ACMS 30600

Final Project – Regression Analysis

Roberto Crespo

Regression for Cirrhosis Death Rate

File Submission

- alcohol.txt – raw data to be explored
- code.txt – R code for the manipulation and analysis of the data
- analysisRegression_rcrespoa.docx – this document

I. Introduction

The data explored is contained within the alcohol.txt file. The dataset contains 7 columns and 46 rows. The dataset is recollection of information from various states, especially regarding population and drinking data. The columns are arranged as follows: **index**, **One**, **urbanPop**, **lateBirths**, **wineC**, **liquorC**, and **cirrhosis**.

The **index** column represents a numbering to distinguish different entries within the dataset.

One is just a column filled with the number 1. There is no true explanation for the need or reason behind this column, it seems counterintuitive.

urbanPop represents the size of the urban population as a percentage of the total population.

lateBirths is the reciprocal of the number of births to women between 45 to 49, times a 100.

wineC is the wine consumption per capita

liquor is the liquor consumption per capita

Finally, **cirrhosis** is the death rate from Cirrhosis

Having access to this information, I am interested in noting if a relationship can be established between the death rate from Cirrhosis and the other variables. That way if a linear relationship can be established, then different expectations regarding the death certainty from Cirrhosis can be formed when considering different approaches, including medical ones. Knowing different consumption patterns of alcohol (wine and liquor) and some demographic information (composure of urban population and late births), we could potentially predict what the death rate from having Cirrhosis will be.

Note: an α=0.05 value will be established, in order to assess statistical significance. The α is the probability required threshold (p-Value) of rejecting the null hypothesis.

II. Data

The dataset comes from the Department of Scientific Computing of Florida State University. The dataset is located within a section of the web page of John Burkardt, who is a visiting research associate at the previously mentioned department. The dataset is contained within a file called x20.txt, but this file contains other strings that serve as the description and references of the data. For this reason, I extracted the raw data from the x20.txt file and copied it to a new file which I named alcohol.txt.

The file alcohol.txt includes all of the columns in the original x20.txt file. When I load the data and begin to work on it, I immediate remove 2 columns, as it can be noted in the code.txt. I preferred eliminating these columns, instead of just not including in them in the model for simplicity purposes. The logic behind eliminating these two columns is the following:

index: This column just represents the index, and other than that it represents no contribution to the data. The data was obtained from different states, and the order in which it was obtained (i.e. the index) does not contribute towards the overall relationship we are trying to find.

One: This column is just filled with the number 1. Logically, this column makes no sense. It remains constant for all entries in the dataset. The x20.txt file does not offer a compelling reasoning behind why this column is part of the data, nor what its purpose is. Statistically, this column should be removed. I ran an F test of a model without it after noting that logically it made no sense and the p-Value was 1, which means that the model without it should be kept. For these reasons, I just decided to eliminate it from the beginning of the analysis.

I found the data particularly interesting, because the relationship I am trying to establish makes sense intuitively. One should expect for the cirrhosis death rate for a particular state to increase if the population consumption of alcohol increases, that being wine or liquor. With the analysis of this data, I will note if the data follows intuition.

The data relationship that will be explored is going to be established by means of a multivariate linear regression.

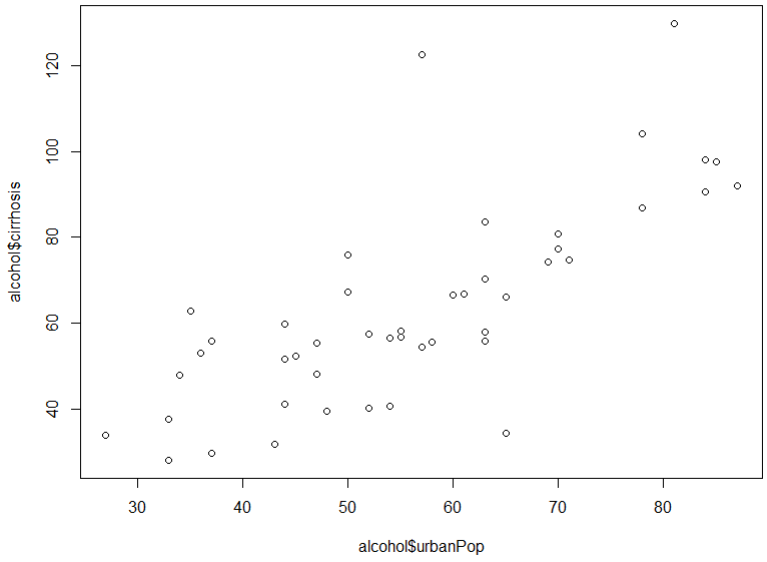
Source of data: https://people.sc.fsu.edu/~jburkardt/datasets/regression/x20.txt

III. Regression Analysis

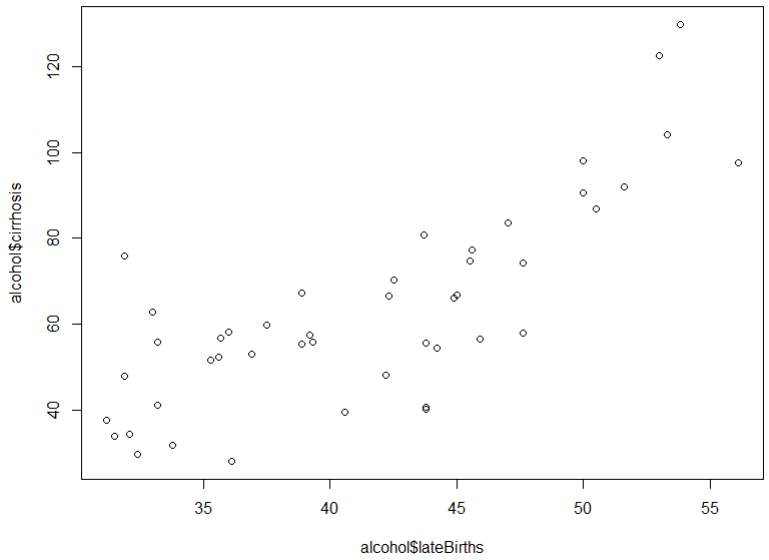
i. Exploratory

Scatterplots of X variables against Y variable

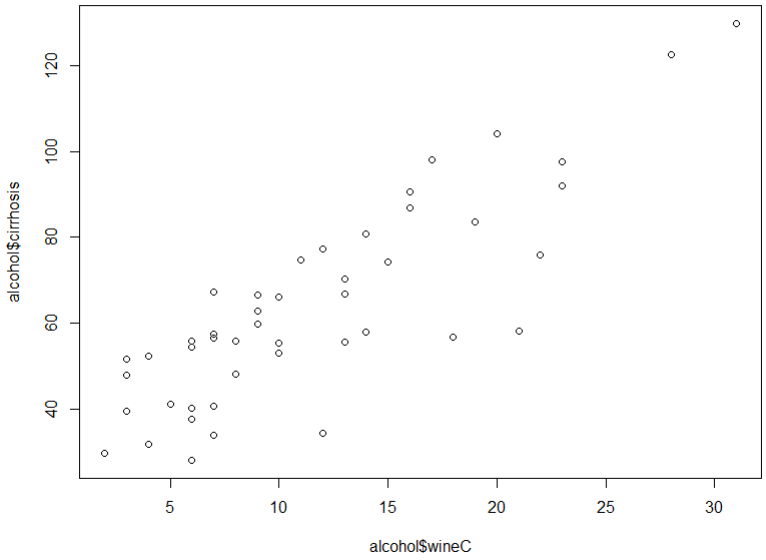
urbanPop



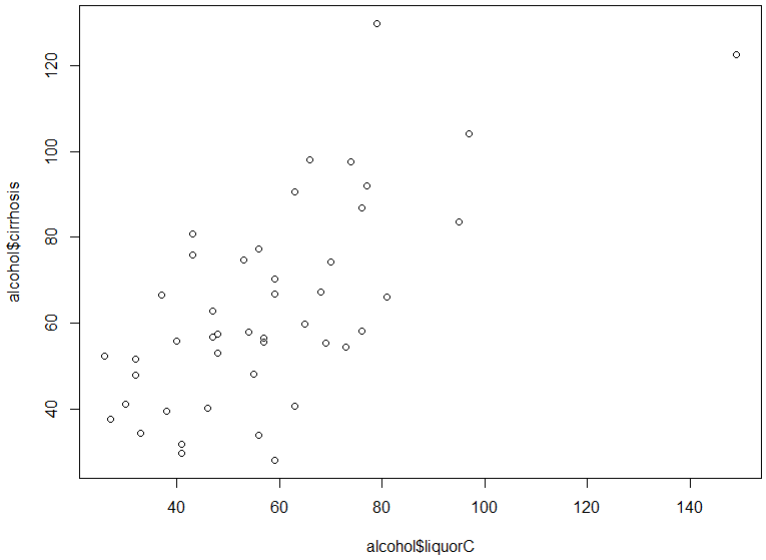
lateBirths



wineC



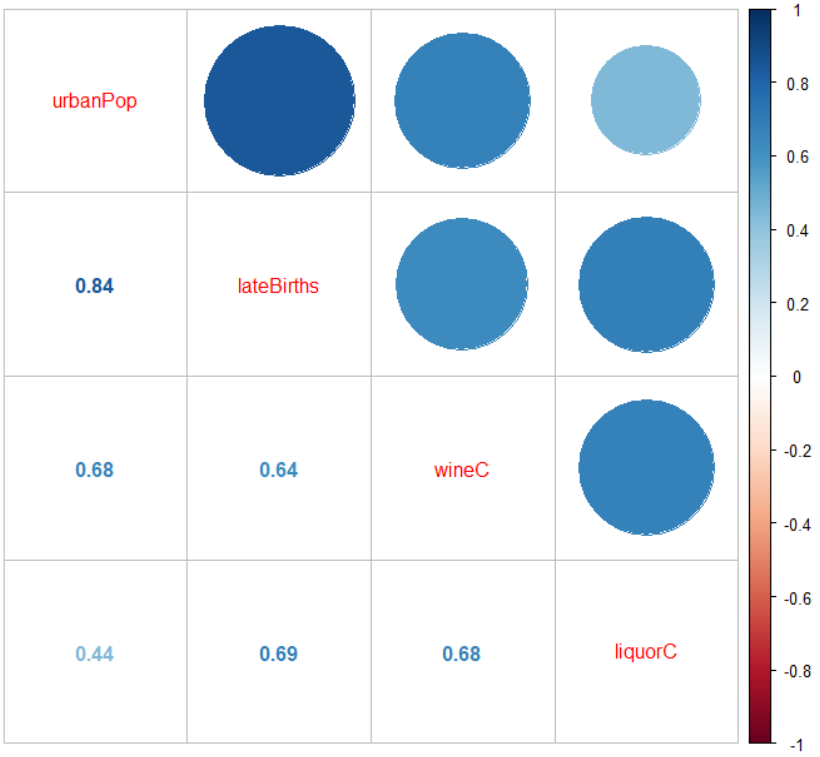
liquorC



Correlation

In order to distinguish the correlation between the X's, I depicted their correlation using a correlation plot within the **corrplot** package. The variables of the dataset are numerically represented by the following:

Here is the output of corrplot():



Insight: It can be noted from the output that the X's that have the most correlation between them are:

urbanPop **and** lateBirths

urbanPop **and** wineC

latebirth **and** liquorC

wineC **and** liquorC

ii. Linear Regression Analysis

Fitting of all Variables into One Model

I'll denote the model with all of the variables as M1.

Formula:

$$Y = \beta_0 + (\beta_{urbanPop} * x_{urbanPop}) + (\beta_{lateBirths} * x_{lateBirths})) + (\beta_{wineC} * x_{wineC})) + (\beta_{liquorC} * x_{liquorC})$$

R:

```
M1<-lm(cirrhosis ~ urbanPop+lateBirths+wineC+liquorC,data=alcohol)
```

M1 Output Highlights

R^2	=0.8136
R_a^2	=0.7954
p-Val	=1.951e-14
numerator DF	=4
denominator DF	=41

Calculated SSE =4611.109

Removal of Variables (F-Test)

Want to see if the model has similar predictive ability if I remove the urbanPop and liquor predictors. To see if I can remove these variable, I'll use an F-Test for testing if the subset has a similar R^2. I'll denote the new model (without the variables previously mentioned) as M2.

Ho: $\beta_{urbanPop} = \beta_{liquorC} = 0$

Ha: $\beta_{urbanPop} \neq \beta_{liquorC} \neq 0$

M2

Formula:

$$Y = \beta_0 + (\beta_{lateBirths} * x_{lateBirths})) + (\beta_{wineC} * x_{wineC})$$

R:

```
M2<-lm(cirrhosis ~ lateBirths+wineC,data=alcohol)
```

M2 Output Highlights

R^2	=0.8128
R_a^2	=0.8041
p-Val	= 2.268e-16
numerator DF	=2
denominator DF	=43

Calculated SSE =4632.076

F test statistic = $\frac{\frac{(SSE_{M2}-SSE_{M1})}{(k-g)}}{\frac{SSE_{M1}}{[n-(k+1)]}}$

F test statistic=$\frac{(4632.076-4611.109)/(4-2)}{4611.109/[46-(4+1)]}$

F test statistic=0.0932

p-Val= pf(.0932,2,41,lower.tail=F)

p-Val=0.911

Conclusion → As the p value is greater than the alpha of 0.05, there is no statistical evidence to reject Ho. Therefore, we accept M2 as the new model. We remove the urbanPop and liquorC variables from the model.

Automated Variable Selection Process

Want to find the most useful model (more predictive ability) that has the least number of predictors. I will select the variables by using the Akaike information criterion (AIC) as a statistical criterion to determine which variables should be kept in the model. The smaller the AIC, the better the model performs. I'll be using the stepAIC() within the MASS library in RStudio. I'll start with the complete model (M1) and then let the function determine the best reduced version of it

R:

```
library(MASS)
stepAIC(M1)
```

The function first removes the liquor predictor, as a model with all predictors but liquorC yields the lowest AIC as opposed to those made without the other predictors respectively. Then a model without the urbanPop predictor yields the lowest AIC. Then after that, further removals of any predictor yields no decrease in AIC so the process ends. The stepAIC() functions suggests, that a model without urbanPop and liquor should be used. Coincidently, this is the model we have previously fitted as M2. This is a good confirmation to our previous conclusion. The first variable screening method of checking whether certain predictors could be removed proved that those variables should be removed, and the model without those variables is the model that is recommended using the stepAIC() function for variable screening.

★ From now on I will utilize M2 as the model of prediction.

Outlier Detection

I now check if there are any outliers, that is any data point that is within three standard deviations away. I'll check this by taking a look at the z-scores of the data points and the predicted values from M2.

R:

```
which(abs(rstandard(M2))>3
```

This model produces no outliers.

Influential Observations Detection

I still need to check for any influential observations for M2. That is any data point that its removal would cause a large change in the fitting of the model. A manner to detect these points is by checking the Cook's distance and seeing if its greater than the 50^{th} percentile of the F distribution.

R:

```
F_thresh<-qf(.5,3,43)

which(cooks.distance(M2)>F_thresh)
```

This model has no influential observations.

Final Model

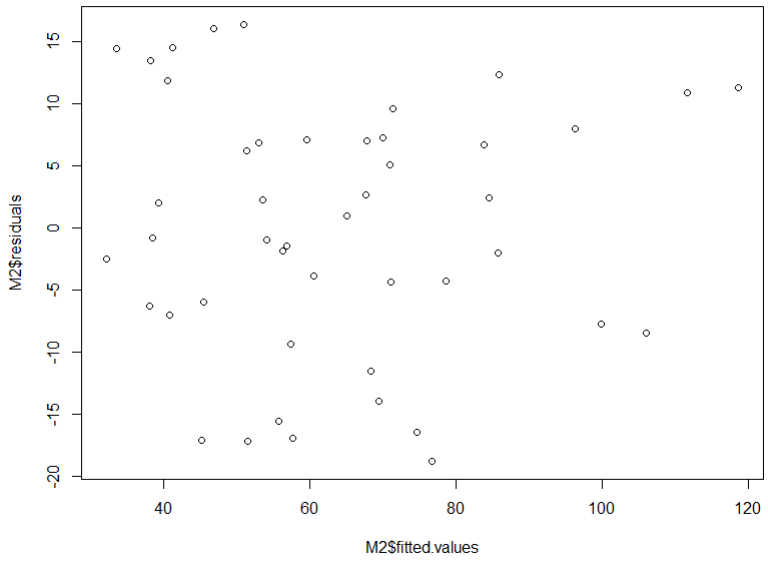
I chose to select M2 as my final predicting model. M2 consist of predicting cirrhosis with lateBirths and wine C. This model was selected over M1 (model with all predictors), because it has little change in predictive ability. R^2 only drops by 0.098%. R_a^2 increases by 0.109%. The F-test variable screening method, proved that the removal of the variables urbanPop and liquor are adequate. The stepAIC() functions seconds that a model without these two variables is the optimal model. The model without those variables is M2, and for these reasons this is the final model.

$$Y = \beta_0 + (\beta_{lateBirths} * x_{lateBirths})) + (\beta_{wineC} * x_{wineC})$$

Model Diagnosis

This model was fitted and explained assuming that certain assumptions were true. Now I will check all assumptions and note if they are upheld it will mean that the model is valid. First looking at the fitted values of the final model plotted against the residuals of the final model can give visual insight. I will denote this plot as plotD. Any patterns observed in it may indicate problems with the model.

PlotD



At first glance, it seems that there is no visible trend or pattern that suggests that at least one assumption is not upheld. It seems to be a good model, but I will go in depth of all specific assumptions and check their integrity. These are the following factors that were assumed to be true and now are checked to see if the assumptions were true:

i. ε ’s follow normal curve

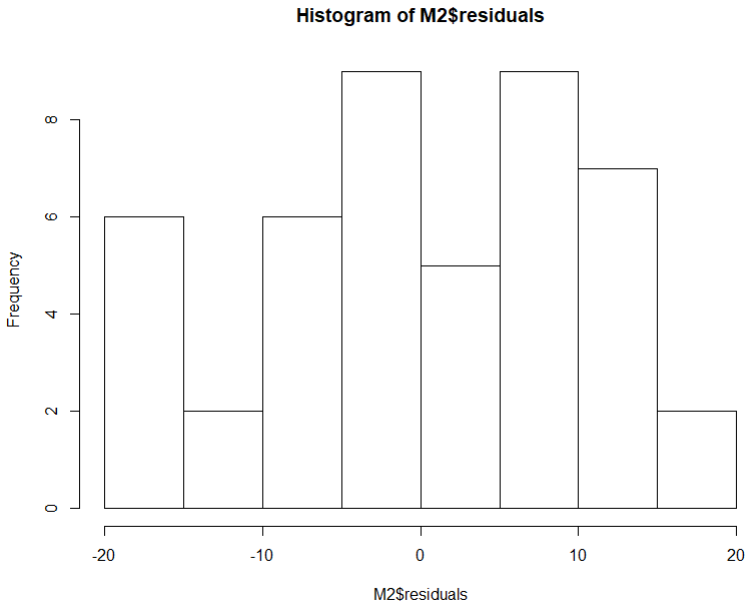
- Histogram

Trying to see if there is some sort of bell shape, that will suggest that the assumption is true.

R:

hist(M2$residuals)

Using further argument of breaks, to denote the vector breakpoints. This particular output is given by hist(M2$residuals,breaks=12)



Conclusion → The curve “seems” to follow normal curve. No significant visual contradiction to the assumption.

- Shapiro-Wilk Test

Is a test used to assess whether ε ’s follow normal curve. The test has as for its null hypothesis that ε ’s are normally distributed. So ideally, in this case, a p value bigger than alpha .05 is desired, as that would suggest that the null can’t be rejected; and thus, the ε ’s follow normal curve.

Ho: ε ’s are normally distributed

Ha: ε ’s are NOT normally distributed

R:

shapiro.test(M2$residuals)

Conclusion → As output of shapiro.test(), the p value of 0.0953 is yielded. This is larger than alpha .05, which means that there is no significant statistical evidence to reject the null hypothesis, so the test suggests that the ε 's are normally distributed.

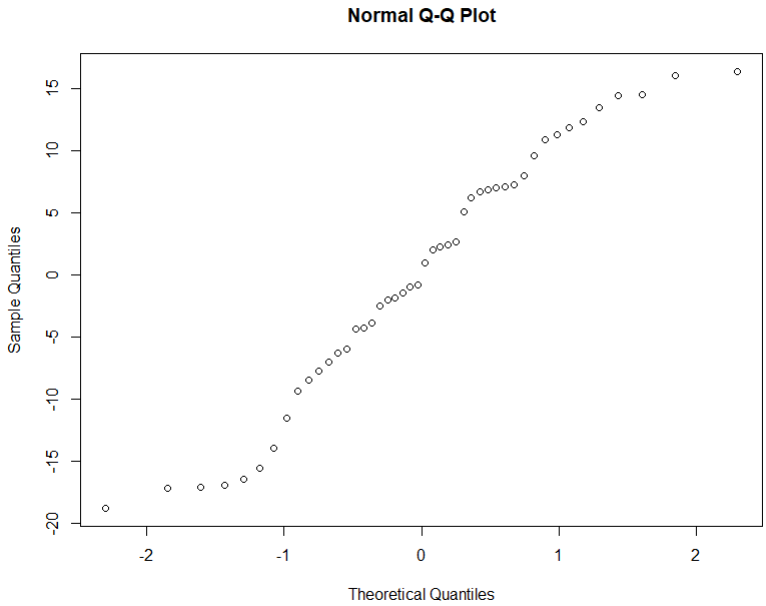
- Normal Probability Plot

A normal probability plot of the residuals of M2 should yield roughly a straight diagonal (approx. 45 degrees) to suggest that the ε 's are normally distributed.

R:

```
qqnorm(M2$residuals)
```

Conclusion → The line visually looks pretty straight. No significant visual contradiction to the assumption.



Overall Conclusion of Assumption i:

All of the three processes to discern whether the ε 's are normally distributed agree. All three conclude that the ε 's follow the normal curve. Assumption is upheld.

ii. ε ’s have constant variance

There is no visual confirmation on plotD that suggests that this assumption is not being upheld. There is no cone or megaphone pattern in the plot of fitted values against residuals that would suggest that there is heteroscedasticity within the model. For this reason, there is no need for a variable transformation as there is no problem with the patterns in the plot.

Overall Conclusion of Assumption i:

As no visual pattern is noted in plotD that would suggest otherwise, the assumption is upheld.

iii. ε ’s are independent (not correlated)

In order to find out if this assumptions remains valid, a Durbin-Watson test must be performed. This test will hold as its null hypothesis that the ε ’s are not correlated, that is that they are independent. Ideally want a p-Value that will fail to reject this null hypothesis to signify that the assumption is upheld. To access the Durbin Watson test we must use the **lmtest** library. The test uses d as a test statistic which is distributed in the domain [0,4]. A 0 represents perfect negative correlation, a 2 represents no correlation, and a 4 represents a perfect positive correlation.

Durbin-Watson Test

Ho: ε ’s are not correlated

Ha: ε ’s are correlated

R

```
install.packages("lmtest")
library(lmtest)
dwtest(M2)
```

Output:

d	=2.5152
p-value	=0.9639

Overall Conclusion of Assumption iii:

According to the Durbin-Watson test, there is no statistical evidence to reject H_o. That means that we accept H_o, which is that the ε ’s are not correlated. Thus, Assumption is upheld.

iv. ε ’s have mean 0

Running simple calculations in R about the residuals for M2 to determine the validity of the assumption, gives us a straight answer. Taking the mean of the residuals yields a number very close to zero.

R

```
mean(M2$residuals)
```

Output:

```
2.295115e-16
```

Overall Conclusion of Assumption iv:

The ε ’s have a mean pretty close to zero, to the point where its basically zero. Assumption is upheld.

Conclusion on Assumptions

All assumptions were upheld; thus, the final model is a valid linear regression. The insights, that are obtained from tests and statistical knowledge from the model, assuming they are statistically backed up, will be statistically valid.

IV. Results

Taking a closer look at the remaining predictors, we will explore their statistical significance. A breakdown of each of the remaining two predictors, including two confidence intervals (at 95% and at 99%), will give insight to what the predictor represents for the final model.

$\beta_{lateBirths}$

Significance

p-Value: 2.08×10^{-5}

This p-Value makes the predictor very statistically significant. The value is far from the predetermined α (.05). Also the p-Value is very close to 0, which seconds the fact that is significant. R output marks it with a ***, meaning it pertains to the most statistically significant category that R classifies.

Breakdown of Test

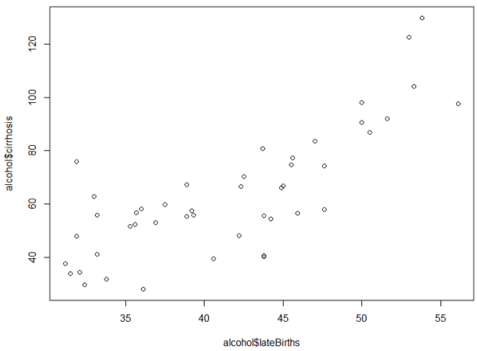
H_0: $\beta_{lateBirths} = 0$

H_a: $\beta_{lateBirths} \neq 0$

$$test\ statistic: t = \frac{\hat{\beta}_{lateBirths} - Hypothesized\ value\ of\ \beta_{lateBirths}}{s_{\hat{\beta}_{lateBirths}}}$$

$$test\ statistic: t = \frac{1.3656 - 0}{0.2858}$$

$$test\ statistic: t = 4.778167$$



p-Value

R

```
2*pt(4.778167,43,lower.tail=F)
```

Output

p-Value: 2.08605×10^{-5}

This p-value is lower than our alpha and very close to zero, so there is enough statistical evidence to reject the null hypothesis. Also, the plot of lateBirths against cirrhosis (response var iable), visually indicates that the alternative hypothesis is true. This plot was previously shown, but for pr actical measurements I have shown it again in chartLB. A positive trend can be noted, suggesting that the coefficient of the beta predictor is positive and is not 0. This is the case, for what we obtain in the final m odel. The value obtained for that coefficient is 1.3656.

95% Confidence Interval

R

1.3656+qt(.975,(46-(2+1)))*.2858

1.3656-qt(.975,(46-(2+1)))*.2858

CI: [0.7892294 - 1.941971]

We are 95% confident that the true Beta predictor lies in our Confidence Interval

$$\beta_{wineC}$$

Significance

p-Value: 2.69×10^{-8}

This p-Value makes the predictor very statistically significant. The value is far from the predetermined α (.05). Also the p-Value is very close to 0, which seconds the fact that is significant. R output marks it with a ***, meaning it pertains to the most statistically significant category that R classifies.

Breakdown of Test

H_0: $\beta_{wineC} = 0$

H_a: $\beta_{winceC} \neq 0$

$$test\ statistic: t = \frac{\hat{\beta}_{lateBirths} - Hypothesized\ value\ of\ \beta_{lateBirths}}{s_{\hat{\beta}_{lateBirths}}}$$

$$test\ statistic: t = \frac{1.9723 - 0}{0.2909}$$

$$test\ statistic: t = 6.77999$$

p-Value

R

2*pt(6.77999,43,lower.tail=F)

Output

p-Value: 2.68205×10^{-8}

This p-value is lower than our alpha and very close to zero, so there is enough statistical evidence to reject the null hypothesis. Also, the plot of wineC against cirrhosis (response variab le), visually indicates that the alternative hypothesis is true. This plot was previously shown. A positive trend can be noted, suggesting that the coefficient of the beta predictor is positive and is not 0. This is the case, for what we obtain in the final model. The value obtained for that coefficient is 1.9723.

95% Confidence Interval

R

```
1.9723+qt(.975,(46-(2+1)))*.2909
```

```
1.9723-qt(.975,(46-(2+1)))*.2909
```

CI: [1.385644 – 2.558956]

```
We are 95% confident that the true Beta predictor lies in our
Confidence Interval
```

V. Conclusion

I believe that this model is a good predictor for the death rate from Cirrhosis. The response variable's behavior is described with accuracy by the model. Having more predictors that can potentially contribute would have been great for building an ever more precise model. More variables that perhaps could indicate patterns of death rate from cirrhosis other than just drinking and city demographics could be insightful. Perhaps knowing if there is a tendency in a patient to develop this type of diseases (i.e does the family has a history of similar diseases). Obviously, significance tests would need to be done to see if such a variable would contribute to the model. Knowing the death rate from cirrhosis from consumption patterns and late births is important for medical purposes. Perhaps having this information and seeing certain patterns of consumptions can lead to preventive measurements.