Homework: Methods and Data Analysis 1

Srishti Saha (ss1078)

05 September, 2019

```
# importing general libraries
library(ggplot2)
```

Question 1: OLD FAITHFUL

```
# importing dataset
oldfaithful<- read.csv("C:\\Users\\srish\\Desktop\\books and study material\\Methods and Data Analysis

library(data.table)
of_dt<-as.data.table(oldfaithful)
#is.data.table(of_dt)
of_dt[, lag.Duration:=c(NA, Duration[-1]),by=Date]

#overview of the dataset
summary(oldfaithful)
```

```
##        X               Date          Interval       Duration
##  Min.   :  1.0   Min.   :1.000   Min.   :42.0   Min.   :1.700
##  1st Qu.: 27.5   1st Qu.:3.000   1st Qu.:59.0   1st Qu.:2.300
##  Median : 54.0   Median :5.000   Median :75.0   Median :3.800
##  Mean   : 54.0   Mean   :4.514   Mean   :71.0   Mean   :3.461
##  3rd Qu.: 80.5   3rd Qu.:6.000   3rd Qu.:80.5   3rd Qu.:4.300
##  Max.   :107.0   Max.   :8.000   Max.   :95.0   Max.   :4.900
```

Part A: Creation of a model

Let us try to create a model to estimate the intervals between eruptions from the duration of the previous one. A simple linear regression would take the following form:

$$Interval_i = \beta_o + \beta_1 Duration_{i-1} + \epsilon_i$$

Here $Duration_{i-1}$ simply represents the duration of the previous eruption. $Interval_i$ is the perios between the current eruption and the previous one. ϵ_i is the error term associated with the mathematical representation. The β_1 term represents the coefficient of the variable (Duration) in the model and β_o is the constant term (intercept).

Part B: Fitting the model

Let us first check the correlations.

```
#check correlation matrix
of_dt<-as.data.frame(of_dt)
cor(of_dt[3:5], use="complete.obs")
```

```
##              Interval  Duration lag.Duration
## Interval     1.0000000 0.8596709    0.8596709
## Duration     0.8596709 1.0000000    1.0000000
## lag.Duration 0.8596709 1.0000000    1.0000000
```

Creating the model and getting model summary:

```r
# regression model for predicting Interval from Duration
lm_model <- lm(Interval~Duration,data=of_dt);
summary(lm_model)
```

```
##
## Call:
## lm(formula = Interval ~ Duration, data = of_dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.644  -4.440  -1.088   4.467  15.652
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.8282     2.2618   14.96   <2e-16 ***
## Duration     10.7410     0.6263   17.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.683 on 105 degrees of freedom
## Multiple R-squared:  0.7369, Adjusted R-squared:  0.7344
## F-statistic: 294.1 on 1 and 105 DF,  p-value: < 2.2e-16
```

Interpreting the model results

1. We first see that p-value associated with Duration is very low (i.e. of the order e^{-16}). Hence, Duration is a significant variable. The value of β_1 equal to the estimate corresponding to *Duration* and is equal to 10.7410. This implies that just on the basis of Duration (duration of previous eruption), having any and all other variables constant, an eruption will have an Interval of 10.74 greater than the baseline.

2. The value of the intercept is 33.8249. This means that for an eruption instance with the duration of the previous eruption equal to 0, the interval will be 33.82.

3. The residual standard error(rse) is 6.683 on 105 degrees of freedom. It is the positive square root of the mean square error and determines how good is the model to fit the line to the given points. A residual standard error of 0 implies that the model fits the data perfectly. 6.68 is a high value which indicates that the predicted values are distant from the true observed values.

```r
# calculating mean squared error
mse<-mean(summary(lm_model)$residuals^2)

sqrt(mse)
```

```
## [1] 6.619861
```

This is very close to the residual standard error given by the model above.

4. The R-squared is 0.7369. This means that 73.69% of the variation in y is explained by the model. However, since we shold look at adjusted R-squared to adjust for higher number of variables added, we see that adjusted R-squared is also in the same range i.e. 0.7344 (This might be because only 1 variable has been used to construct the model.)

Part C: 95% confidence interval of slope

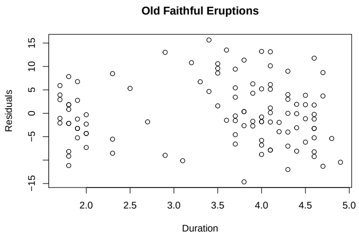
```
confint(lm_model, level=0.95)
```

```
##                 2.5 %   97.5 %
## (Intercept) 29.343441 38.31297
## Duration     9.499061 11.98288
```

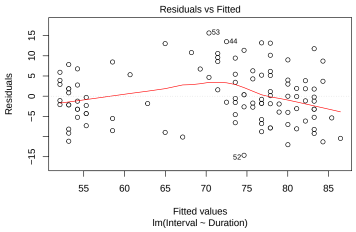
For *Duration* the 95% confidence interval is (9.50,11.98). This means the fitted value of β_1 is 10.74 with an interval of (9.50,11.98). This interval represents the range in which we can say that the true value of slope of the line lies with a 95% confidence.

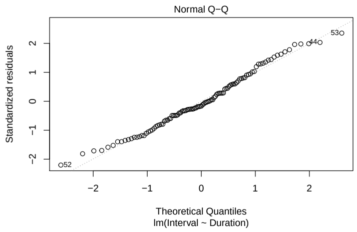
Part D: Residual Plots and assumptions of Linear Regression

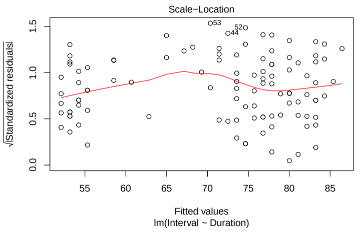
```
model_resid= resid(lm_model)

#residual versus x
plot(of_dt$Duration, model_resid,
     ylab="Residuals", xlab="Duration",
     main="Old Faithful Eruptions")
```

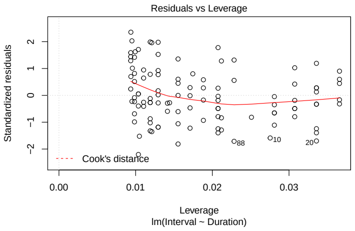


```
#abline(0, 0)                       # the horizontal reference

#residual versus fitted
plot(lm_model)
```









- Linearity: Based on the residual plot, the residuals seem to be linearly related to Duration. There is no evident patten in the residual plot which indicates that the model will capture the pattern and the information within Y and X. However, there might be a very slight non-linearity on the basis of the slight parabolic spread of the points. Thus, the assumption of Linearity is met (maybe slightly). If the relationship was completely linear, the relation of y and x would have met complete Linearity.
- Independence and Equal Variance: The residual versus fitted plot does not show an evident conical (spread out or converging) patterm. Howrever, it is slightly non-linear. This might indicate that there is no heteroskedasticity, but there is a slight non-linear pattern that is not captured by the fitted values and x.
- Normality: The Q-Q plot can be used to interpret Normality. If the standard residuals are linearlity related to the theoretical quantiles (no curve), then the condition of Normality is met. Here, normality is met.

Part E: 95% prediction intervals

```
newdata1 <- data.frame("Duration" = c(2, 2.5, 3, 3.5, 4))
pred_int<-predict(lm_model, newdata = newdata1, interval = 'prediction',level=0.95)
pred_int
```

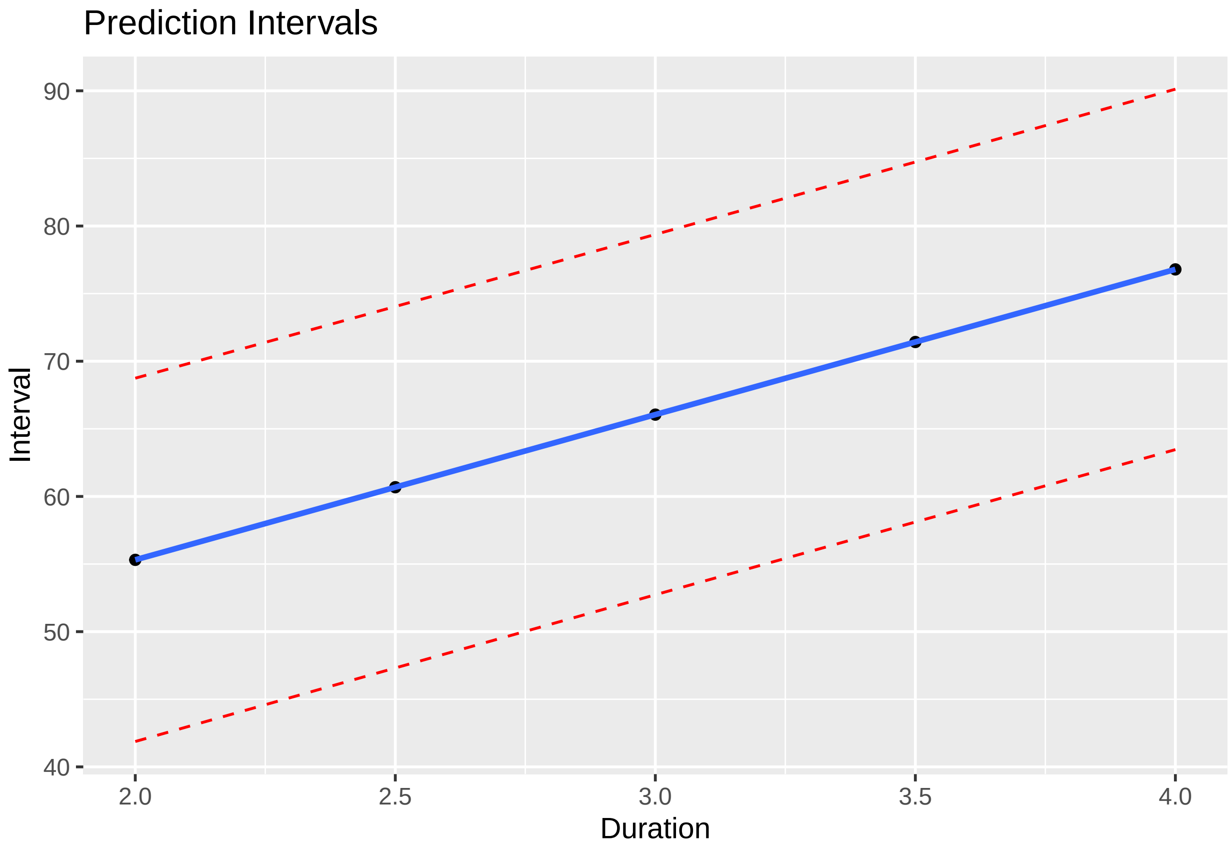
```
##        fit      lwr      upr
## 1 55.31015 41.87495 68.74535
## 2 60.68064 47.31512 74.04616
## 3 66.05112 52.72668 79.37557
## 4 71.42161 58.10936 84.73385
## 5 76.79209 63.46310 90.12108
```

The above snippet gives us the 95% Prediction Interval.

```
predictions<-cbind(newdata1,pred_int)

predictions
```

```
##   Duration      fit      lwr      upr
## 1      2.0 55.31015 41.87495 68.74535
## 2      2.5 60.68064 47.31512 74.04616
## 3      3.0 66.05112 52.72668 79.37557
## 4      3.5 71.42161 58.10936 84.73385
## 5      4.0 76.79209 63.46310 90.12108
```

```
ggplot(predictions, aes(x=Duration,y=fit)) + geom_point()+geom_line(aes(y=lwr),color="red",linetype="da
```



Question 2: RESPIRATORY RATES FOR CHILDREN

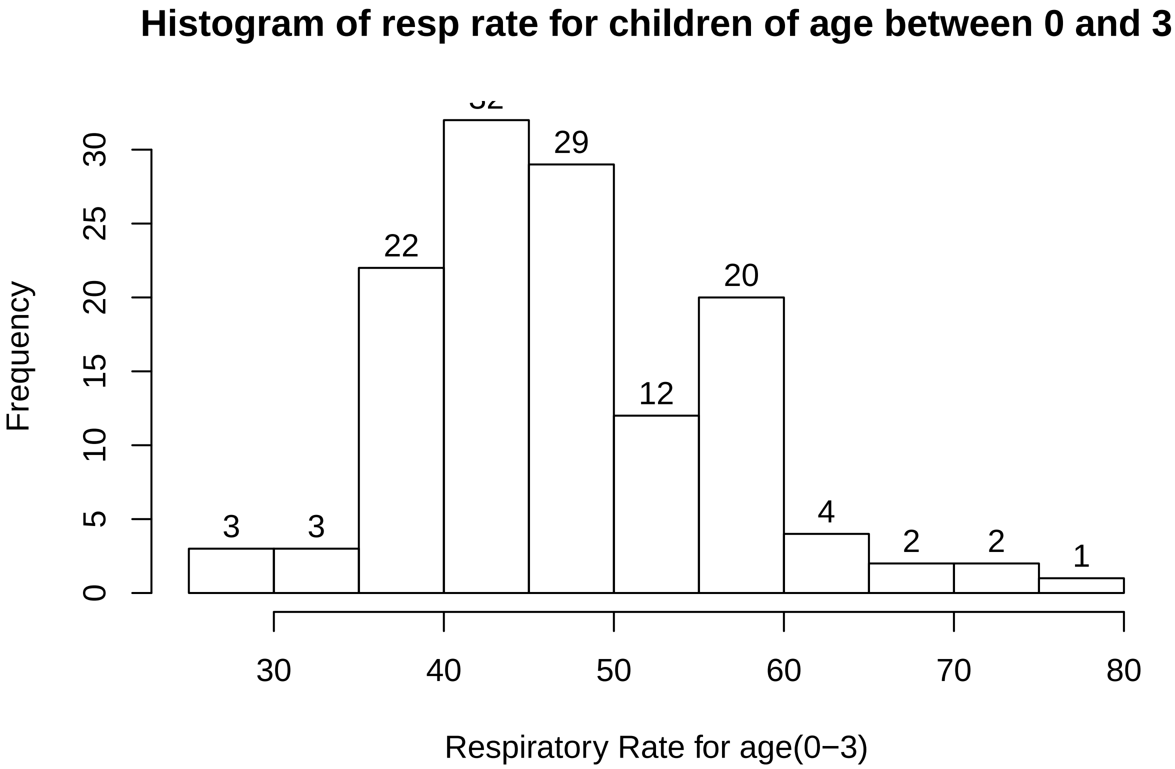
```
# importing dataset
respiratory<- read.csv("C:\\Users\\srish\\Desktop\\books and study material\\Methods and Data Analysis
summary(respiratory)
```

```
##        X              Age             Rate      
##  Min.   :  1.0   Min.   : 0.10   Min.   :18.00  
##  1st Qu.:155.2   1st Qu.: 3.80   1st Qu.:30.00  
##  Median :309.5   Median :10.55   Median :36.50  
##  Mean   :309.5   Mean   :13.39   Mean   :37.74  
##  3rd Qu.:463.8   3rd Qu.:22.00   3rd Qu.:44.00  
##  Max.   :618.0   Max.   :36.00   Max.   :78.00  
```

Part A: Data Understanding and plot for respiratory rate and age

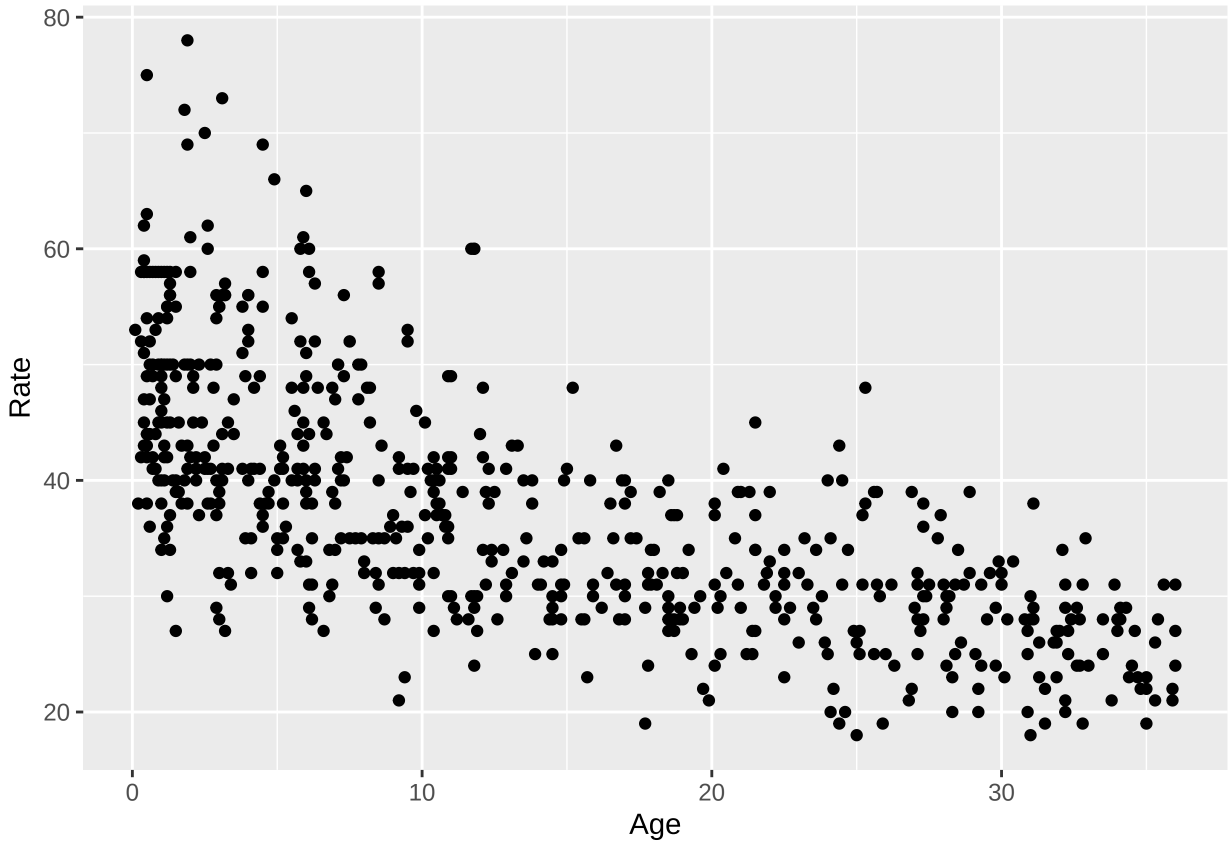
Let us first look at the distribution of respiratory rates in the subset of the dataset where age is between 0 and 3

```
#creating subset of datset
respiratory_0to3<- respiratory[respiratory$Age>0 & respiratory$Age<3,]

#plotting histogram of rate
hist(respiratory_0to3$Rate ,main='Histogram of resp rate for children of age between 0 and 3', xlab= 'R
```



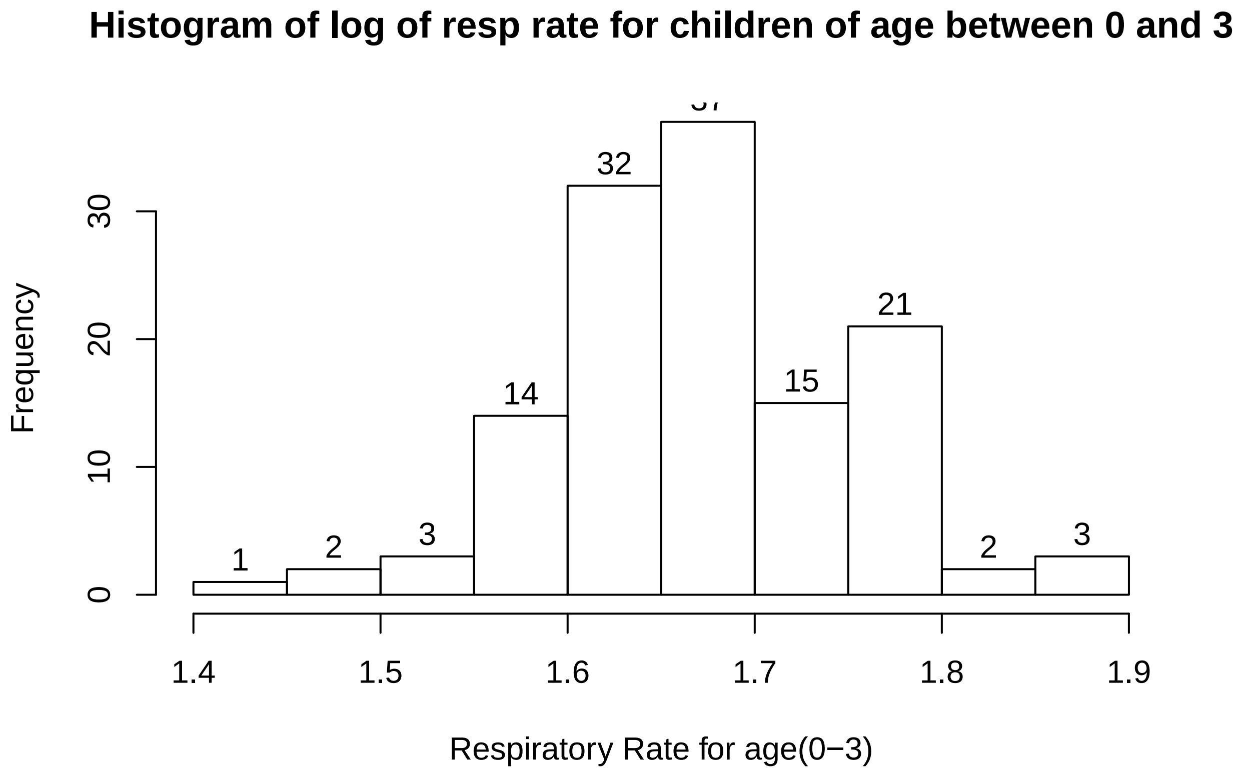
From the above plot, we see that there is an evident right skew to the plot. Let us also plot a scatter plot between the two variables from the entire dataset.

```
#scatter plot
ggplot(data = respiratory, aes(x = Age, y = Rate)) + geom_point() +
labs(y = "Rate")
```

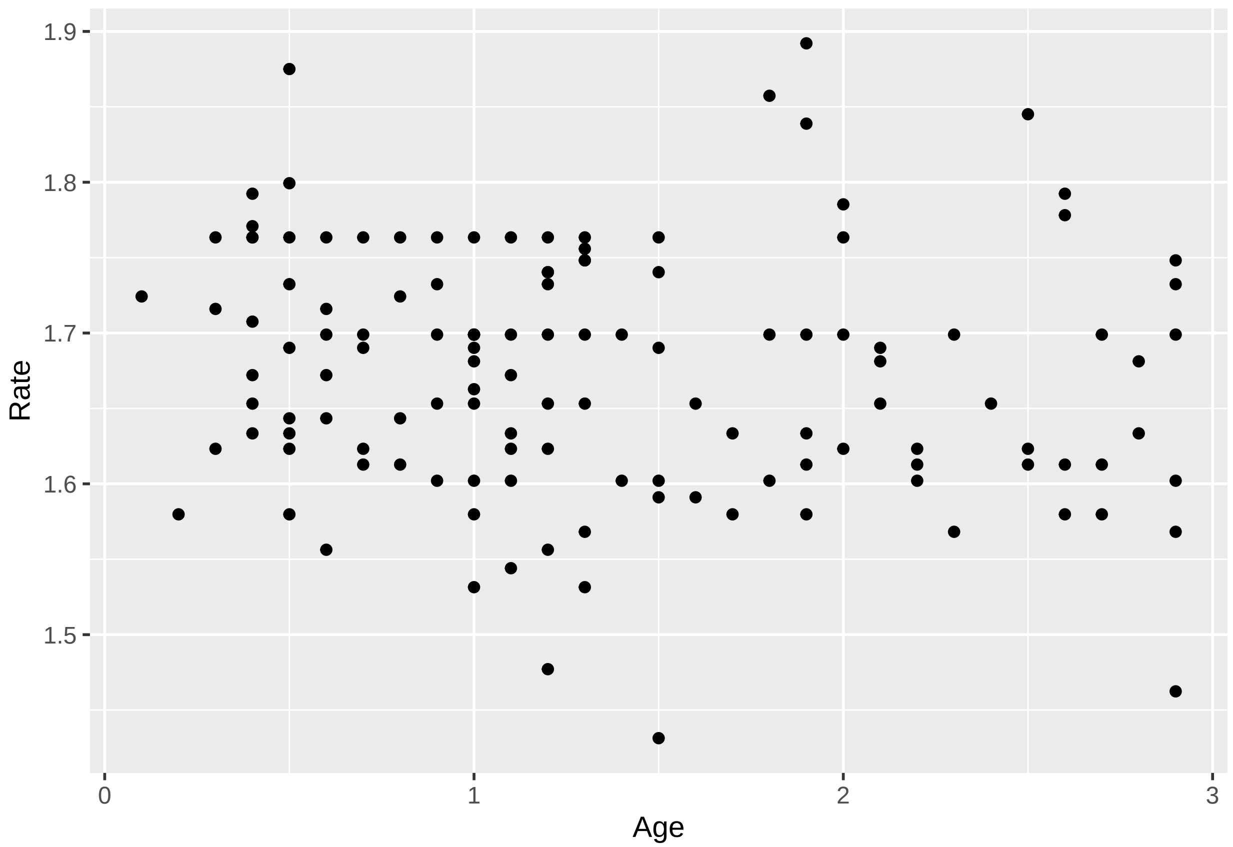


We see a non-linear trend. As suggested later in the question as well, let us do a log transformed Rate.

```
#log-transformation of both variables
respiratory$log.Rate <- log10(respiratory$Rate)
respiratory$log.Age<- log10(respiratory$Age)

# plotting histogram and scatter plot for log transformed rate
#creating subset of datset
respiratory_0to3<- respiratory[respiratory$Age>0 & respiratory$Age<3,]

#plotting histogram of rate
hist(respiratory_0to3$log.Rate ,main='Histogram of log of resp rate for children of age between 0 and 3
```



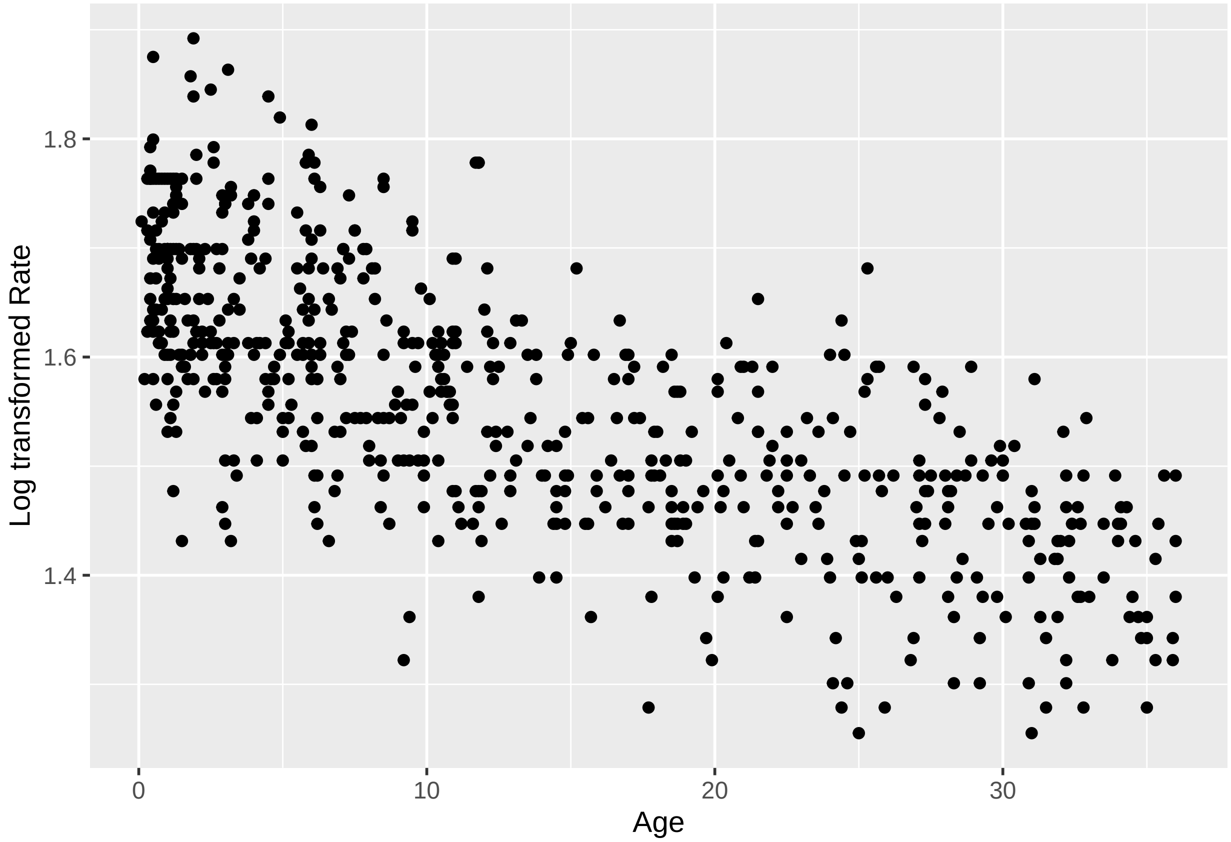
```
#scatter plot
ggplot(data = respiratory_0to3, aes(x = Age, y = log.Rate)) + geom_point() +
labs(y = "Rate")
```



Part B: Regression that predicts respiratory rates from age

In the above part, let us plot the scatter plot of the log transformed Rates versus Age from the entire dataset to observe the transformed relationship.

```
#scatter plot of log-transformed respiratory rate
ggplot(data = respiratory, aes(x = Age, y = log.Rate)) + geom_point() +
labs(y = "Log transformed Rate")
```



This plot shows a linear relationship between rate and age.

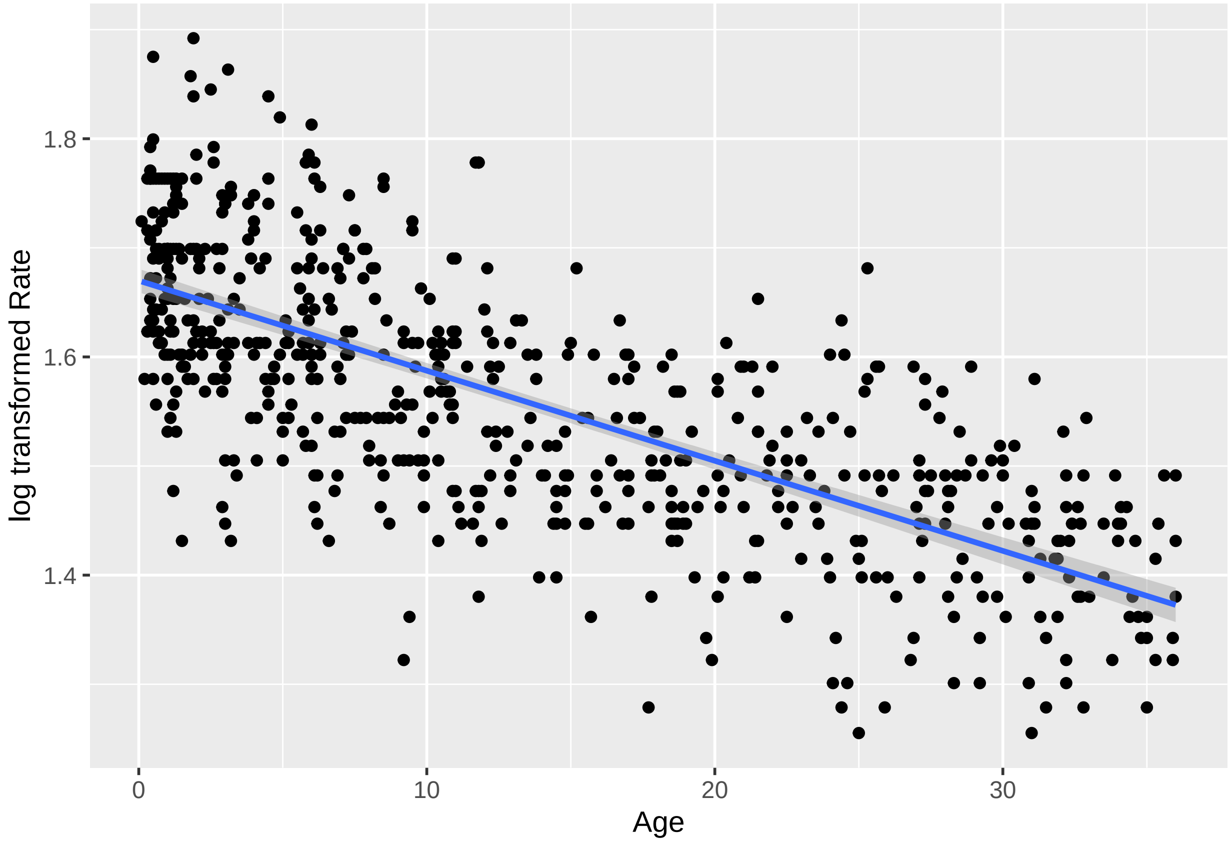
```
# regression model for predicting log.Rate votes from Age
lm_model_resp <- lm(log.Rate~Age,data=respiratory);
summary(lm_model_resp)
```

```
##
## Call:
## lm(formula = log.Rate ~ Age, data = respiratory)
##
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.271744 -0.057330 -0.001746  0.058581  0.237866 
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.6699138  0.0054841  304.50   <2e-16 ***
## Age         -0.0082555  0.0003195  -25.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0853 on 616 degrees of freedom
## Multiple R-squared:  0.5201, Adjusted R-squared:  0.5193 
## F-statistic: 667.6 on 1 and 616 DF,  p-value: < 2.2e-16
```

Model Evaluation:

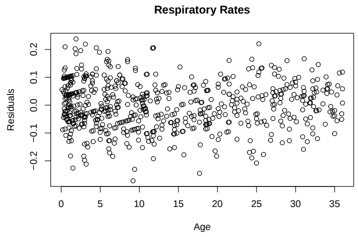
1. We see that Age is a significant variable to predict log.Rate (log transformed rate). The value of β_1 is -0.008.
2. R-squared is 0.5201 while the adjusted R-squared is 0.5193. This indicates ~52% of variation in RRespiratory rate (log transformed) is explained by Age (moderate effect).
3. Residual Standard Error is 0.0853 which is close to 0. Thus the true points are close to the predicted values. The model is a fairly accurate model.
4. A low p-value of the F-statistic (of the order e^{-16}) indicates that the null hypothesis that the 2 variables (Rate and Age) are not related can be rejected. hence, the model is justified by the alternate hypothesis that the Rate (log-transformed) is related to Age.

```
ggplot(data = respiratory, aes(x = Age, y = log.Rate)) + geom_point() +
labs(y = "log transformed Rate") + stat_smooth(method = lm)
```

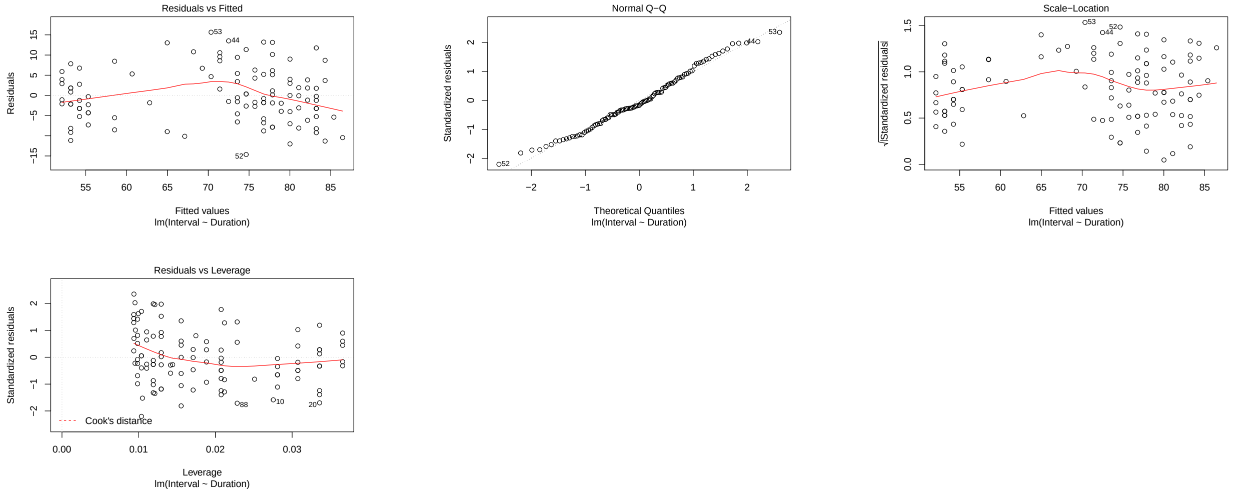


We see that the regression fits a linear line between log transformed Rate and Age. Let us assume model assumptions.

```
model_residuals= resid(lm_model_resp)

#residual versus x
plot(respiratory$Age, model_residuals,
     ylab="Residuals", xlab="Age",
     main="Respiratory Rates")
```



```
#abline(0, 0)                          # the horizontal reference

#residual versus fitted
plot(lm_model)
```



- Linearity: Although the raw form of Rate and Age are not linearly related, the log-transformed Rate and Age show a nearly linear relationship. From the residuals versus x plot, as well, the relationship looks linear.

- Independence and Equal Variance: There is no heteroskedasticity as revealed by the residuals versis fitted plot. It shows no pattern except for a very slight non-linear relationship. The assumption of independence and equal variance is met.

- Normality: The Q-Q plot shows a straight linear relationship. This reinforces the assumption of normality.

Part C: Prediction Intervals

```
# creating data for three individual children: a 1 month old, an 18 months old, and a 29 months old
# for ease of computation, assuming every month has 30 days
age_1 = log10(1 * 30)
age_2 = log10(18 * 30)
age_3 = log10(29 * 30)
newdata2 <- data.frame("Age" = c(age_1, age_2, age_3 ))
pred_interval_rates <- predict(lm_model_resp,newdata=newdata2, interval="prediction", level=0.95)
```

```
pred_interval_rates
```

```
##        fit      lwr      upr
## 1 1.657719 1.489912 1.825526
## 2 1.647357 1.479583 1.815130
## 3 1.645647 1.477878 1.813415
```

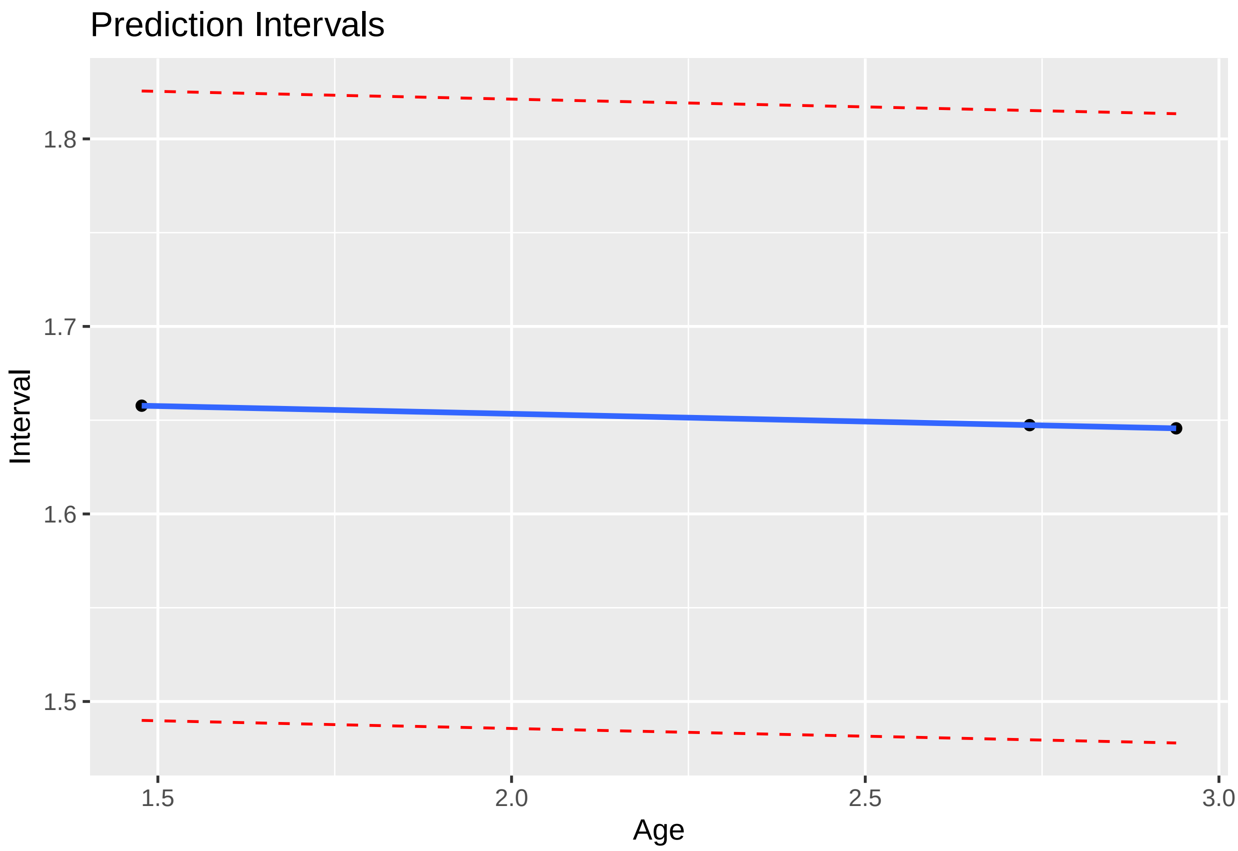
The above snippet gives us the 95% Prediction Interval for each point in the new dataset.

Plotting as in previous question:

```
PI_data = cbind(newdata2, pred_interval_rates)
PI_data
```

```
##        Age      fit      lwr      upr
## 1 1.477121 1.657719 1.489912 1.825526
## 2 2.732394 1.647357 1.479583 1.815130
## 3 2.939519 1.645647 1.477878 1.813415
```

```
ggplot(PI_data, aes(x=Age,y=fit)) + geom_point()+geom_line(aes(y=lwr),color="red",linetype="dashed")+ge
```



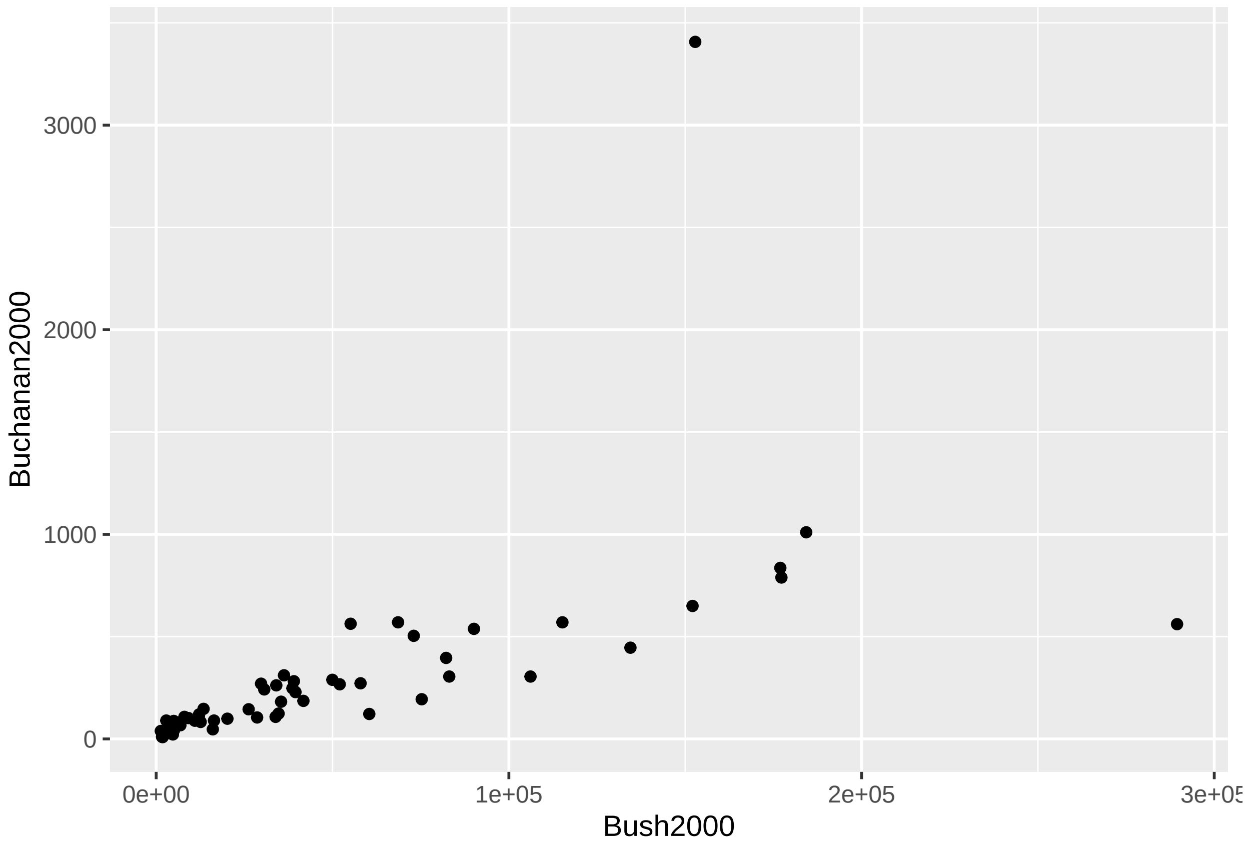
Question 3: THE DRAMATIC U.S. PRESIDENTIAL ELECTION OF 2000

```
#import dataset

elections<- read.csv("C:\\Users\\srish\\Desktop\\books and study material\\Methods and Data Analysis 1\
summary(elections)
```

```
##        X              County    Buchanan2000       Bush2000
##  Min.   : 1.0   Alachua : 1   Min.   :   9.0   Min.   :  1316
##  1st Qu.:17.5   Baker   : 1   1st Qu.:  46.5   1st Qu.:  4746
##  Median :34.0   Bay     : 1   Median : 114.0   Median : 20196
##  Mean   :34.0   Bradford: 1   Mean   : 258.5   Mean   : 43356
##  3rd Qu.:50.5   Brevard : 1   3rd Qu.: 285.5   3rd Qu.: 56542
##  Max.   :67.0   Broward : 1   Max.   :3407.0   Max.   :289456
##                 (Other) :61
```

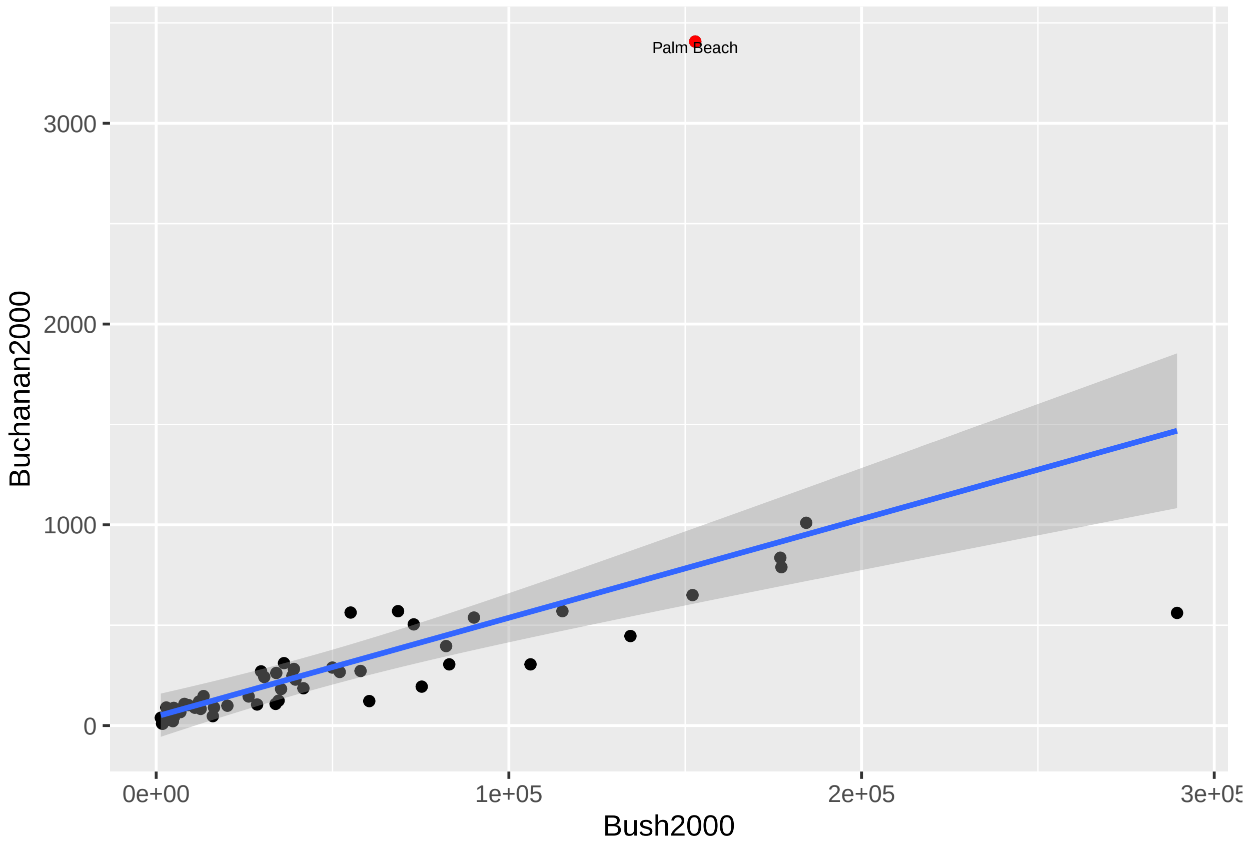
Part A: Scatter Plot of Buchanan2000 and Bush2000

```
#scatter plot
ggplot(data = elections, aes(x = Bush2000 , y =  Buchanan2000)) + geom_point() +
labs(y = "Buchanan2000")
```



Let us highlight the data for Palm Beach County to observe that point and investigate this plot in detail.

```
#subset dataset
elections_pbc <- subset(elections, County == "Palm Beach")

# plot scatter plot to show that Buchanan received more votes than expected in Palm Beach County
ggplot(data = elections, aes(x = Bush2000, y = Buchanan2000)) +
geom_point() +
geom_point(data = elections_pbc , colour="red") +
geom_text(data = elections_pbc, label="Palm Beach", vjust=1, size = 2) +
stat_smooth(method = lm)
```



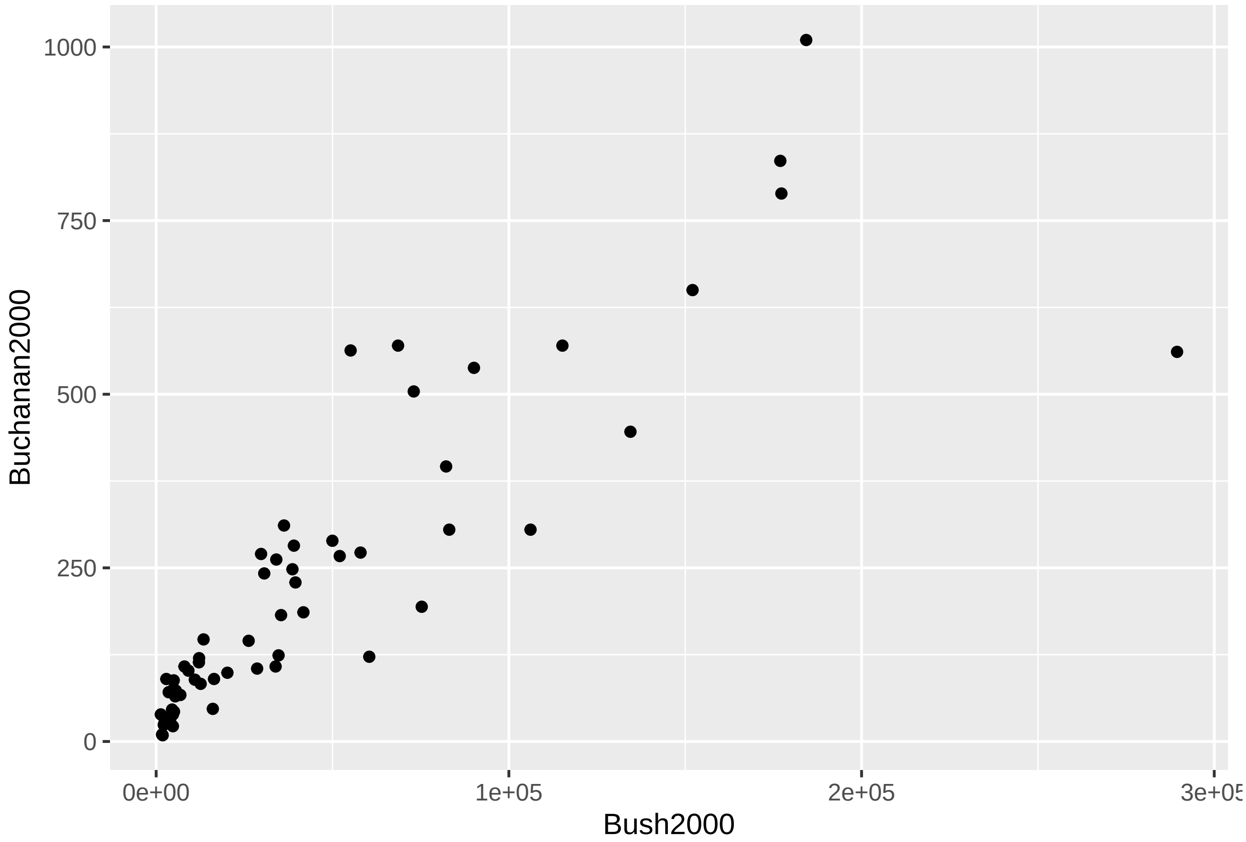
In the above plot, we see that Buchana received ~3,500 votes in Palm Beach County in 2000 which is exceptionally high, as compared to the general range of ~(0,1,000) across other counties. This shows with a high probability that he got more votes than expected in Palm Beach County.

Part B: Regression analysis on data without Palm Beach County

Let us first treat the data.

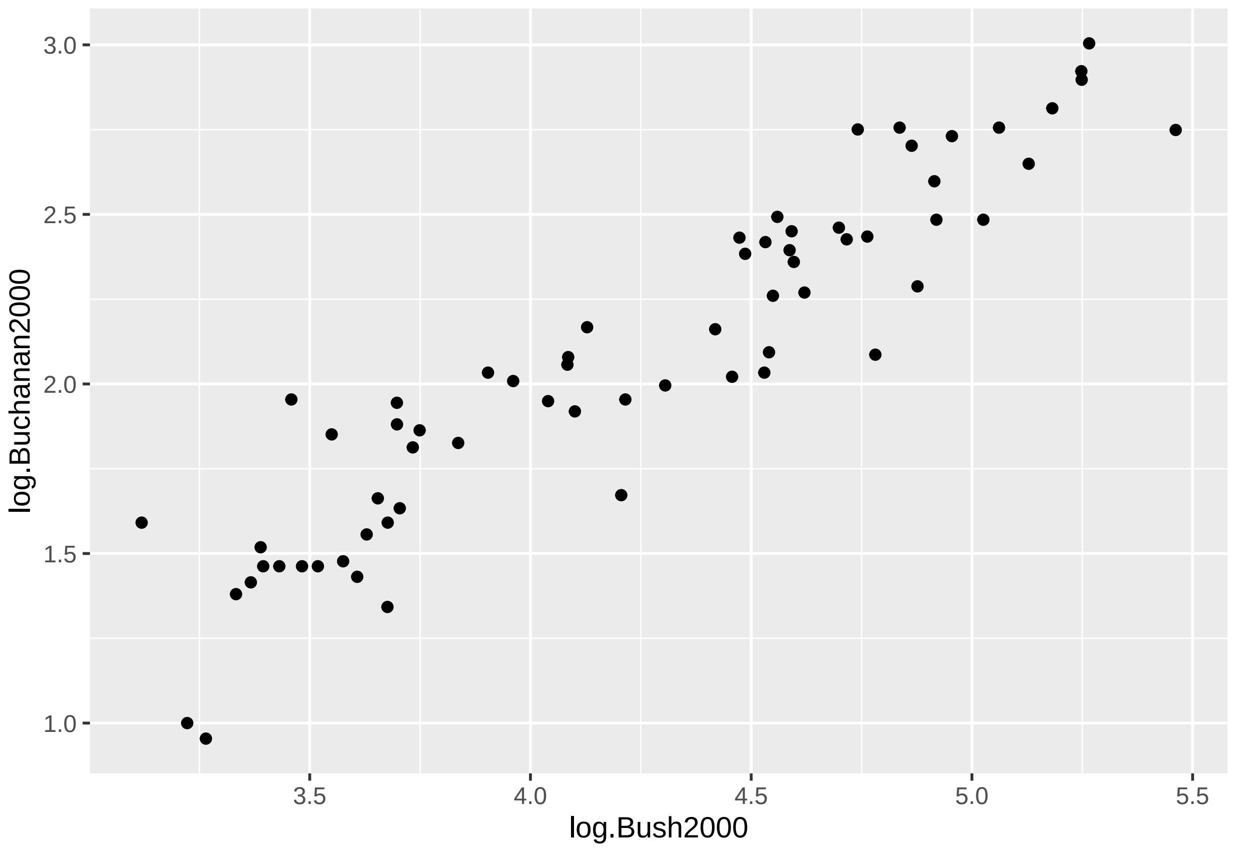
```
#removing row where County= "Palm Beach"
elections_no_pb = elections[!(elections$County=="Palm Beach"),]

# checking distribution of Buchanan2000 and Bush2000
#scatter plot
```

```
ggplot(data = elections_no_pb, aes(x = Bush2000, y = Buchanan2000)) + geom_point() +
labs(y = "Buchanan2000")
```



Due to the shape of the plot it looks like there is a non-linear relationship between the two variables. Hence, applying a log-transformation on these variables

```
#log transformation of both variables
elections_no_pb$log.Bush2000 <- log10(elections_no_pb$Bush2000)
elections_no_pb$log.Buchanan2000<- log10(elections_no_pb$Buchanan2000)

# investigate scatter plot again
ggplot(data = elections_no_pb, aes(x = log.Bush2000, y = log.Buchanan2000)) + geom_point() +
labs(y = "log.Buchanan2000")
```



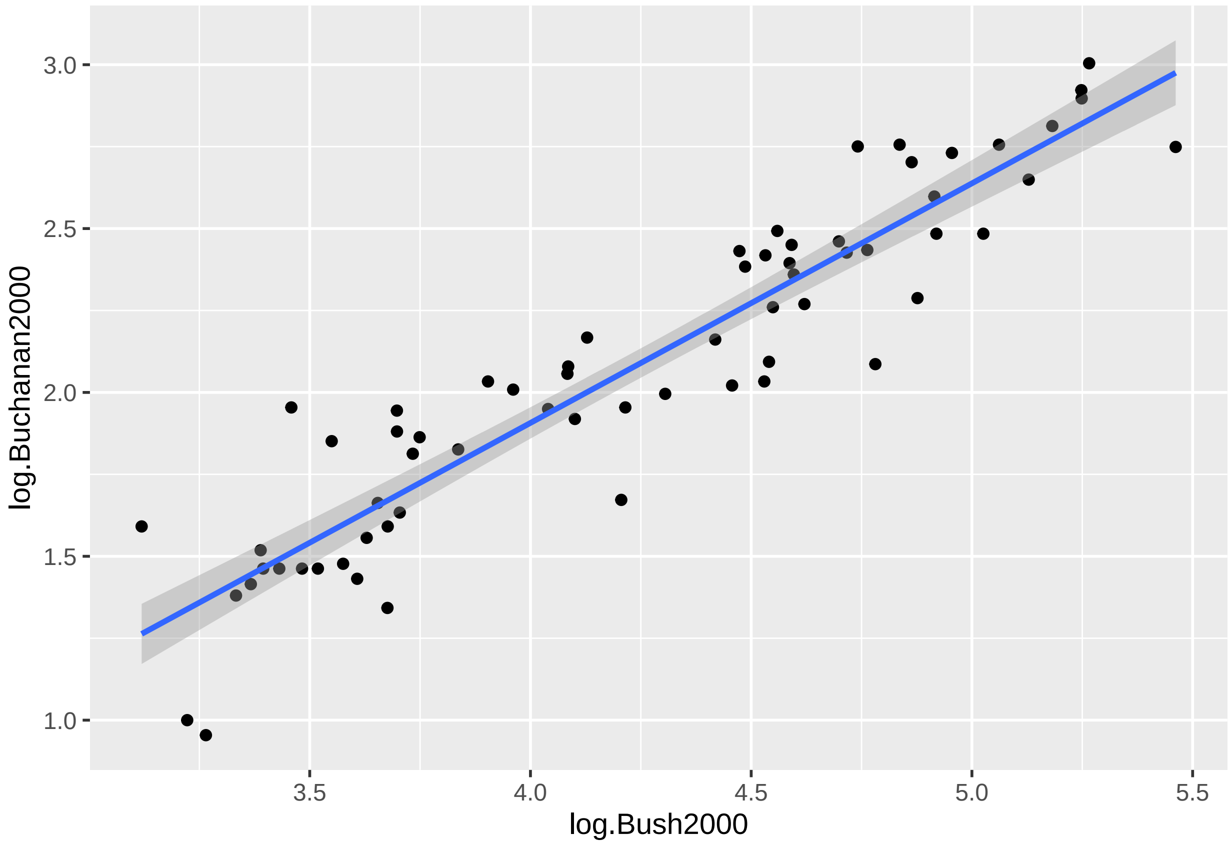
This looks like a linear relationship between the two variables. Now let us fit a model.

```
# regression model for predicting Buchanan votes from Bush votes
lm_model_elections <- lm(log.Buchanan2000~log.Bush2000,data=elections_no_pb);
summary(lm_model_elections)
```

```
##
## Call:
## lm(formula = log.Buchanan2000 ~ log.Bush2000, data = elections_no_pb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41532 -0.09223  0.01087  0.12204  0.44323
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.01689    0.15392  -6.607 9.07e-09 ***
## log.Bush2000  0.73096    0.03597  20.323  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1823 on 64 degrees of freedom
## Multiple R-squared:  0.8658, Adjusted R-squared:  0.8637
## F-statistic:   413 on 1 and 64 DF,  p-value: < 2.2e-16
```

Let us now put this into a scatter plot with a regression line:

```
#scatter plot to observe relationship between log transformed vote counts along with the regression lin
ggplot(data = elections_no_pb, aes(x = log.Bush2000, y = log.Buchanan2000)) +
geom_point() + stat_smooth(method = lm) + labs(x = "log.Bush2000", y = "log.Buchanan2000")
```



The model seems to capture the linear relationship between the log transformed votes for both Bush and Buchanan.

Equation to predict Buchanan's votes from Bush's votes

The general equation will be of the form:

$$log.Buchanan2000_i = \beta_0 + \beta_1 log.Bush2000_i$$

From the above model results, we can see that the equation would be:

$$log.Buchanan2000_i = (-1.017) + (0.731) log.Bush2000_i$$

# Part C: Regression Model and Assumptions

```
#regression results
summary(lm_model_elections)
```

```
##
## Call:
```

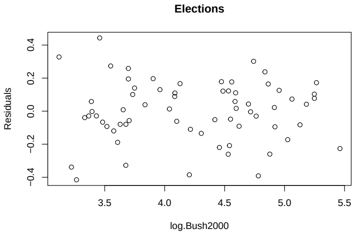
```
## lm(formula = log.Buchanan2000 ~ log.Bush2000, data = elections_no_pb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41532 -0.09223  0.01087  0.12204  0.44323
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.01689    0.15392  -6.607 9.07e-09 ***
## log.Bush2000  0.73096    0.03597  20.323  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1823 on 64 degrees of freedom
## Multiple R-squared:  0.8658, Adjusted R-squared:  0.8637
## F-statistic:   413 on 1 and 64 DF,  p-value: < 2.2e-16
```

Interpret model results

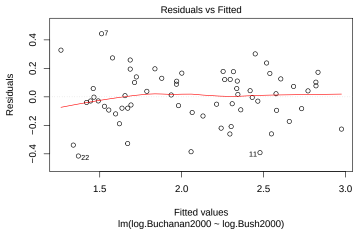
1. The log transformed Bush2000 variable is significant to predict Buchanan's votes, purely on the basis of the p-value (e^{-16}). The estimate (β_0) is 0.731 which means that just on the basis of Bush's votes, having any and all other variables constant, Buchanan will have 0.73 log(votes) greater than Bush. **Kindly note this is a log transformed relationship.**
2. The R-squared is 0.8658 while the adjusted R-squared (after penalizing for extra variables if any) is 0.8637. This explains the portion of variation in log.Buchanan2000 explained by log.Bush2000, which happens to be ~86% in this model. Thus, we have a model that explains a fair amount of the variation in the y-variable.
3. The residual standard error is 0.1823. This indicates how close is our prediction to the true value. A residual standard error of 0 indicates perfect predictions. We have an error of 0.1823 which is fair low.
4. Observing the F-statistic can help in determining whether or not we accept our null hypothesis (that the 2 variables are not related). Our model has a very low p-value of the order (e^{-16}) which rejects the null hypothesis. Thus, the votes received by the two candidates is related, thus justifying our model.

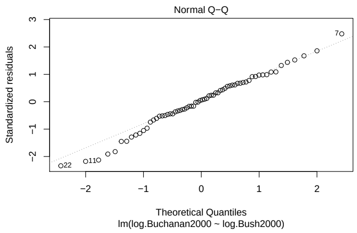
Let us now look at regression plots.

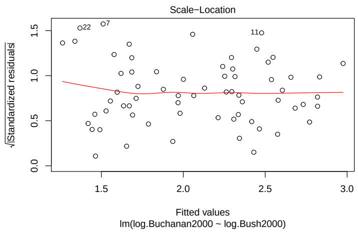
```
model_resid_elect= resid(lm_model_elections)

#residual versus x
plot(elections_no_pb$log.Bush2000, model_resid_elect,
     ylab="Residuals", xlab="log.Bush2000",
     main="Elections")
```

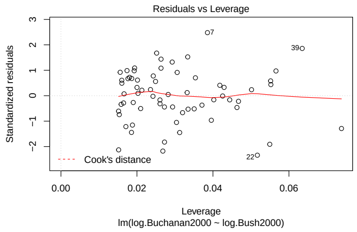


```
#abline(0, 0)                   # the horizontal reference

#residual versus fitted
plot(lm_model_elections)
```









The residual plot shows how accurately our model fits the data. Here, the residuals are roughly centered around 0 with no evident pattern (almost linear). So, the model is a good fit to the data.

The residual versus x plot also shows a linear relationship, thus satisfying the assumption of Linearity. The residual versus fitted plot confirms no evident pattern (almost linear) and thus proves independence and equal variance. The Q-Q plot proves normality with a nearly linear trend.

Part D: 95% prediction interval

```
#prediction for the number of Buchanan votes in Palm Beach
# number of votes for Bush from Palm Beach County in 2000: 152846
#log transform
log.Bush_pbc = log10(152846)
#create data frame
newdata3 <- data.frame("log.Bush2000" = log.Bush_pbc)

#pred interval
pred_interval_el <- predict(lm_model_elections, newdata=newdata3, interval="prediction", level = 0.95)
pred_interval_el
```

```
##        fit      lwr      upr
## 1 2.772598 2.399328 3.145869
```

Since these results are for the log-transformed vote count, let us get the actual number of votes for Buchanan:

```
#from the prediction interval results
Buchanan_true_predicted_val = 10^pred_interval_el[1]
Buchanan_pred_range_min = 10^pred_interval_el[2]
Buchanan_pred_range_max = 10^pred_interval_el[3]
```

```
data.frame("Buchanan_tru_pred" = Buchanan_true_predicted_val,"Buchanan_min_pred"=Buchanan_pred_range_mi
```

```
##   Buchanan_tru_pred Buchanan_min_pred Buchanan_max_pred
## 1          592.3769          250.8001          1399.164
```

Now, from the above results, we see that the minimum of the range of votes predicted for Buchanan is 250.8 (~251) while the maximum is 1399.16. Assuming that some of Gore's votes were mistakenly counted as/added to Buchanan's, this means that we can use the predicted votes for Buchanan and estimate for Gore. In a scenario if and only if the additional votes were Gore's, he should have receives at least 251 (from 250.8) votes and at most 1399 more (on top of his actual number of votes) votes than he actually got. The model predicts ~592 votes for Buchanan (in the range with 95% confidence). The 95% prediction interval says that we can say with 95% confidence that the true predicted value of Buchanan's vote will lie in the range (251,1399). The other votes might actually be miscounted.