**Yuchao Zhao Final: R Written yuchzhao**

Please answer the following questions by analyzing the associated datasets. For all tests, please:

* check whether the data meet the requirements/assumptions of the test you plan to run
* complete any transforms needed to make the data meet the required assumptions
* run the test
* interpret the results (do not include only the R output)
* check model fit in the case of linear regressions and/or glms
* if you have the option between running a linear model with a transformed y variable or a glm, choose the linear model with a transformed y variable. only run a glm when you have to.

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?**

**Two sample one tail T-test:**

t.test(male$unrulybaby,female$unrulybaby,alternative='greater',var.equal=TRUE)

In this analysis I change “unruly\_child” from factor to numeric, which allow

me to compare male and female by using T-test.

**Assumption:**

continuous dependent variable

assume sample is randomly selected

observations are independent both within and across the population

do not pass shapiro test for normality, but the sample size >30

have same variance (p-value = 0.4512)

**Result:**

Because P-value<0.05, I will accept the alternative hypothesis that true

differences in means is greater than 0. Which means there is a significant

association between gender and whether people think it’s rude to bring an

 Unruly child on the plane. And male tends to think its more rude.

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

**Welch two sample one tail T-test:**

**Variables:** tuition (dependent), public & private nonprofit (independant)

**Assumption:**  continuous dependent variable

assume sample is randomly selected

observations are independent both within and across the population

do not pass shapiro test for normality, but the sample size >30

do not have same variance (p-value < 2.2e-16)

**Result:** p-value < 2.2e-16, indicate we can reject the null hypothesis and accept

the alternative that true difference in means is less than 0, which means

the mean of tuition in public college is less than the tuition mean in private

 nonprofit college. Therefore, private non-profit type has a higher tuition.

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

**one way ANOVA test**

**Variables:**

Hscore (dependent) Region (independent)

**Assumptions check:**

1. samples assume to be independent

2. Hscore is not normally distributed (shapiro. test p-value = 0.01248), but the

samples size> 30 and each region has more than 15 observations, so we can

relax the normality assumption;

3.each population have same variance (Levene's Test for Homogeneity of

Variance p-value=0.5179)

**Result:**

P-value=1.28e- 12, significant enough to reject the null hypothesis which means

 there is significant difference in happiness by region.

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

**Multiple linear regression**

lm(sqrt(Corruption)~GDP+Generosity+Freedom, data=happy)

**Variables:**

Corruption (dependent) GDP&Generosity&Freedom (independent)

**Assumptions check:**

1. dependent variable is not normally distributed (shapiro. test p-value =

5.128e-11). Therefore, the dependent variable need to be transformed in order

to keep the normality of residuals. I will try “sqrt” transform firstly.

2. there is no significant correlation among independent variables.

VIF=1.189712 1.367952 1.89185

3. there is linear relationship between dependent and independent variables

****

**Model fit check**

residual is normally distributed (p-value = 0.7366)

there is no significant autocorrelation of errors (p-value = 0.1442)

not homoscedastic (p-value = 0.01311)

**Try log transformed model**

residual is not normally distributed (p-value = 0.0006382)

there is no significant autocorrelation of errors (p-value = 0.3701)

homoscedastic (p-value = 0.06878)

**Since each of the transformed model do not perfectly pass all fit test, I decide**

**to use the model which has the lowest AIC value, “sqrt” transformed model**

** has the lowest AIC.**

**Result:** Among the three factors, freedom is significantly associated with a

 country’s corruption levels.

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?**

lm(sqrt(Corruption)~Freedom+Freedom\*Region, data=happy)

**1. Model fit check**

residual is normally distributed (p-value = 0.9516)

there is no significant autocorrelation of errors (p-value = 0.607)

not homoscedastic (p-value = 0.01121)

**Try log transformed model**

residual is not normally distributed (p-value = 0.01481)

there is no significant autocorrelation of errors (p-value =0.5627)

homoscedastic (p-value = 0.2501)

2. **Since each of the transformed model do not perfectly pass all fit test, I**

**decide to use the model which has the lowest AIC value, “sqrt”**

 **transformed model has the lowest AIC.**

3. **Result:** The output indicate when freedom keep same, the region change from

Africa Mideast(intercept) to Europe will result in 0.18188 unit decrease

of corruption. The change into other two regions is not statistical

significant. The interaction result indicate the effect of freedom factor

significantly depends on the region effect when in “Europe”.



**6) 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.**

Because the dependent variable is (0 or 1), so I will use a **GLM model with the**

**family of binomial.**

glm(malignant~radius\_mean+texture\_mean+smoothness\_mean, data=cancer,

family=binomial(link="logit"))

**1.Variables:**

malignant (dependent variable”)

radius\_mean & texture\_mean & smoothness\_mean (independent variables)

**2. Assumptions:**

There is no significant correlation among independent variables.

VIF=1.1582 1.125114 1.037347

**3. Model fit check**

It can be seen from the model comparison that by including those three variables

increase the likelihood of the model and this increase is significant. It also has the

 lowest AIC value.

**4. Result:** the output indicate all three factors are significantly associated with

whether a breast cancer tumor is malignant or not. The increase of those

three variables will all increase the log odds of chance of malignant in

 different level.

**7.）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.**

To answer the relative importance of independent variables, I use the package called **relimp**. This is a functions to facilitate inference on the relative importance of predictors in a linear or generalized linear model. Considering my original model is a GLM, so this function is suitable. The relative importance is measured by the **ratio of effect standard deviations among each dependent variables**.

From the output it can been seen radius/texture=3.008 (radius important than texture), radius/smoothness=2.42 (radius important than smoothness), texture/smoothness=0.804 (smoothness is important than texture ). Therefore, **the most important factor in explaining whether a breast cancer tumor is malignant or not in this model is “radius”**. The secondary is “smoothness”.

