Alice Elliott

NRE 538 Final Exam Take-Home

4/18/2017

This exam is open book and open internet but you are NOT allowed to work with anyone else or ask anyone other than Meha or Oscar any questions about the exam. It is due at noon on Sunday, April 23.

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
* 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? **Flying**

*This is a chi-square test since we are comparing two categorical variables.*

*Therefore we do not need to test for anything except:*

*\*Whether the observations are independent (we can assume that they are)*

*\* Check that there will be no “structural zeroes” by making sure that none of the cells in the table have an expected value of less than 5:*

> chisq.test(tbl)$expected

unruly

gender No Somewhat Very

Female 77.07473 182.4626 182.4626

Male 69.92527 165.5374 165.5374

*All of the table cell values are greater than 5.*

*Since the data meet all the assumptions: Run the chi-square test with our alpha level at .05.*

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

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

*Null hypothesis: There is no association between gender and whether people think it’s rude to bring an unruly child onto a plane.*

*Alternate hypothesis: There is an association between gender and whether people think it’s rude to bring an unruly child onto a plane.*

*Since the p-value is under .05, we can safely reject the null hypothesis and say that there is an association.*

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

*This requires an unpaired two-sample t-test since we are comparing a continuous variable to a categorical variable.*

*Assumptions:*

*\*Sample is randomly selected from population*

*\*Independent observations (the same people don’t appear in multiple records)*

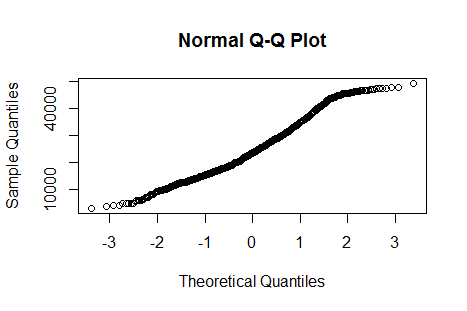
*\*Values of follow a normal distribution—check:*

> shapiro.test(college$tuition)

Shapiro-Wilk normality test

data: college$tuition

W = 0.97407, p-value = 3.193e-15



*Since the p value is less than .05, we must reject the null hypothesis that the data are normal.*

*We’ll try to transform it with log or square root (since it’s right skewed) but that doesn’t fix the problem since the p value still remains less than .05:*

> shapiro.test(sqrt(college$tuition))

Shapiro-Wilk normality test

data: sqrt(college$tuition)

W = 0.9915, p-value = 2.932e-07

> shapiro.test(log(college$tuition))

Shapiro-Wilk normality test

data: log(college$tuition)

W = 0.97079, p-value = 2.852e-16

*But the qqplot seems relatively normal, and since we have a large sample size (1407), we will work with it for now, with the understanding that the results are not as robust, and potentially biased, but the central limit theorem ensures that the distribution of the means of the sample will approximate normality due to the size of the sample.*

*\* Equal variance between two populations—check:*

> var.test(public$tuition,private$tuition)

F test to compare two variances

data: public$tuition and private$tuition

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

*Since the p-value is less than .05, this means that we must reject the null hypothesis that the variances are equal. This means that we must use the Welch’s t-test.*

*Let’s run the t-test with an alpha level of .05 after subsetting each type of college:*

> t.test(public$tuition,private$tuition,paired=FALSE,var.equal=FALSE)

Welch Two Sample t-test

data: public$tuition and private$tuition

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

*Null hypothesis: There is no significant difference in tuition type based on type of institution.*

*Alternate hypothesis: There is a significant difference in tuition based on type of institution.*

*Since the p-value from the test is less than .05, we can safely reject the null hypothesis that there is no significant difference in tuition.*

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

*This requires an ANOVA, because we are again comparing a categorical variable to a continuous variable, but this time there are multiple categories.*

*Assumptions:*

*\*Independent samples: probably true, since people were probably not sampled multiple times for this survey*

*\*Equal variances: check—*

> leveneTest(happy$Hscore,happy$Region)

Levene's Test for Homogeneity of Variance (center = median)

Df F value Pr(>F)

group 3 0.7605 0.5179

153

*Since the p-value is greater than .05, we can accept the null hypothesis that the variances are equal.*

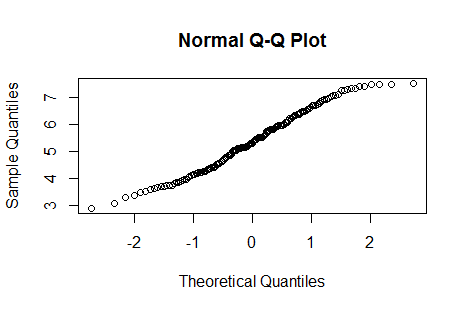
*\*Normally distributed populations—check:*

shapiro.test(happy$Hscore)

Shapiro-Wilk normality test

data: happy$Hscore

W = 0.97783, p-value = 0.01248

**

*Since the p value is less than .05, we must reject the null hypothesis that the data are normal.*

*No simple transforms seem to fix this (first square root, then log, then square):*

> shapiro.test(sqrt(happy$Hscore))

Shapiro-Wilk normality test

data: sqrt(happy$Hscore)

W = 0.9784, p-value = 0.0145

> shapiro.test(log(happy$Hscore))

Shapiro-Wilk normality test

data: log(happy$Hscore)

W = 0.97323, p-value = 0.003788

> shapiro.test(happy$Hscore^2)

Shapiro-Wilk normality test

data: happy$Hscore^2

W = 0.96191, p-value = 0.0002581

*But the qqplot seems relatively normal, and since we have a large sample size (157), we will work with it for now, with the understanding that the results are not as robust, and potentially biased, but the central limit theorem ensures that the distribution of the means of the sample will approximate normality due to the size of the sample.*

*Now let’s run the test:*

> anova = lm(Hscore~Region, data=happy)

> summary(anova)

Call:

lm(formula = Hscore ~ Region, data = happy)

Residuals:

Min 1Q Median 3Q Max

-2.1138 -0.6861 -0.1210 0.5650 2.8019

Coefficients:

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

(Intercept) 4.6021 0.1251 36.776 < 2e-16 \*\*\*

RegionAmericasCarribean 1.5397 0.2280 6.753 2.84e-10 \*\*\*

RegionAsiaAustralia 0.7473 0.2313 3.231 0.00151 \*\*

RegionEurope 1.3209 0.1839 7.182 2.81e-11 \*\*\*

---

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

Residual standard error: 0.953 on 153 degrees of freedom

Multiple R-squared: 0.3166, Adjusted R-squared: 0.3032

F-statistic: 23.62 on 3 and 153 DF, p-value: 1.279e-12

*Null hypothesis: There is no significant difference in happiness based on region.*

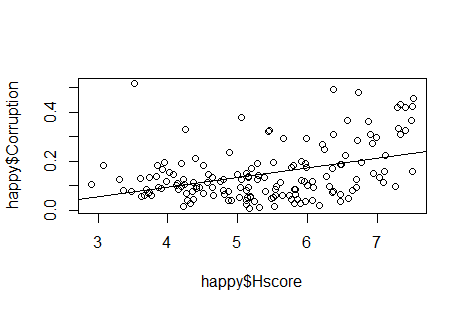
*Alternate hypothesis: There is a significant difference in happiness based on region.*

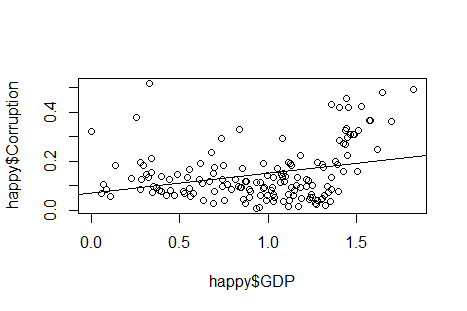
*Since the p-value for each category is less than .05, we can safely reject the null hypothesis that there is no significant difference in happiness.*

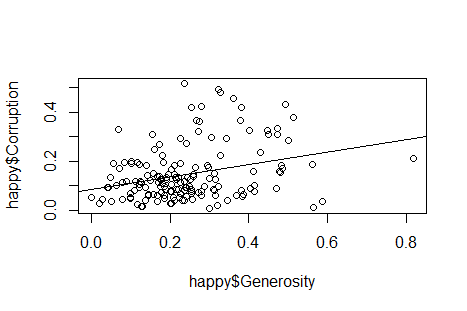
1. What factors are significantly associated with a country’s corruption levels (Corruption)? Choose three continuous independent variables to include in your model. **happy**

*We will build a linear model to answer this question. We will include the factors Freedom, GDP, and Generosity.*

*First, we will check the assumption that there is a linear relationship between the variables.*

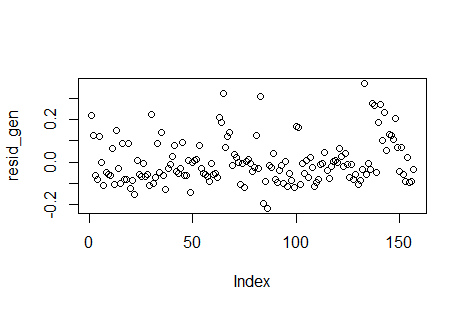
**



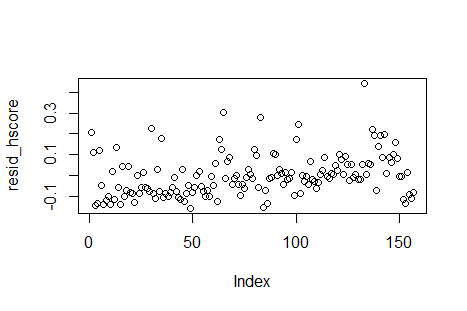


*These look pretty linear. Let’s check for homoscedasticity of the errors.*

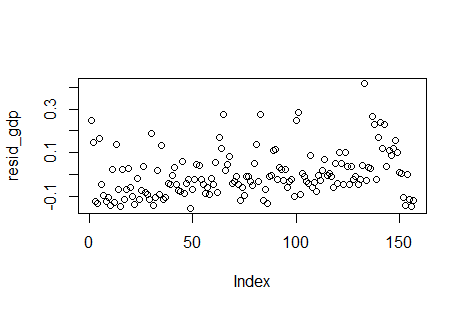
*Generosity:*

**

*Hscore:*

**

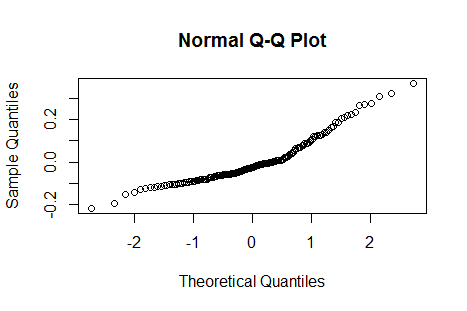
*GDP:*

**

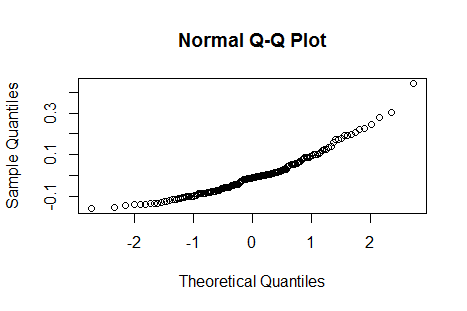
*Most of the variables look homoscedastic.*

*Let’s check for normality of the errors:*

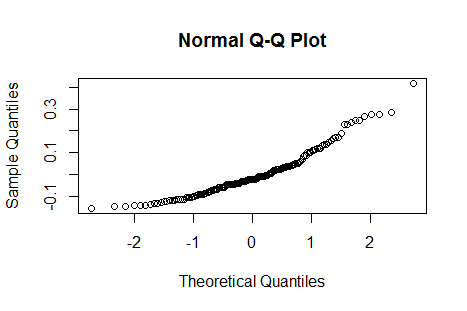
*Generosity:*

**

*Hscore:*

**

*GDP:*

**

*Most of the error distributions look normal, except for GDP, which could be slightly non-normal. We will proceed for now under the assumption that they are all normal, with the understanding that our results relating to GDP could potentially be slightly biased.*

*\*Check for multicollinearity:*

> VIF(lm(Corruption~Generosity+Hscore+GDP,data=happy))

[1] 1.286397

*It’s less than 10, so we’re good.*

*Now we will run our model:*

> mod = lm(happy$Corruption~happy$Generosity+happy$Hscore+happy$GDP)

> summary(mod)

Call:

lm(formula = happy$Corruption ~ happy$Generosity + happy$Hscore +

happy$GDP)

Residuals:

Min 1Q Median 3Q Max

-0.21877 -0.07438 -0.00977 0.04960 0.43580

Coefficients:

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

(Intercept) -0.08716 0.04250 -2.051 0.041984 \*

happy$Generosity 0.21164 0.06183 3.423 0.000795 \*\*\*

happy$Hscore 0.03223 0.01182 2.727 0.007140 \*\*

happy$GDP 0.01044 0.03231 0.323 0.747070

---

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

Residual standard error: 0.09886 on 153 degrees of freedom

Multiple R-squared: 0.2226, Adjusted R-squared: 0.2074

F-statistic: 14.61 on 3 and 153 DF, p-value: 2.053e-08

*Generosity, GDP, and Hscore are positively associated with corruption. For every 1 increase in generosity, corruption increases by .21. For every 1 increase in Hscore, corruption increases by .03. For every 1 increase in GDP, corruption increases by .01. Only Generosity and Hscore have a significant association with 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? **Happy**

*Let’s choose Generosity.*

> mod2 = lm(Corruption~Generosity\*Region, data=happy)

> summary(mod2)

Call:

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

Residuals:

Min 1Q Median 3Q Max

-0.18452 -0.06900 -0.01507 0.03649 0.36336

Coefficients:

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

(Intercept) 0.09182 0.03192 2.876 0.00461 \*\*

Generosity 0.25491 0.13267 1.921 0.05659 .

RegionAmericasCarribean 0.02367 0.06002 0.394 0.69388

RegionAsiaAustralia 0.01826 0.06061 0.301 0.76366

RegionEurope -0.04622 0.04356 -1.061 0.29031

Generosity:RegionAmericasCarribean -0.24457 0.25173 -0.972 0.33284

Generosity:RegionAsiaAustralia -0.12683 0.18502 -0.685 0.49410

Generosity:RegionEurope 0.24336 0.17433 1.396 0.16481

---

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

Residual standard error: 0.1046 on 149 degrees of freedom

Multiple R-squared: 0.1528, Adjusted R-squared: 0.113

F-statistic: 3.84 on 7 and 149 DF, p-value: 0.0007228

*Only the intercept is marked as significant. This means that the influence of Generosity on Corruption does not vary by region.*

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**

*This is a binomial distribution, because the outcome can only be 0 (non-malignant) or 1 (malignant).*

*Therefore we will use a generalized linear model that uses a binomial distribution and the logit link function because it would be difficult to make the data fit the assumptions for a normal linear regression.*

*We will test the following assumptions:*

*\*Independent observations: we will assume this is true since each of these tumors are (hopefully) from a different patient*

*\*Normal distribution of transformed dependent variable: this is difficult to test, so we will assume this will be true if we use the correct distribution and link function*

*\*Dependent variable is from the binomial family—this is true since the only possible outcomes are 0 or 1*

*We will also test to make sure none of the independent variables are highly correlated with each other:*

> VIF(lm(malignant~radius\_mean+texture\_mean+smoothness\_mean,data=cancer))

[1] 2.727459

*Since the VIF is under 10, we will proceed.*

> mod3 = glm(malignant~radius\_mean+texture\_mean+smoothness\_mean,data=cancer,family='binomial')

> summary(mod3)

Call:

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

family = "binomial", data = cancer)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.19102 -0.19403 -0.03799 0.04025 2.92583

Coefficients:

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

(Intercept) -42.01941 4.45943 -9.423 < 2e-16 \*\*\*

radius\_mean 1.39699 0.15403 9.069 < 2e-16 \*\*\*

texture\_mean 0.38056 0.05711 6.663 2.68e-11 \*\*\*

smoothness\_mean 144.67423 19.04687 7.596 3.06e-14 \*\*\*

---

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: 187.29 on 565 degrees of freedom

AIC: 195.29

Number of Fisher Scoring iterations: 8

*\*For every increase in the mean radius, the log odds of the tumor being malignant increases by 1.4*

*\*For every increase in the mean texture, the log odds of the tumor being malignant increases by .38*

*\*For every increase in the mean smoothness, the log odds of the tumor being malignant increases by 144.7.*

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.**

*We will use AIC values to compare each independent variable to see which one is most important.*

> mod\_radius = glm(malignant~radius\_mean, data=cancer, family='binomial')

> mod\_texture = glm(malignant~texture\_mean, data=cancer, family='binomial')

> mod\_smooth = glm(malignant~smoothness\_mean, data=cancer, family='binomial')

>

> AIC(mod\_radius,mod\_texture,mod\_smooth)

df AIC

mod\_radius 2 334.0108

mod\_texture 2 650.5191

mod\_smooth 2 677.9485

*Since “mod\_radius” had the lowest AIC value, we can conclude that the “radius\_mean” independent variable is the most important in explaining whether a breast cancer tumor is malignant.*