STAT R 501 Final Project

Thomas Wright

Do rainy days make you drown your sorrows?

Introduction

Many things can drive an individual to drink. Stress, elation, apathy or moments of crisis – but do rainy days cause people to turn to the bottle?

Quantifying the relationship between drinking and rain is of paramount importance to people living in rainy climates (such as Seattle) or those looking to pick a booze-themed vacation. An entrepreneur looking to open a bar could benefit from knowing whether rainfall impacts his potential clientele's consumption habits or whether a sunny local would be a better destination. Marketing campaigns can leverage the knowledge to maximize ROI – either by choosing climates to market in or using weather patterns to plan initiatives around.

This inferential study will test the significance of the relationship between rainfall and total alcohol consumed – whether it is positive, negative or non-existent. Furthermore, three secondary analyses will be conducted: the linearity of rainfall, normality of errors and the impact of confounding variables.

Data

The dataset was constructed from three sources: 538.com, the UN and the World Bank – with 538 supplying the drinking data. It contains 192 observations (which are countries) with 6 explanatory variables and a numeric response.

Variables

- Country
  - An ID variable
- total_litres_of_pure_alcohol
  - Total liters of alcohol consumed per person in a year
- Rainfall
  - Average annual rainfall across the country in Inches
- unemployment rate
- GDP
- Literacy Rate
- Temp (C)
- Region
  - Factor for what region of the world the country is in
- Island
  - Binary for whether it is an island or not

Methodology

The study starts with a visual exploration of the potential explanatory variables before constructing a simple model – using it to check for linearity and preliminary heteroskedascity.

Additional variables are then added to the model to check for confounding variables. If the p-values and estimates from the simple model change, then the added variables are likely confounding factors and should be included in the final model.

Finally the study constructs a full model, with all available explanatory variables and their interaction terms.

Simple Model:

The simple model takes the form of:

$Y = \boldsymbol{\alpha + \beta X + \varepsilon}$

- Y=total alcohol consumed
- X = total annual rainfall (mm)

Sensitivity analysis:

To begin the sensitivity analysis, a quadratic model is constructed. The two models are compared to see if the fit is improved by including the new term.

Testing for Linearity

- $Y = \boldsymbol{\alpha + \beta X + \varepsilon}$
- $Y = \boldsymbol{\alpha + \beta_1 X + \beta_2 X^2 + \varepsilon}$

Testing and correcting for heteroskedascity

To test for heteroskedascity, first the study inspects the residual plots of the fitted models. If it appears that there may be unequal variance present, a Breusch Pagan test is employed. If the p-value exceeds a 5% threshold, then the null hypothesis of homoskedascity is rejected.

Correcting for heteroskedascity

There are two options for correcting the unequal variance:

1. Obtaining estimates with White-Huber standard errors

   The White-Huber S.E. are robust to heteroskedascity and can be used to obtain un-biased Beta estimates.

2. Modeling with a Poisson generalized linear model (glm).

   A Poisson distribution sets its variances equal to its mean – allowing it to increase with dependent variable. However, in this case the variance appears to be equal to the inverse of the mean (or something close to it).

   As such, a quasi-Poisson distribution of errors must be used. In R, this is obtained with the argument of family = Poisson (link ="log") in the glm function.

Confounding variables and interaction terms:

Interaction terms are needed in models where there may be confounding factors. For example, if people are more likely to drink with it rains, and it more likely to rain when it is hot, then temperature must be included as an interaction term – to isolate the impact of rainfall.

As such, the study constructed two models:

- Temperature as another explanatory variable
- Temperature as an additional interaction term with rainfall

If the estimates and p-values from the simple model change, it suggests that there are confounding variables that must be included to get an accurate assessment of rainfall's impact

Full Model:

Finally the study turns to a full model, with all explanatory variables[1] and their interaction terms included. Significant terms are reported, while only rainfall is examined in detail (as it is the main inferential goal of the study).

$$Y = \boldsymbol{\alpha} + \boldsymbol{\beta}_i \mathbf{X}_i + \boldsymbol{\varepsilon}$$

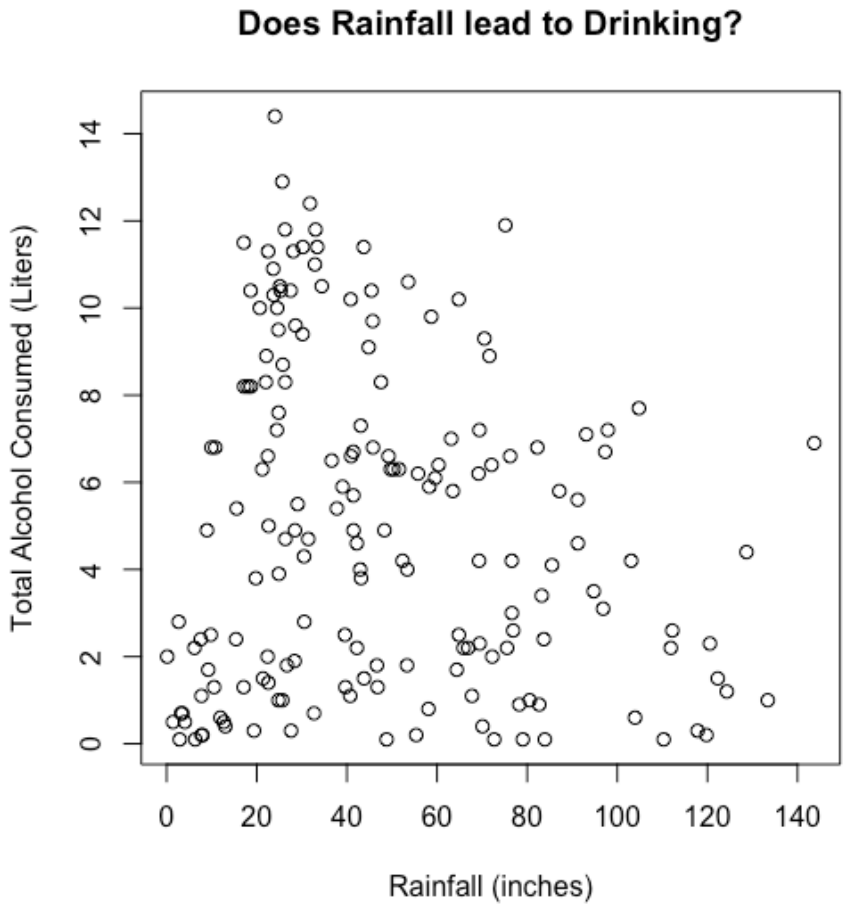
[1] Region was excluded from the model as it contained too many levels and obscured the interpretability of the model. The study suggests using the lasso method (outside the

- Where i = variable from the list of:
    - unemployment rate
    - Population
    - GDP
    - Literacy
    - Temp
    - Rainfall
    - Rainfall^2

Results

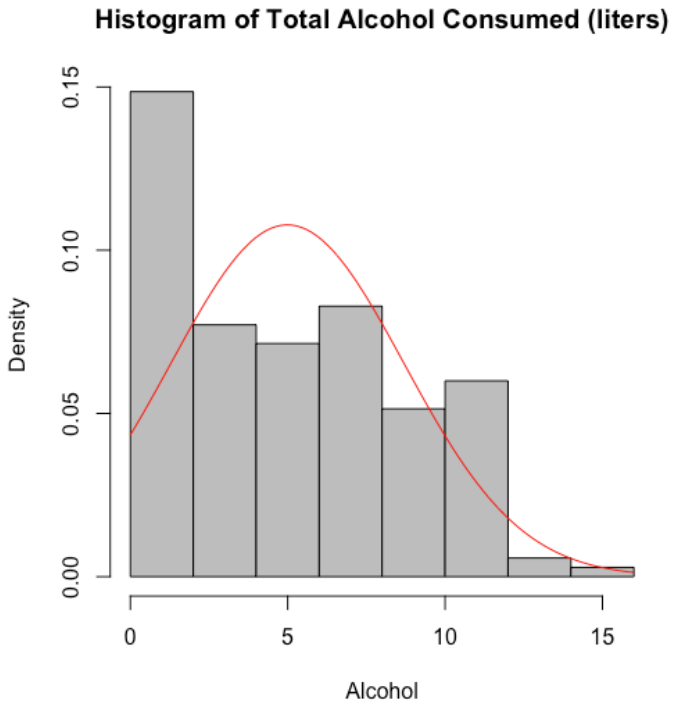
Visualizations

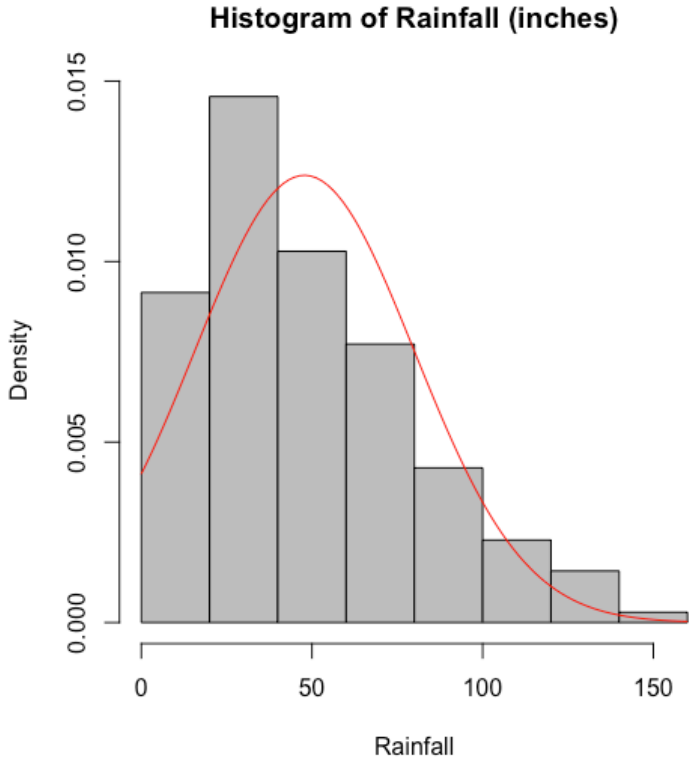
Looking at the main variables of interest, Rainfall and Total.Alcohol consumed:



It appears that there is an initial increase in drinking with rainfall, which peaks between 20-40 inches, and begins to decrease after that. This plot hints that a quadratic term may be needed to adequately model the relationship.

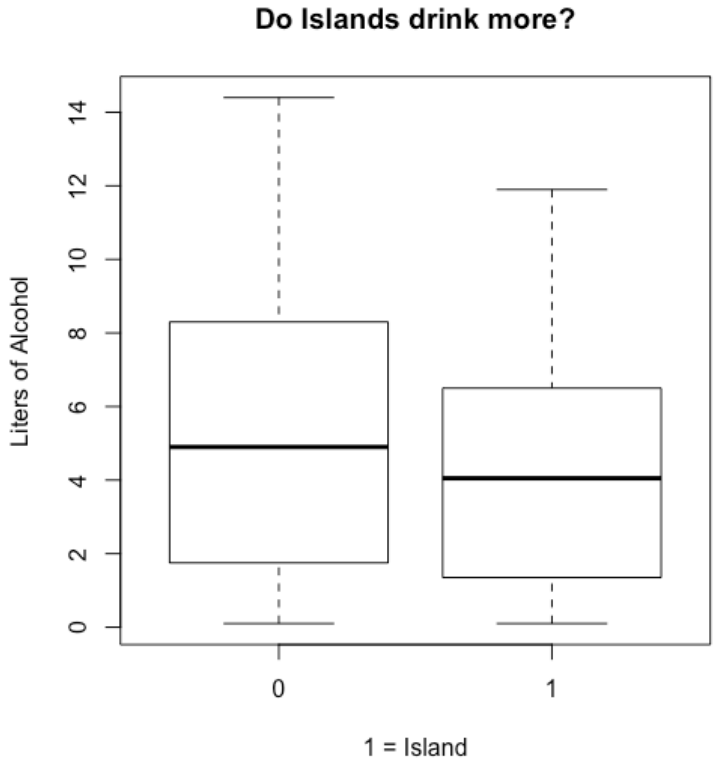
Turning to the histograms of the two variables reveals that Total.Alcohol is not distributed normally, while rainfall is roughly normal. However, with a log transformation not improving the distribution and the large sample size (n = 196), the study will rely on the Central Limit Theorem to ensure the assumption of normality.


Histogram of Total Alcohol Consumed (liters)
Density
0.00
0.05
0.10
0.15
0
5
10
15
Alcohol


Histogram of Rainfall (inches)
Density
0.000
0.005
0.010
0.015
0
50
100
150
Rainfall

Boxplots of categorical variables

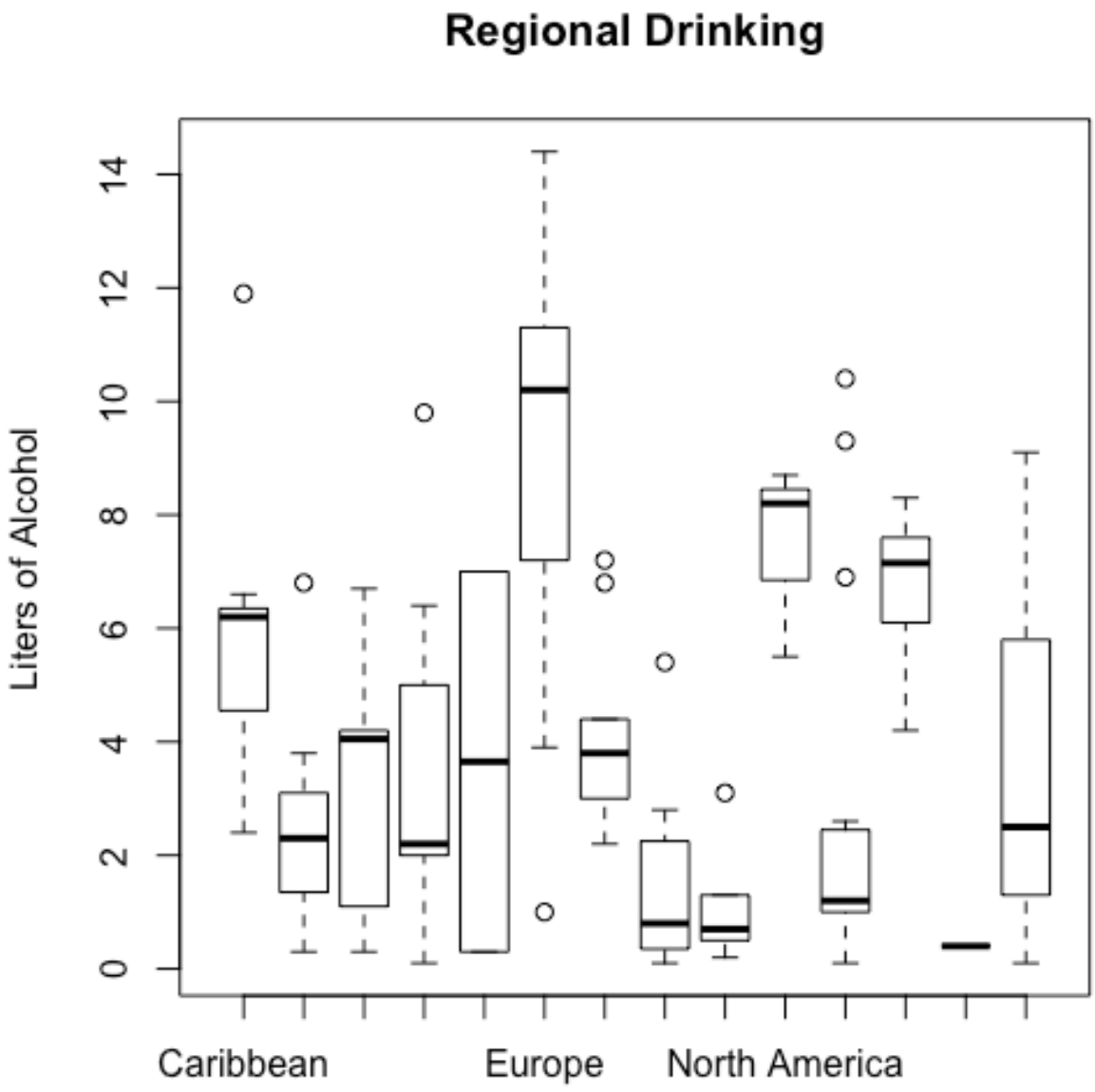
Islands are often popular destinations for epicurean vacations – drawing on images of poolside bars and fancy little umbrellas in coconuts. But do Islanders actually drink more?



The response variable is total liters of alcohol consumed per person – without taking into account tourism. This should increase the per person measurement for countries with large amounts of visitors (as the response is total alcohol/ population).

However, it appears that Islanders actually drink less.

Taking a look at consumption by region reveals that Europe and North America imbibe the most. The eastern APAC (Asian Pacific Region) has the highest variance – perhaps due to the differing cultures that span from Oceania to the Philippines.



Statistical Analysis

Simple Model

The results of the simple model – with Rainfall as the only explanatory variable – shows a significant, negative relationship between the two variables. However, with the adjusted R^2 of .01476, it is clear that this model has essentially no inferential or predictive power.

```
summary(Simple)

##
## Call:
## lm(formula = Project$Total.Alcohol ~ Project$Rainfall)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6288 -3.4315 -0.4045  2.7252  9.0183
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.776243   0.498205  11.594   <2e-16 ***
## Project$Rainfall -0.016427   0.008651  -1.899   0.0592 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.674 on 173 degrees of freedom
## Multiple R-squared:  0.02042,    Adjusted R-squared:  0.01476
## F-statistic: 3.606 on 1 and 173 DF,  p-value: 0.05924
```

Linearity

Adding a quadratic term to the model improves the fit, increasing the adjusted R^2 (which remains abysmally low). While the relationship turns positive for low rainfall values, after ~45 inches of rainfall the relationship switches back to negative – depressing the amount of alcohol consumed.

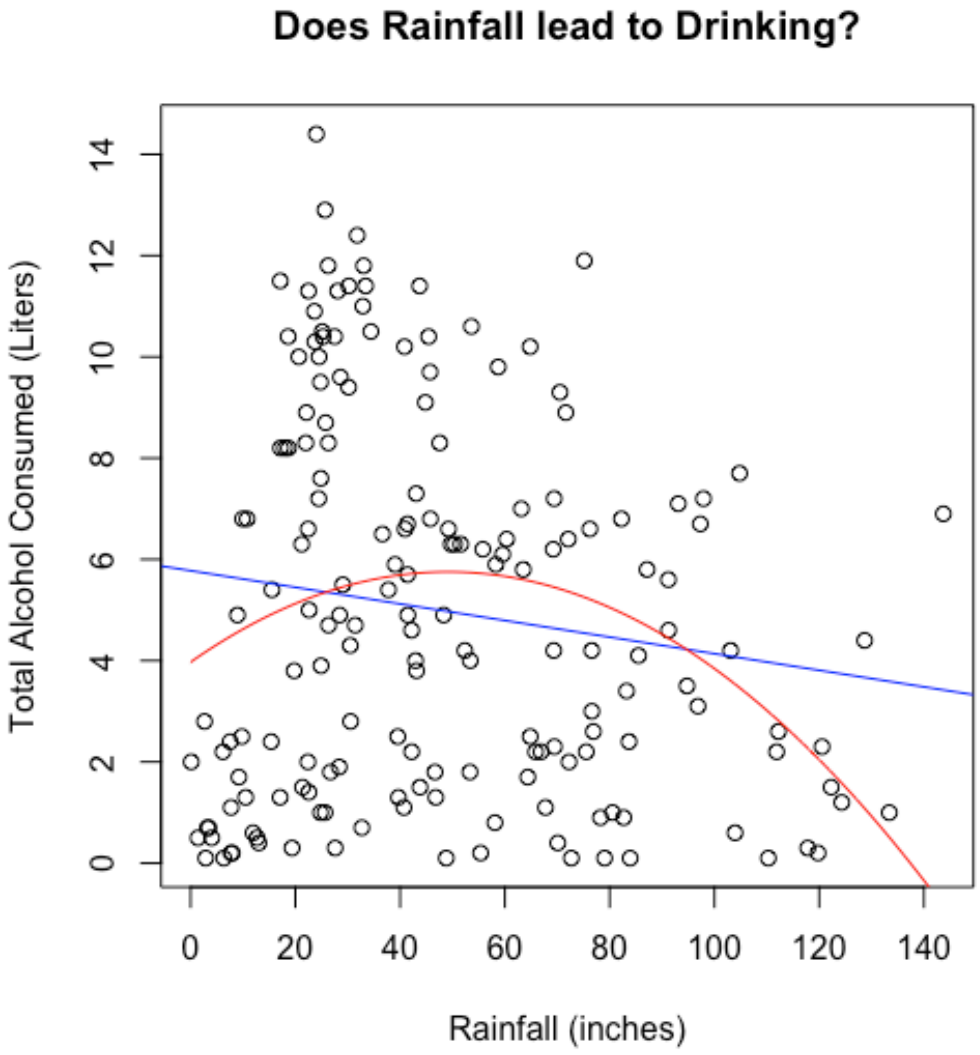
```
summary(quadratic.model)

##
## Call:
## lm(formula = Project$Total.Alcohol ~ Project$Rainfall + Project$Rainfall.2
)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.653 -3.334 -0.258  3.011  9.113
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.9705254  0.7398855   5.366 2.57e-07 ***
## Project$Rainfall    0.0725630  0.0287908   2.520  0.01263 *
## Project$Rainfall.2 -0.0007384  0.0002284  -3.232  0.00147 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.578 on 172 degrees of freedom
```

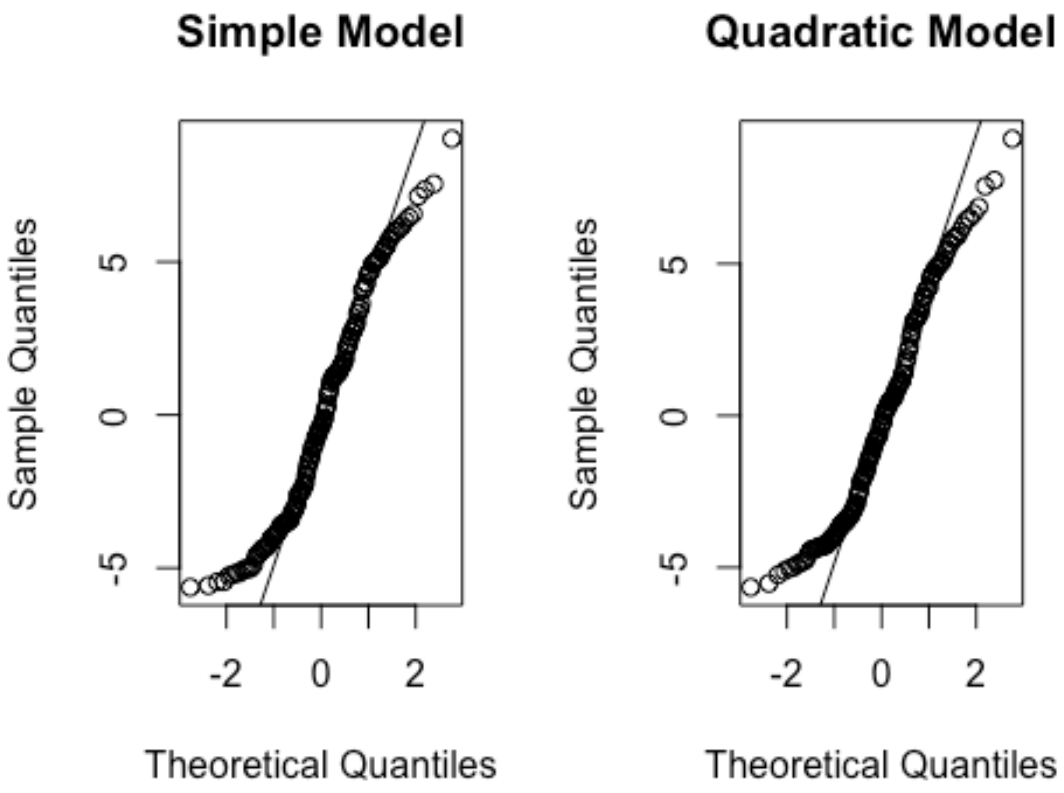
```
## Multiple R-squared:  0.07651,    Adjusted R-squared:  0.06578
## F-statistic: 7.125 on 2 and 172 DF,  p-value: 0.001064
```

Diagnostic Plots

Plotting the two regression firm confirms that the fit has improved, with the positive relationship at low rainfall levels now captured in the model.



The QQ plots of both models confirm that the sample size was large enough to overcome the abnormality of the distributions of the component variables.



Heteroskedascity

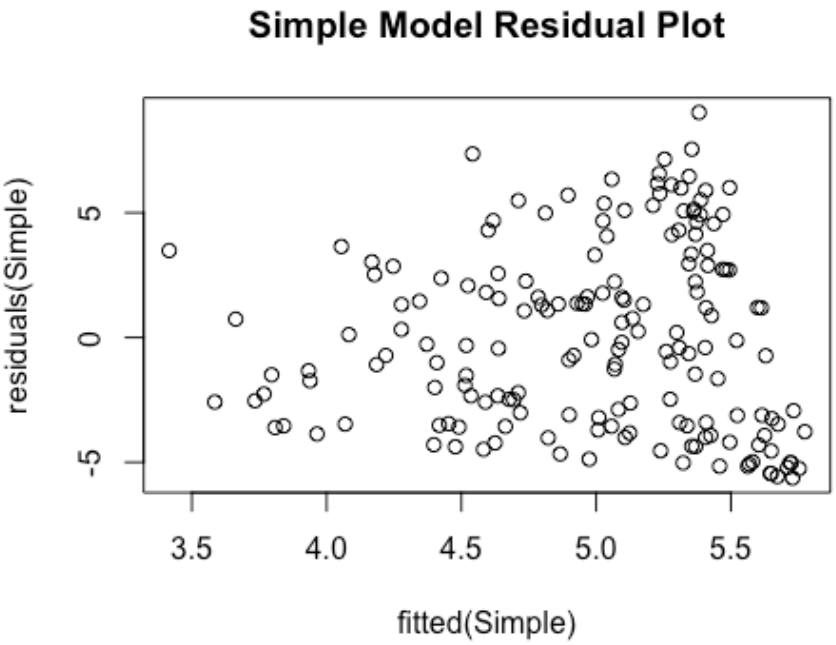
Looking at the residual plots shows that the models may be suffering from Heteroskedascity. The Breusch-Pagan tests reveal that while the simple model suffers from heteroskedascity (p-value of ~0) while the quadratic model does not (p value of .07).

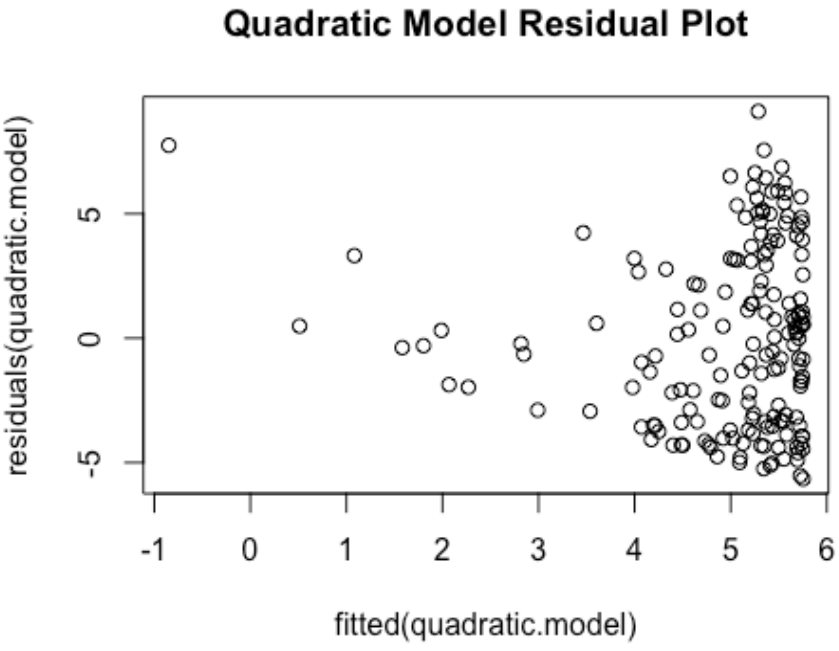
```
bptest(Simple)

##
##  studentized Breusch-Pagan test
##
## data:  Simple
## BP = 16.958, df = 1, p-value = 3.822e-05

bptest(quadratic.model)

##
##  studentized Breusch-Pagan test
##
## data:  quadratic.model
## BP = 5.3334, df = 2, p-value = 0.06948
```





Correcting the model

Using Huber-White Standard Errors[2] adjusts the estimates to:

```
coeftest(Simple,vcov=hccm(Simple))

##
## t test of coefficients:
##
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.7762434  0.5479718 10.5411  < 2e-16 ***
## Project$Rainfall -0.0164268  0.0081468 -2.0164  0.04531 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Alternatively, the study also employed a generalized linear model with a Poisson distribution (coded as a quasi-Poisson that selected a dispersion parameter of 1).

```
summary(Simple.P)

##
## Call:
## glm(formula = Project$Total.Alcohol ~ Project$Rainfall, family = poisson(l
ink = log))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2509  -1.6959  -0.1772   1.1226   3.2094
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.766094   0.059508   29.68  < 2e-16 ***
## Project$Rainfall -0.003439   0.001102   -3.12  0.00181 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 528.55  on 174  degrees of freedom
## Residual deviance: 518.51  on 173  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 5
```

While the models produce different estimates (as they use different estimation techniques), they both show that the significant and negative relationship remains.

---

[2] A simple regression model is equivalent to a one-sample t-test. This allows the study to use the corrected SE, which are not an argument option in lm or glms.

Confounding Variables

The coefficient on Temperature has a large magnitude, particularly relative to rainfall – with a 5 degree temperature swing resulting in a difference of one liter of alcohol consumed per person annually. With variables that have high significance and impact on the response variable, there may be a confounding presence – with some of the impact of rainfall being captured in temperature (as those places where it rains more are more likely to be warmer). As such, an interaction term should be included in the model.

```
p2 <- glm(Total.Alcohol~Rainfall+Rainfall.2+Temp,data=Project)
summary(p2)
```

```
##
## Call:
## glm(formula = Total.Alcohol ~ Rainfall + Rainfall.2 + Temp, data = Project
)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -9.0087  -2.3364  -0.1233   2.2765   7.5292
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.9154568  0.7620608   9.075 2.67e-16 ***
## Rainfall     0.0876527  0.0252756   3.468 0.000664 ***
## Rainfall.2  -0.0006423  0.0002003  -3.207 0.001603 **
## Temp        -0.2201689  0.0300560  -7.325 8.99e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 9.800831)
##
##     Null deviance: 2384.3  on 174  degrees of freedom
## Residual deviance: 1675.9  on 171  degrees of freedom
## AIC: 902.01
##
## Number of Fisher Scoring iterations: 2
```

Including the interaction term not only reduces the magnitude of temperature, but also the significance – with the variable no longer having a relationship with alcohol consumption. However, the interaction term is highly significant – still capturing the impact of temperature.

```
p1 <- glm(Total.Alcohol~Rainfall*Temp*Rainfall.2,data=Project)
summary(p1)
```

```
##
## Call:
## glm(formula = Total.Alcohol ~ Rainfall * Temp * Rainfall.2, data = Project
)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -8.923  -2.256  -0.036   2.320   8.033
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -2.988e+00  3.215e+00  -0.930 0.353943
## Rainfall                  8.441e-01  2.518e-01   3.353 0.000989 ***
## Temp                      1.419e-01  1.362e-01   1.042 0.298949
## Rainfall.2               -1.600e-02  5.867e-03  -2.727 0.007073 **
## Rainfall:Temp            -2.600e-02  1.032e-02  -2.519 0.012712 *
## Rainfall:Rainfall.2       8.405e-05  4.218e-05   1.993 0.047929 *
## Temp:Rainfall.2           5.225e-04  2.315e-04   2.257 0.025321 *
## Rainfall:Temp:Rainfall.2 -2.861e-06  1.625e-06  -1.761 0.080025 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 9.188807)
##
##     Null deviance: 2384.3  on 174  degrees of freedom
## Residual deviance: 1534.5  on 167  degrees of freedom
## AIC: 894.59
##
## Number of Fisher Scoring iterations: 2
```

Full Model

For the final model, the study included all of the explanatory variables (barring region as discussed above).

Rainfall remained significant, along with literacy, temperature and a variety of interaction terms.

```
summary(Full.Int)
```

```
##
## Call:
## glm(formula = Total.Alcohol ~ (.)^2, data = Project2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.7957  -1.5545   0.0765   1.8281   7.6916
##
## Coefficients:
```

```
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -4.030e+01  1.088e+01  -3.706 0.000298 ***
## Rainfall                         5.815e-01  2.245e-01   2.590 0.010564 *
## Rainfall.2                      -6.513e-03  2.257e-03  -2.886 0.004498 **
## GDP                             -2.635e-02  3.519e-02  -0.749 0.455181
## Literacy                         4.715e-01  1.093e-01   4.312 2.97e-05 ***
## Temp                             1.340e+00  3.340e-01   4.013 9.56e-05 ***
## Island                          -5.892e+00  5.028e+00  -1.172 0.243105
## unemployment.rate2              -2.614e+00  2.222e+00  -1.176 0.241369
## Rainfall:Rainfall.2              1.710e-05  6.608e-06   2.588 0.010622 *
## Rainfall:GDP                     1.130e-03  5.427e-04   2.082 0.039078 *
## Rainfall:Literacy               -1.729e-03  1.461e-03  -1.183 0.238728
## Rainfall:Temp                   -1.290e-02  6.333e-03  -2.036 0.043546 *
## Rainfall:Island                 -8.421e-03  6.967e-02  -0.121 0.903968
## Rainfall:unemployment.rate2      8.136e-03  3.370e-02   0.241 0.809591
## Rainfall.2:GDP                  -7.271e-06  4.462e-06  -1.630 0.105306
## Rainfall.2:Literacy              8.822e-06  1.104e-05   0.799 0.425549
## Rainfall.2:Temp                  1.106e-04  7.703e-05   1.436 0.153001
## Rainfall.2:Island               -3.795e-05  5.337e-04  -0.071 0.943409
## Rainfall.2:unemployment.rate2 -8.615e-05  2.731e-04  -0.315 0.752846
## GDP:Literacy                    -4.452e-04  3.113e-04  -1.430 0.154859
## GDP:Temp                        -5.592e-05  7.739e-04  -0.072 0.942496
## GDP:Island                      -3.488e-03  1.201e-02  -0.290 0.771933
## GDP:unemployment.rate2           1.763e-02  6.114e-03   2.883 0.004529 **
## Literacy:Temp                   -1.252e-02  3.296e-03  -3.797 0.000214 ***
## Literacy:Island                  3.967e-02  4.037e-02   0.983 0.327331
## Literacy:unemployment.rate2      3.950e-03  1.972e-02   0.200 0.841523
## Temp:Island                      4.342e-02  8.952e-02   0.485 0.628393
## Temp:unemployment.rate2          4.182e-02  3.743e-02   1.117 0.265628
## Island:unemployment.rate2        6.906e-01  8.724e-01   0.792 0.429885
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 7.226287)
##
##     Null deviance: 2384.3  on 174  degrees of freedom
## Residual deviance: 1055.0  on 146  degrees of freedom
## AIC: 871.02
##
## Number of Fisher Scoring iterations: 2
```

Discussion

Rainfall has a significant, quadratic relationship with the liters of total alcohol consumed per person in a country. At low levels of rainfall, the relationship is positive before turning negative at high levels of rainfall. This could be a factor of particularly dry places leading to dehydration, which alcohol worsens.

Surprisingly, GDP does not have an impact on alcohol consumption – suggesting that it may be a recession proof industry. Although, unemployment and GDP have a positive interaction term – suggesting that as either income or free-time increases so does alcohol consumption.

Temperature has a positive relationship with consumption– with hotter places (that are not overly dry) possibly requiring a larger amount of cooler beverages. Literacy has a positive relationship as well – suggesting that as education increases as does consumption (curious considering GDP's lack of significance).

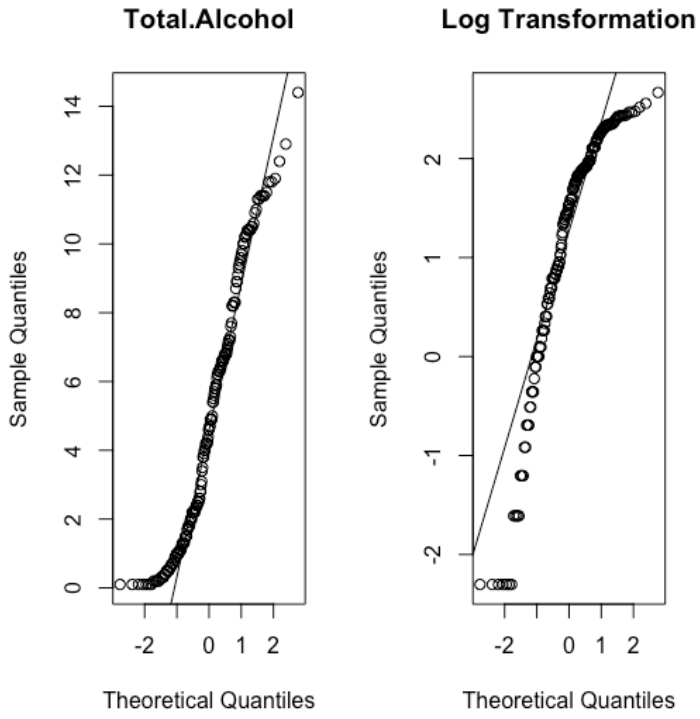
Island nations were also not found to drink significantly more or less when included in the full model – suggesting that the boxplot showing a difference in mean consumption between the two groups (island or not) was due to confounding factors captured in the model.

This result will allow entrepreneurs and markets to better determine where to place their dollars. Those nations or locals that are particularly dry or in danger of a monsoon make more target markets. While temperature is important, those markets are likely saturated. This model suggests that finding a temperature climate with a highly educated populace and moderately warm weather could prove to be an underserved market (such as parts of the Pacific Northwest of England).

A further study modeling discrepancies between liquor bought at an establishment or for consumption at home would give insight into whether rainfall drives customers inside bars or drives them away. Then an investor could make a fully informed decision of what target climate they would like to open a bar in and marketing campaigns could figure out if storm fronts meant it was time to promote a new bar or offer coupons for the local liquor store.

Appendix

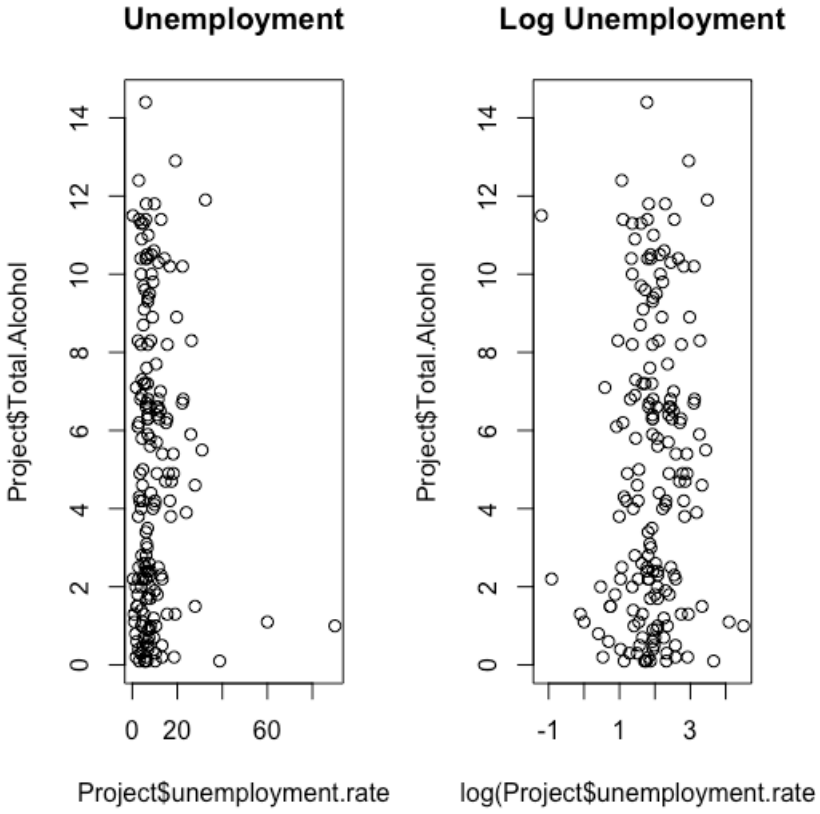
Looking at performing a log transformation on Total.Alcohol

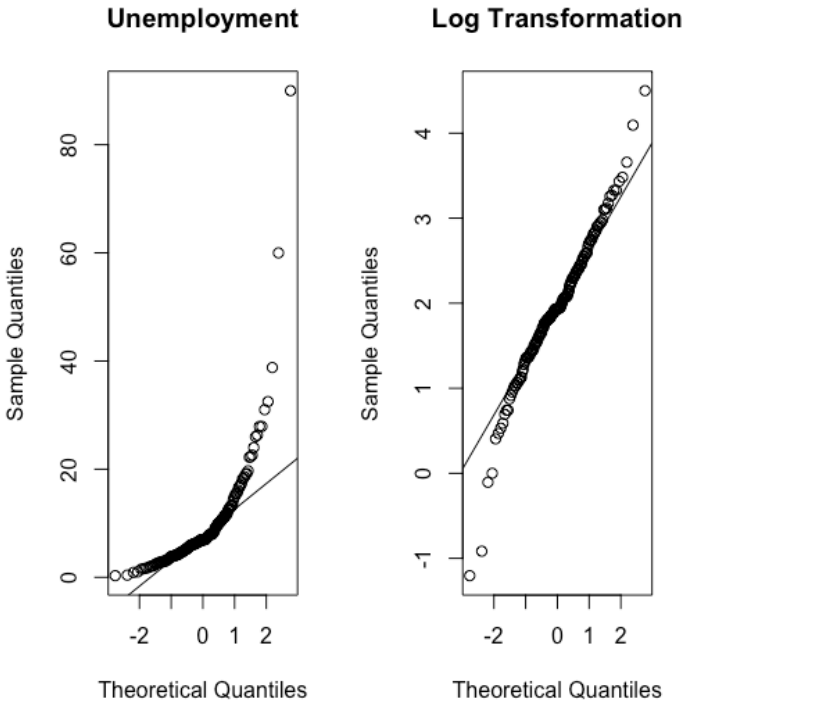


The log transformation does not help so Total.Alcohol was left as is.

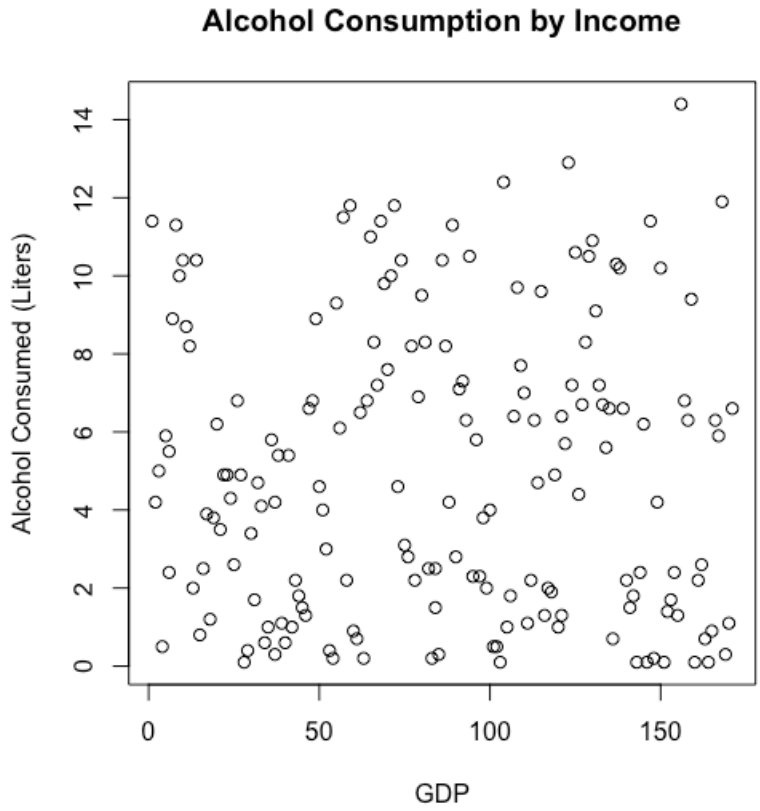
Looking at performing a log transformation on unemployment

The log transformation improves the model and is kept in the analysis.


Unemployment
Log Unemployment
Project$Total.Alcohol
0
2
4
6
8
10
12
14
0
20
60
-1
1
3
Project$unemployment.rate
log(Project$unemployment.rate

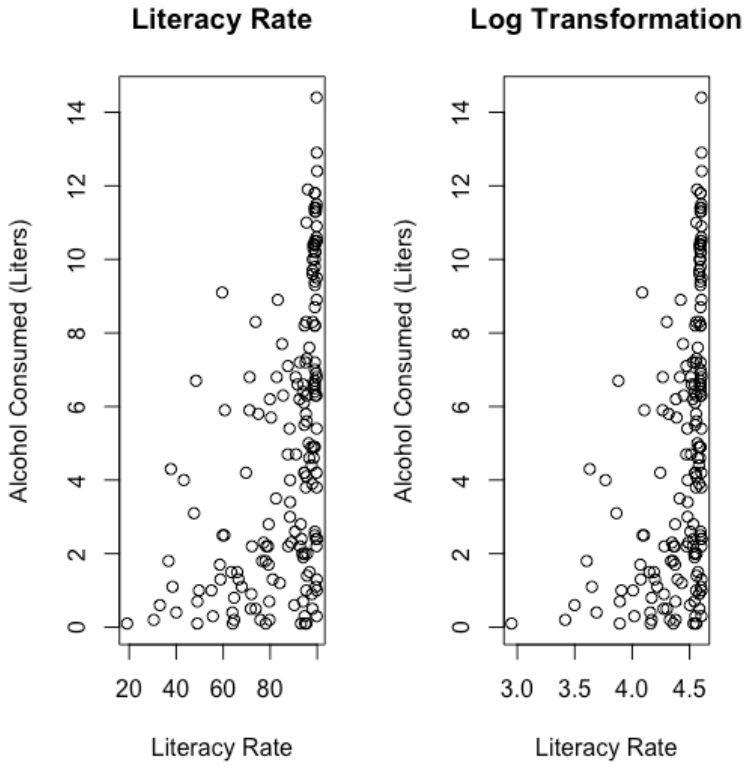

Unemployment
Log Transformation
Sample Quantiles
0
20
40
60
80
-1
0
1
2
3
4
-2
0
1
2
Theoretical Quantiles
Theoretical Quantiles

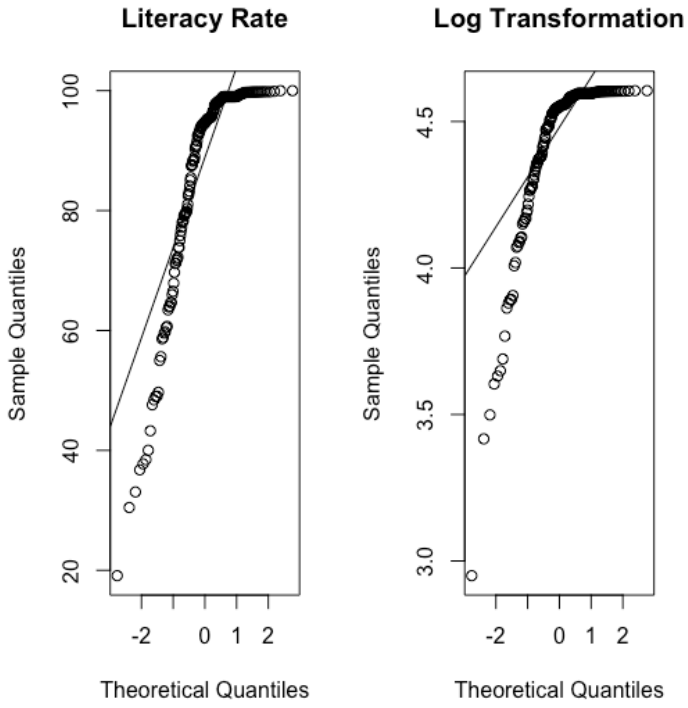
GDP



No transformations needed.

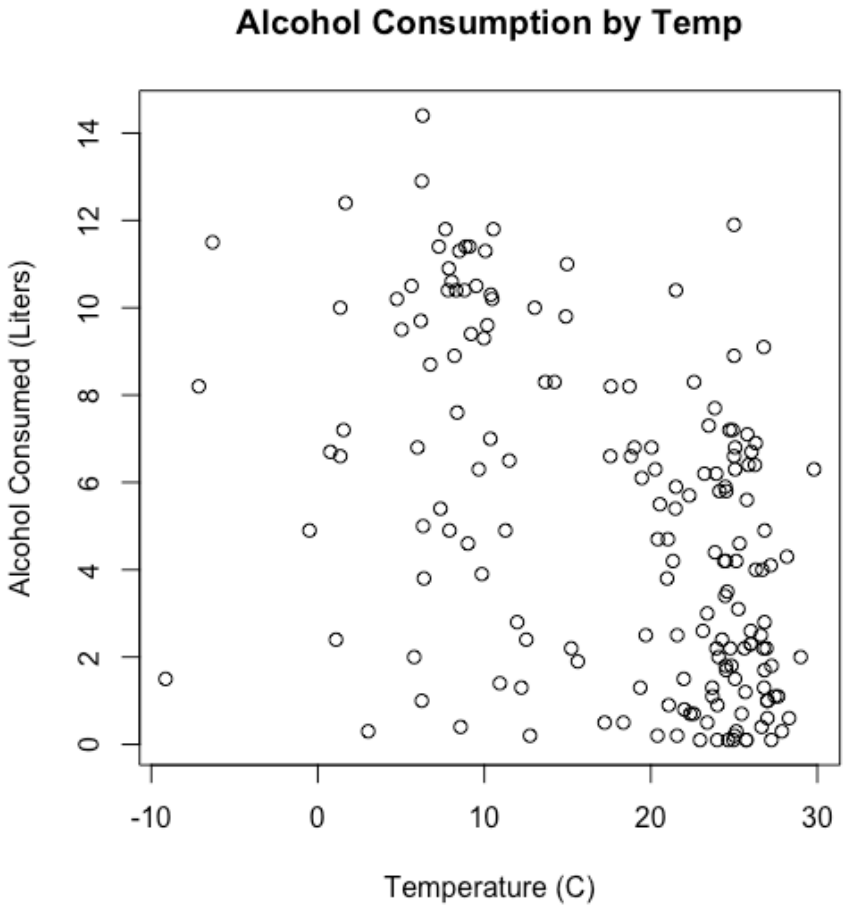
Looking at Literacy





This variable does not require a log transformation or quadratic term.

Temperature



There is no need for a quadratic transformation and the temperature ranges into the negative – meaning that a log transformation is not possible.