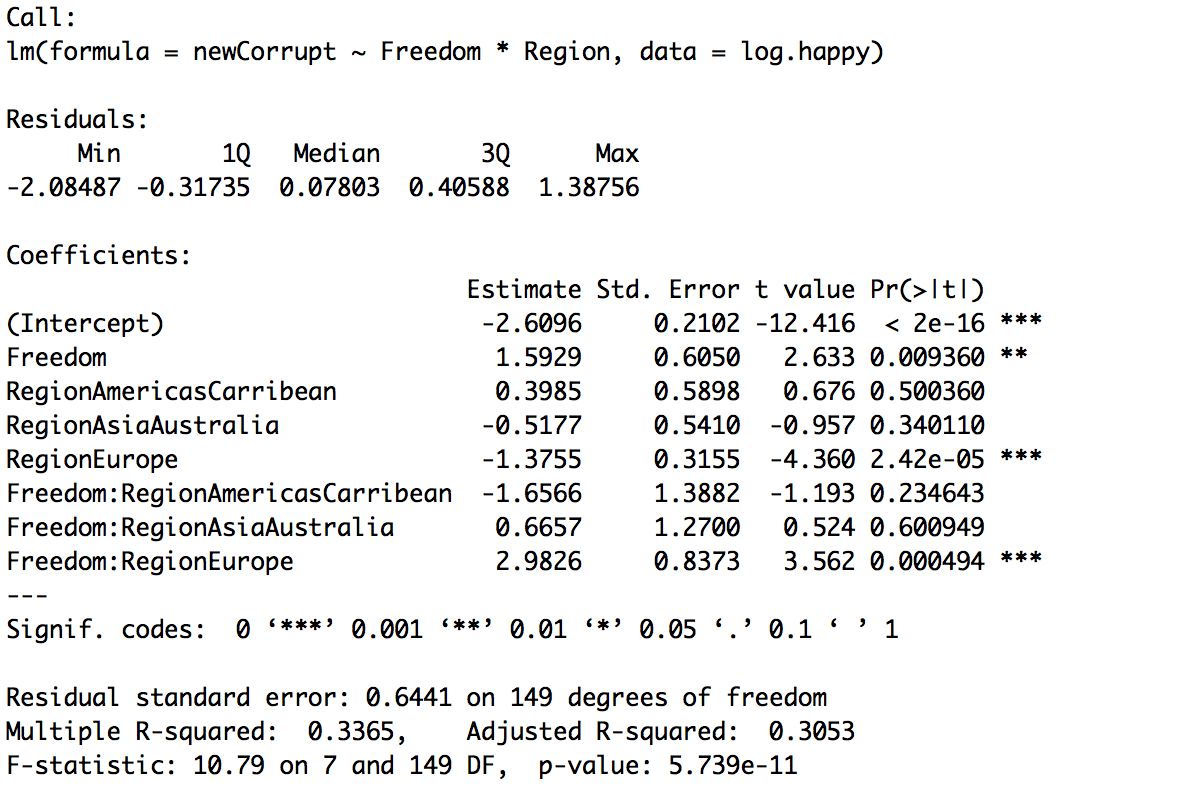


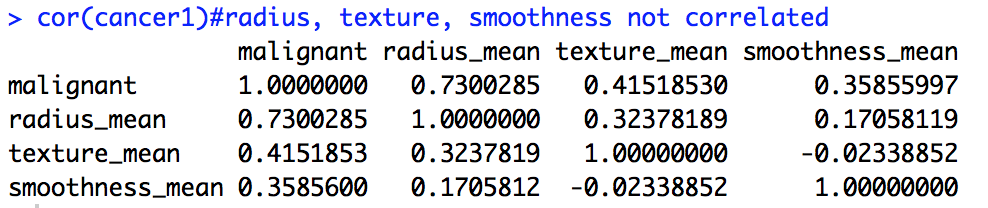
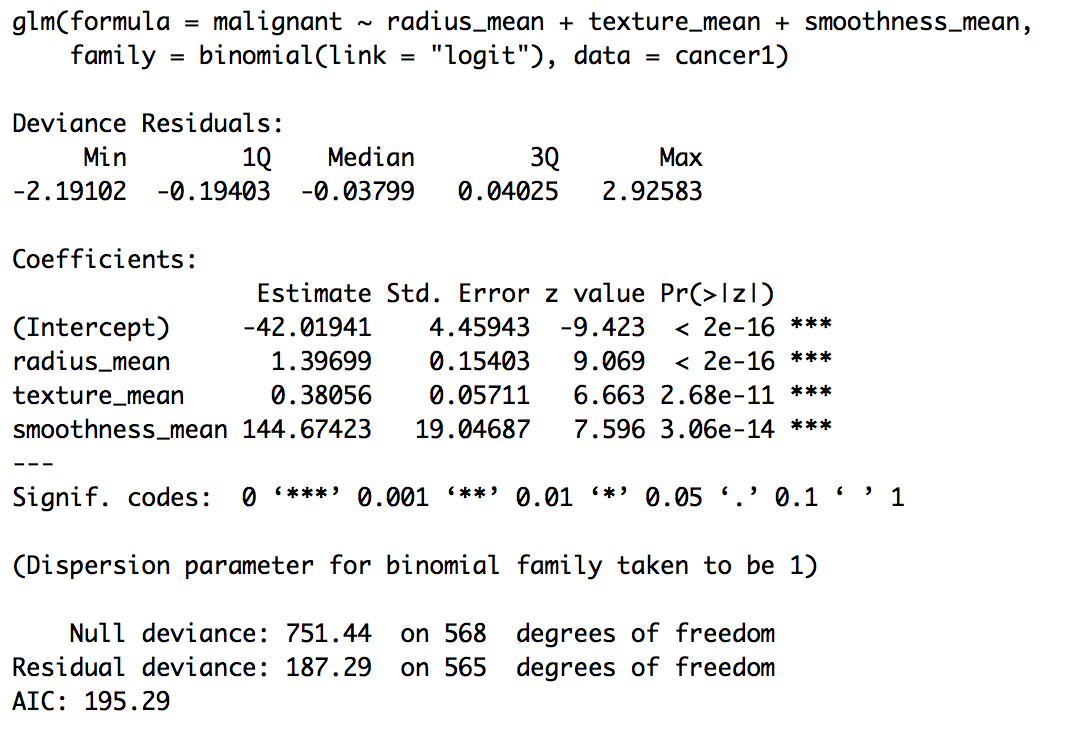
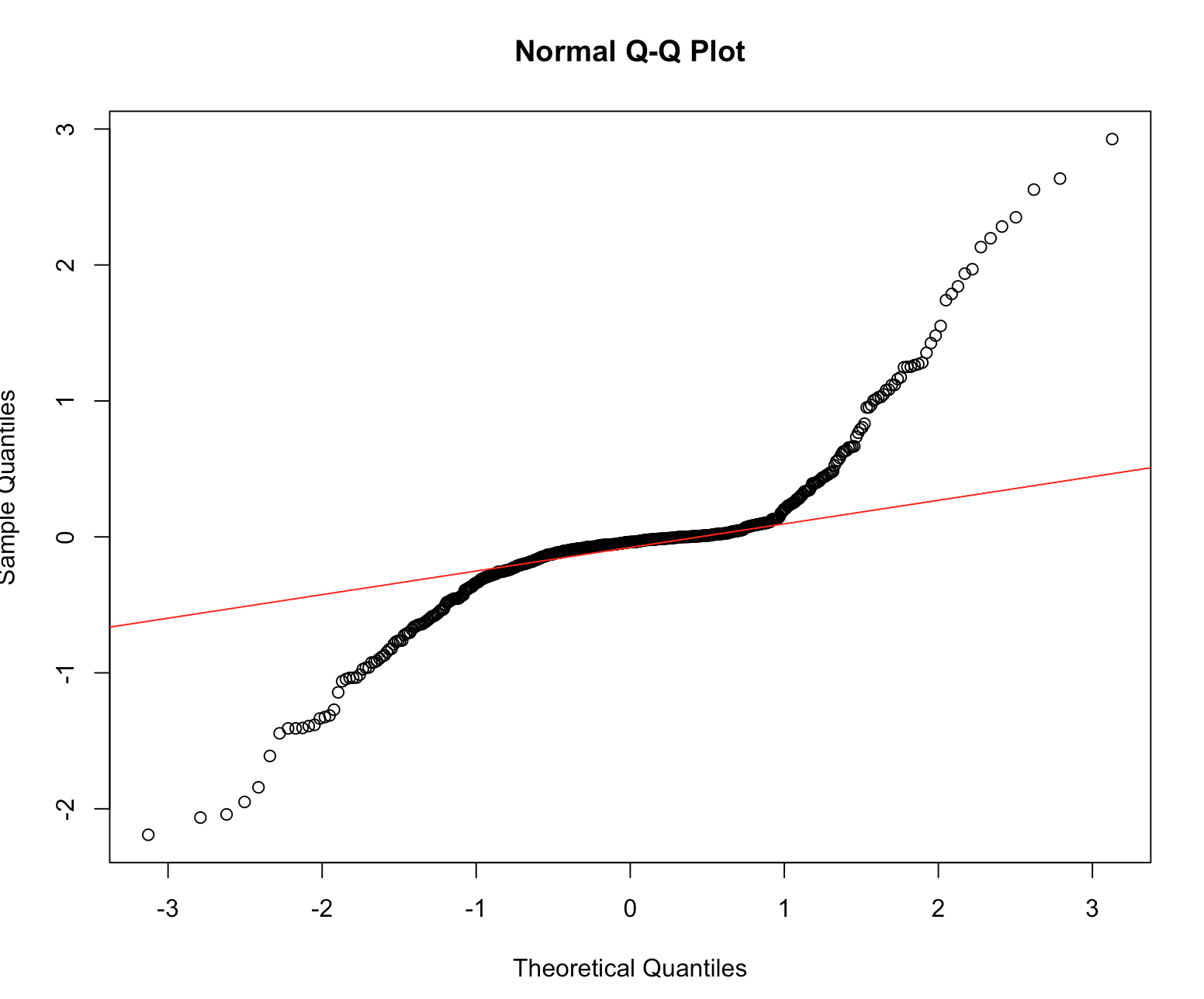
1. **Is there a significant difference in happiness (Hscore) by region (Region)?** 
   1. **One-way ANOVA (4 levels)**
   2. **Assumptions**
      1. Continuous data = yes, can assume from dataset
      2. Normal distribution of y variable (Hscore) = yes passes, qqplot looks good and there is a large sample size of 157, so central limit theorem applies despite data failing the Shapiro test
      3. Random sample = yes, can assume from dataset
      4. Independent observations = yes, can assume from dataset
      5. Equal variances = yes passes, Levene test gives p-value of .51 therefore we cannot reject the null hypothesis that the variances are equal.
   3. **Results**
      1. Yes, there is a significant difference in happiness by region where p~0. Therefore we can reject the null hypothesis that there is no difference in Hscore between the regions.
      2. Using a Tukey HSD test and the boxplot we see that Americas-Caribbean has the highest mean Hscore.





1. **What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model.** 
   1. **Multiple linear regression** (Corruption ~ Freedom + GDP + Generosity)
   2. **Assumptions**
      1. Continuous data = assumed from dataset
      2. Y variable (corruption) normal = no fails in original form 🡪 using a log transform fixes normality with qqplot and Shapiro test p-value =.085
      3. Linear relationship = yes, from the plots above, we can see that a linear relationship could be appropriate between the X and Y variables, but even more so when corruption is log transformed (right plot)
      4. Independent observations = assume from dataset
      5. Correlations = no multi-collinearity where correlation coefficients < .5 and VIF values are < 2
   3. **Results**
      1. ****The intercept is significantly different than 0, and this is the estimate of log(corruption) when all other variables are held constant
      2. Freedom has a significant association with corruption while GDP and generosity do not have significant relationships with corruption.
      3. As freedom increases by 1 unit, log(corruption) increases by 2.17
   4. **Model fit**
      1. Independent errors = yes passes, dwtest where p~.37, so we can’t reject the null hypothesis that there are independent errors.
      2. Homoscedasticity = yes passes, bptest where p~.0688, so we can’t reject the null hypothesis that there is no pattern in the residuals.
      3. Normal errors = not perfect, but errors appear pretty normal through the qqplot:
      4. Adjusted R2 = .1952, this model explains about 20% of the variation in the data, and the small p-value indicates a low percentage that a better model would occur by chance.
2. **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?** 
   1. Model: Corruption ~ Freedom\*Region
   2. Yes, there is a significant interaction effect for Freedom and Region Europe, where Region Europe:Freedom’s influence on Corruption is much stronger than Freedom’s single effect on Corruption. The slope (effect) of Freedom on Corruption is 1.59 where the slope (effect) of the Freedom:Europe interaction on Corruption is 4.57.



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.** 
   1. **GLM** – binomial, because dependent variable is binary
   2. **Assumptions**
      1. Lack of collinearity = passes where all independent variables correlated < .5 with one another and VIF values ~ 2 or under
      2. Observations are independent = assumed through dataset
      3. Broader exponential family = binomial
   3. **Results**
      1. Mean radius, mean texture, and mean smoothness all have a significant association with whether a breast cancer tumor is malignant or not (log/logit odds), based on significant p-values~0 for each variable. Additionally, as each of these variables increases, the log odds of malignancy also increases.
   4. **Model fit**
      1. Independent errors = no violates, dw test p-value = 3.616e-05, therefore we reject the null hypothesis that there is 0 autocorrelation.
      2. Normal errors = no violates this assumption through dramatic shape of qqplot, therefore p-values may be inaccurate. Note: this may be able to be fixed/improved by removing outlier points.
      3. Homoscedasticity = not assumed for GLMs, but interesting to note that this model’s errors pass the bptest where p=.069
      4. Residual deviance = deviance much lower for this model (751.44) compared to a null model (187.29), meaning that my model is a better fit than one just using the intercept.
2. **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.** 
   1. After standardizing the variables by mean and standard deviation, Radius seems to be the most important in explaining whether a tumor is malignant or not (log/logit odds), based on its lowest residual deviance and AIC score among univariate models using radius, texture, or smoothness’s impact on probability of malignancy. After radius, there is a large increase in the deviance value and AIC score to show the lesser importance of texture and then smoothness’s explanatory power. It’s worth noting that none of these variables show a significant increase in the likelihood of the model in the anova output, but there are still clear differences among the deviance values and significant ones in the AIC output.

