Audrey Pangallo

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**

So, given that there are two categorical variables here, I am going to run a chi-squared test for independence. My null hypothesis would be that there is no association between either gender and whether or not they mind children, and my alternate hypothesis would be that there is an association between gender and whether or not they tolerate unruly children. I am going to assume these observations are independent. And here is the readout for the contingency table:

No Somewhat Very

Female 91 193 158

Male 56 155 190

As you can see all the spaces are filled. Here is the readout for the test:

Pearson's Chi-squared test

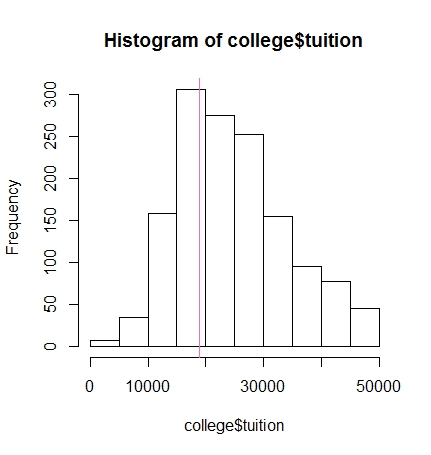
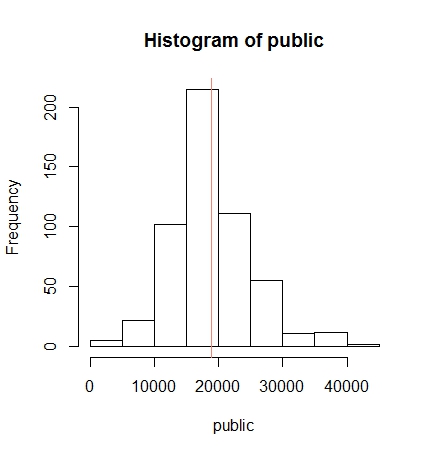
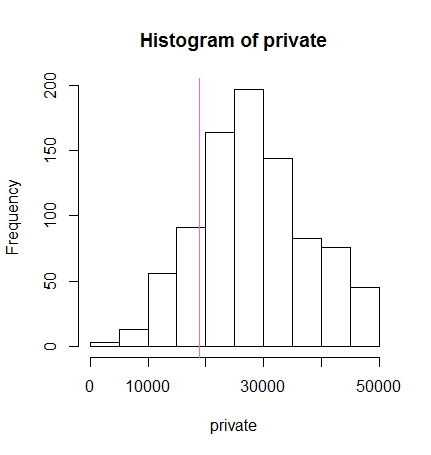
data: flytab

X-squared = 13.463, df = 2, p-value = 0.001193

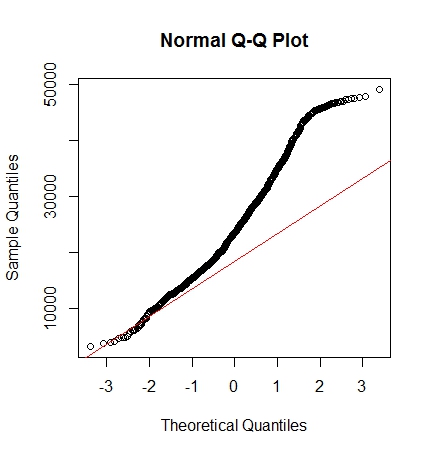
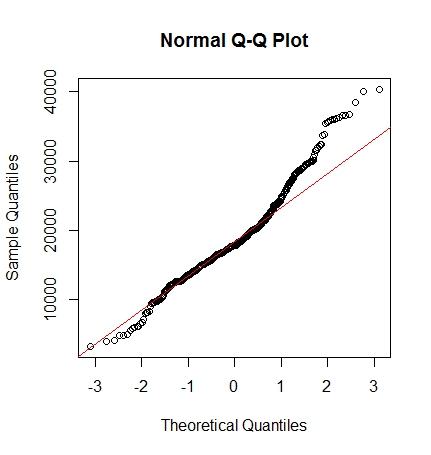
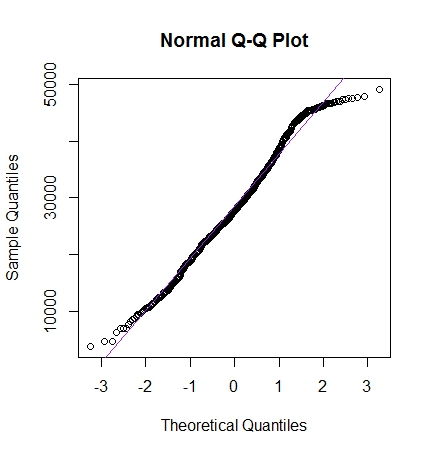
As you can see this test showed significance at the 0.05 level, (and, we could even say at the 0.01 level too… which is nice). Therefore I will reject the null and say there is an association between gender and tolerance of unruly children on planes.

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

I am going to run a t-test. The first thing I did here was do a subset of public tuition and private tuition and checked the histograms of those subsets. I also checked the histogram for just tuition in general.



When I first did this, I checked the normality of both public and private, BUT given what Dr. Jain said in class about checking the normality of the variable before you subset, I went back and checked that too. I’ve put all the qqplots in this answer just in case, because I care about my grade and want to make sure I didn’t leave anything out. Regardless, the qq plot looks terrible, and if you look at the qqplots, you’ll see that public tuition is what’s causing the problem. Which is odd, given how the histogram for private looked

. 

They go in this order: Private, public, and then all the tuition. I believe, from looking at the histogram, that public tuition is being skewed by outliers sitting on the outer end of the distribution ($40,000 tuition… must be UofM). I left this for a moment and checked my variance

data: public and private

F = 0.43469, num df = 534, denom df = 871, p-value

< 2.2e-16

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.3737381 0.5070016

sample estimates:

ratio of variances

0.4346851

If I am remembering how to read this correctly, equal variances would mean the ratio is equal to 1. Here it isn’t, so it looks like there’s unequal variances here. This left me with two options, run a t test in which the data isn’t quite normal, or run one for unequal variances, which if you look at the code, I decided to do both and compare the results.

data: public and private

t = -22.79, df = 1397.9, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-10258.878 -8632.749

sample estimates:

mean of x mean of y

18855.88 28301.69

Here is the Welch’s t test ^

Wilcoxon rank sum test with continuity correction

data: public and private

W = 92975, p-value < 2.2e-16

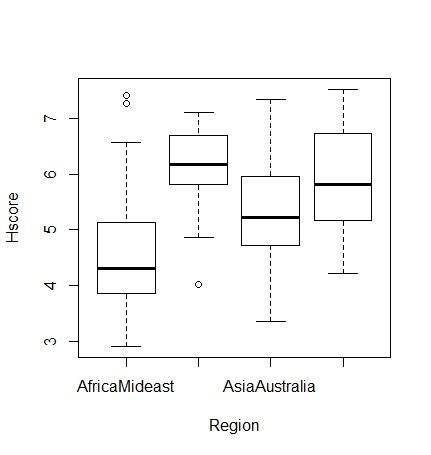
alternative hypothesis: true location shift is not equal to 0

Here is the wilcox.test ^

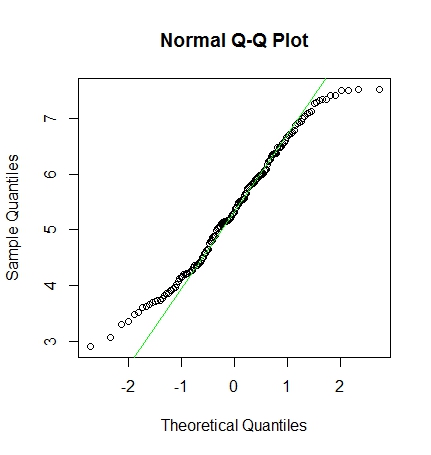
As I mention in the code, I’m not sure how you’ll grade this, but what I was looking for was anything in the readout that seemed off. The means are kind of far off, but with unequal variances, and from looking at the histograms, that seemed expected to me. Additionally, both the p values are exactly the same. So, I feel like if I had to just pick one, I would probably pick the welch’s t test. And from the readout conclude that we can reject the null and say there is a significant difference between public and private tuition. From the histograms, I would say private is more expensive.

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

With on continuous variable of interest and several categorical variables, I think I will run a one way anova for this question. My null hypothesis is that there is no difference between region and HScore, and my alternate is that there will be a difference between regions. The first thing I did was visualize my data with a boxplot:



All the regions are not labeled, I assume because R didn’t have enough room on the x axis. The next thing I did was check for normality of the Hscore:



Looks pretty good to me. The sample size is large enough here that I’m not worried about this being perfect. I am going to assume independent sampling, and equal variances. After this I ran the anova:

Df Sum Sq Mean Sq F value Pr(>F)

Region 3 64.37 21.456 23.62 1.28e-12 \*\*\*

Residuals 153 138.96 0.908

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

It looks like we have a significant difference. To check where that significant difference came from, I also ran a tukeyHSD:

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = Hscore ~ Region, data = happy)

$Region

diff lwr

AmericasCarribean-AfricaMideast 1.5397193 0.94745533

AsiaAustralia-AfricaMideast 0.7472543 0.14642988

Europe-AfricaMideast 1.3208593 0.84313961

AsiaAustralia-AmericasCarribean -0.7924650 -1.49989372

Europe-AmericasCarribean -0.2188600 -0.82522748

Europe-AsiaAustralia 0.5736050 -0.04112656

upr p adj

AmericasCarribean-AfricaMideast 2.13198329 0.0000000

AsiaAustralia-AfricaMideast 1.34807874 0.0081572

Europe-AfricaMideast 1.79857901 0.0000000

AsiaAustralia-AmericasCarribean -0.08503628 0.0214193

Europe-AmericasCarribean 0.38750748 0.7847082

Europe-AsiaAustralia 1.18833656 0.0768934

It looks like the significant differences are in the highlighted areas: Americas and Africa/Mideast, Asia and Africa/Mideast, Europe and Africa/Mideast (poor Africa and Middle East… they’re so much lower on the boxplot), and finally Asia/Australia and Americas.

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

I’m going to pick the variables GDP, Generosity, and Family for my model. Since corruption is also a continuous variable, this will be a linear regression. The first thing I will check is the correlation between my variables to make sure it’s ok.

Corruption GDP Generosity Family

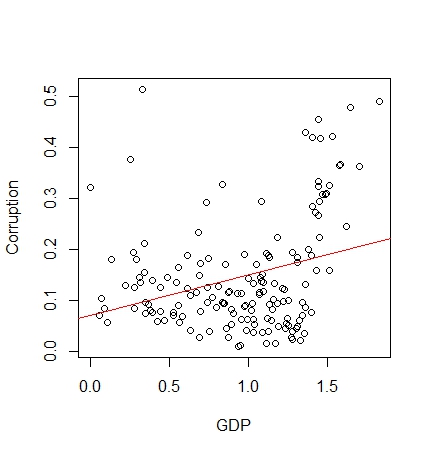
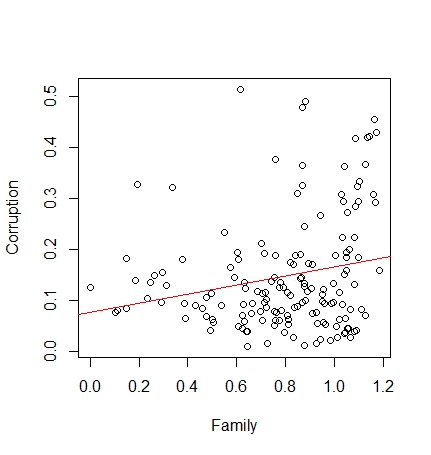
Corruption 1.0000000 0.29418478 0.30592986 0.21356094

GDP 0.2941848 1.00000000 -0.02553066 0.66953969

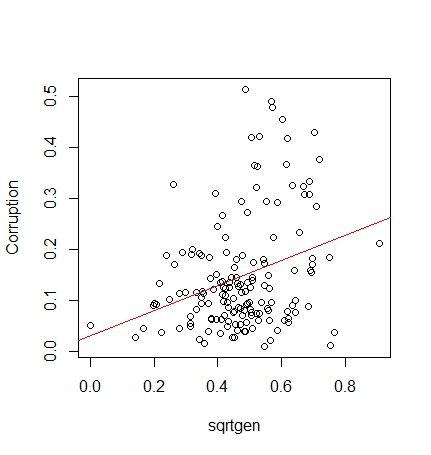
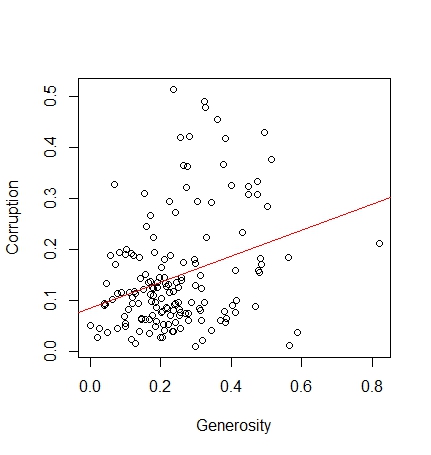
Generosity 0.3059299 -0.02553066 1.00000000 0.08962885

Family 0.2135609 0.66953969 0.08962885 1.00000000

Family and GDP seem to have a high correlation, however nothing was highly correlated with corruption. The next thing I did was plot my variables against Corruption to see if I things look mostly linear. Everything looked fine except for generosity, and so I transformed that with a square root and replotted it in order to make it look more linear. Here is GDP and Family:



And here is my before and after for generosity:



This looks like it was totally useless right now, but I found through experimentation trying to improve my model that the residuals look a bit better if I transform this data.

Call:

lm(formula = Corruption ~ GDP + sqrtgen + Family, data = happy)

Residuals:

Min 1Q Median 3Q Max

-0.20606 -0.06869 -0.00787 0.05617 0.41818

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.05066 0.03670 -1.380 0.169551

GDP 0.09045 0.02658 3.403 0.000851 \*\*\*

sqrtgen 0.26186 0.05849 4.477 1.47e-05 \*\*\*

Family -0.01478 0.04117 -0.359 0.720145

---

Signif. codes:

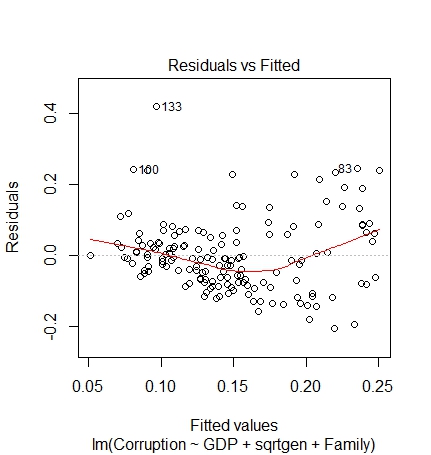
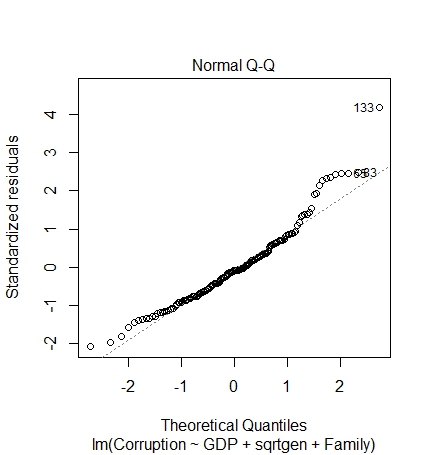
0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1007 on 153 degrees of freedom

Multiple R-squared: 0.1928, Adjusted R-squared: 0.177

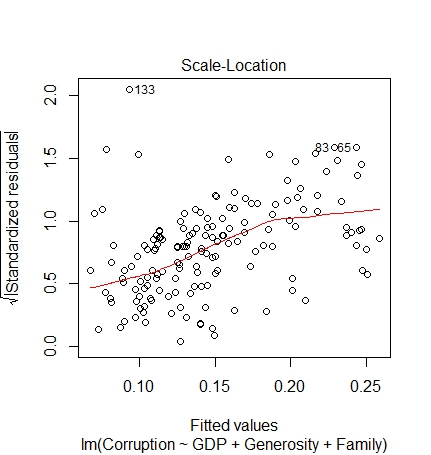
F-statistic: 12.18 on 3 and 153 DF, p-value: 3.423e-07

There are two variables that are significant, however the R squared was super low. Since the model itself was also showing a significant p- value, I decided to check my regression assumptions out before scrapping this.



So, as you can see, the normality is fine but the residuals look a little wonky. However, it’s much better than what they look like without the transformation.

One reason this model may not be as good is because it’s heteroscedastic, as depicted in the scale location graphic:



A homoscedastic plot should have a line that’s straighter. The BPtest confirmed this graph:

studentized Breusch-Pagan test

data: corrupt

BP = 8.3644, df = 3, p-value = 0.03905

I’d go back and try to transform my data, but I’ve run 6 models for this question, and I’ve had little luck getting a model that’ perfect. So, while recognizing that this isn’t the greatest model, I’m going to answer the following question as if it was.

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? happy**

Well, since this specified what I’m supposed to do, here’s the output:

Call:

lm(formula = Corruption ~ GDP \* Region, data = happy)

Residuals:

Min 1Q Median 3Q Max

-0.15786 -0.06441 -0.01891 0.04301 0.39168

Coefficients:

Estimate Std. Error t value

(Intercept) 0.10310 0.02539 4.061

GDP 0.06219 0.03024 2.056

RegionAmericasCarribean 0.01149 0.09389 0.122

RegionAsiaAustralia -0.09087 0.06347 -1.432

RegionEurope -0.21830 0.07436 -2.936

GDP:RegionAmericasCarribean -0.05910 0.09189 -0.643

GDP:RegionAsiaAustralia 0.08746 0.06378 1.371

GDP:RegionEurope 0.16558 0.06437 2.572

Pr(>|t|)

(Intercept) 7.87e-05 \*\*\*

GDP 0.04148 \*

RegionAmericasCarribean 0.90276

RegionAsiaAustralia 0.15429

RegionEurope 0.00386 \*\*

GDP:RegionAmericasCarribean 0.52106

GDP:RegionAsiaAustralia 0.17233

GDP:RegionEurope 0.01108 \*

---

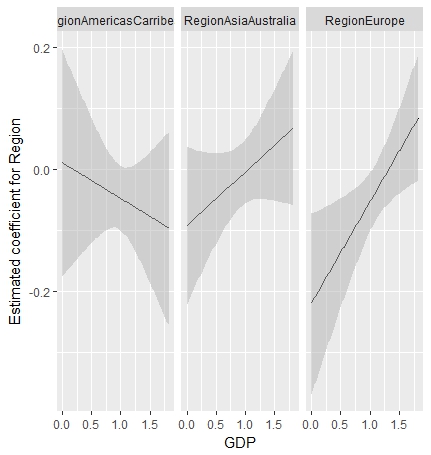
Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1036 on 149 degrees of freedom

Multiple R-squared: 0.1687, Adjusted R-squared: 0.1297

F-statistic: 4.321 on 7 and 149 DF, p-value: 0.0002209

It appears we have one significant interaction between region and GDP- Europe. 

And here is the interaction plot. The third one, the one that shows the significant term, is much steeper and heading in a positive direction, showing that when the European region and GDP are considered together, there is more likely to be corruption.

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. cancer**

Variables I am choosing: radius\_mean, texture\_mean, area\_mean. I know from being a human bio student and the child of doctors that texture is a significant factor in tumors because of how cancer cells are shaped, so I’m interested to see how this model goes. I went through and transformed some of my variables to make them more normal (radius\_mean and area\_mean had a much better looking qqplot when they were transformed)

And immediately the correlation plot has me questioning everything I have learned about cancer:

malignant radius\_mean texture\_mean area\_mean

malignant 1.0000000 0.7300285 0.4151853 0.7089838

radius\_mean 0.7300285 1.0000000 0.3237819 0.9873572

texture\_mean 0.4151853 0.3237819 1.0000000 0.3210857

area\_mean 0.7089838 0.9873572 0.3210857 1.0000000

Texture appears to be the ONLY variable not highly correlated here.

I made two types of models. A regular regression, and a glm, because malignancy is a binomial dataset. I ran a regular lm model first, took a look at the assumptions for regression using the plot( ), and then found that all the plots except the qqnorm looked weird. Then what I did was tried to run it again without transformed data. They STILL looked weird, and so I decided to try a GLM because I believe the odd plots are from the binomial data. The GLM output doesn’t look that much different from the lm, both models found texture\_mean to be significant (so my biology education hasn’t failed me!), here, I’ve put both outputs here to show you that, and then a qqplot from the lm model (the glm qq is also very normal looking)

Call:

lm(formula = malignant ~ lograd + texture\_mean + logar, data = cancer)

Residuals:

Min 1Q Median 3Q Max

-0.65456 -0.23925 -0.04174 0.22373 0.97568

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4.942557 1.011813 -4.885 1.35e-06 \*\*\*

lograd -1.256543 1.929012 -0.651 0.515

texture\_mean 0.022717 0.003295 6.894 1.45e-11 \*\*\*

logar 1.283633 0.951740 1.349 0.178

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3192 on 565 degrees of freedom

Multiple R-squared: 0.5673, Adjusted R-squared: 0.565

F-statistic: 246.9 on 3 and 565 DF, p-value: < 2.2e-16

GLM:

Call:

glm(formula = malignant ~ lograd + texture\_mean + logar, family = binomial,

data = cancer)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.09259 -0.37312 -0.09793 0.16405 3.16020

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -61.07913 14.65983 -4.166 3.09e-05 \*\*\*

lograd -17.20236 26.84916 -0.641 0.522

texture\_mean 0.20837 0.03811 5.468 4.55e-08 \*\*\*

logar 15.88023 13.27568 1.196 0.232

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

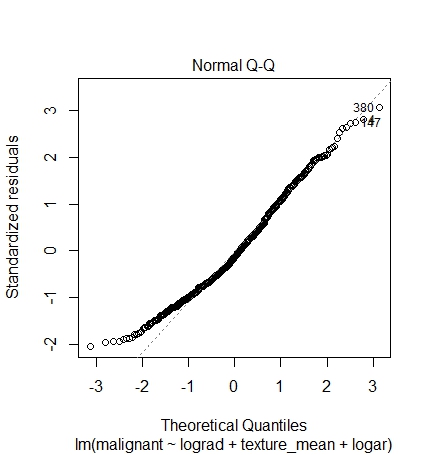
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 751.44 on 568 degrees of freedom

Residual deviance: 291.98 on 565 degrees of freedom

AIC: 299.98

Number of Fisher Scoring iterations: 7



As you can see, both models were finding significance in texture\_mean. I just have no way to check the assumptions on the lm model that makes sense. GLM’s allow you to relax some of the assumptions (homoscedasicity and error normality), so I feel like the GLM will better represent this dataset, although you can see that the things look pretty normal in the qqplot.

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.**

I’d answer this but I’m not sure how it’s different from the previous question… it’s nearly identical in wording.