STAT 252 R3

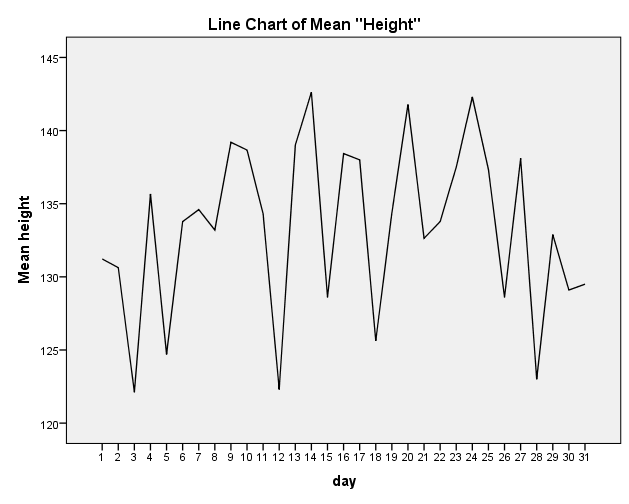
LAB 3

DONG, Boyuan

STAT 252 Lab3

1.

(a)



Trend: The height is raising first half month and then decreasing later half month.

Overall no distinct trend.

Typical mean height: The typical mean is approximately 130 height.

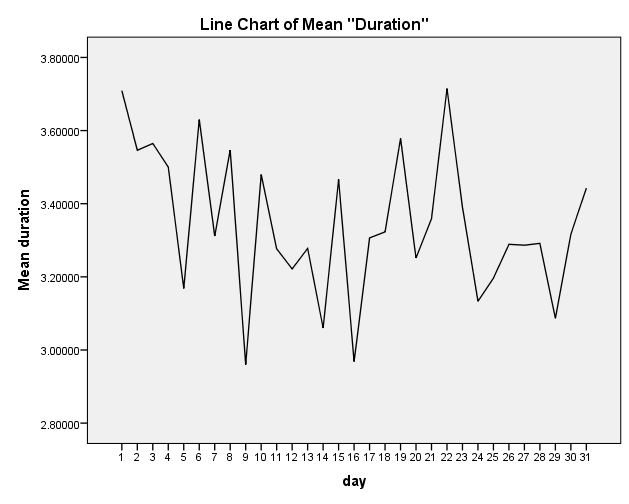
Mean height variation: There is noticeable variation above and below 130 height, ranging from about 122 height to 143 height.

Height distortion: The daily weather might affect the height of eruptions.

(b)

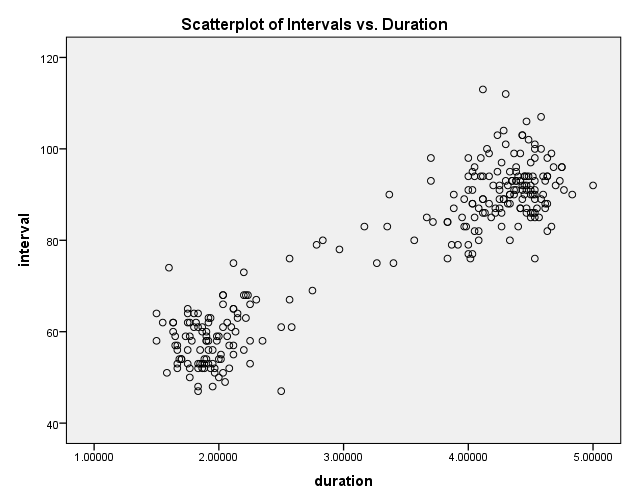
Trend: There’s no distinct trend overall.

Typical duration: The typical duration is approximately 3.30000 minutes ranging from 2.95000 to 3.75000 minutes.



2

(a)



(b)

Interval between eruptions:

Shape: The data are most centered when duration below 3.00000 minutes above 1.00000 minutes and waiting time is blow 80 minutes above 40 minutes, also when duration above 4.00000 minutes below 5.00000 and waiting time is above 80 minutes below 120 minutes.

Overall the waiting time is increasing as the duration time is increasing.

Center: about at 80 minutes

Variability: There is noticeable variation above and below 80 minutes interval, ranging from about 40 minutes to 120 minutes.

Duration of eruptions:

Shape: The data are most centered when duration below 3.00000 minutes above 1.00000 minutes and waiting time is blow 80 minutes above 40 minutes, also when duration above 4.00000 minutes below 5.00000 and waiting time is above 80 minutes below 120 minutes.

Overall the waiting time is increasing as the duration time is increasing.

Center: about at 3.20000 minutes

Variability: There is noticeable variation above and below 3.00000 minutes, ranging from about 1.00000 minutes height to 5.00000 minutes.

Regularity of old faithful’s eruptions:

The longer waiting time is the longer duration it eruptions.

(c)

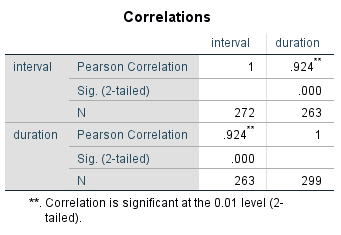
Pattern: There seems to be a positive linear increasing pattern, but it is quite weak and definitely spread out, indicating a positive linear model may not be ideal for this relationship between the two variables.

Outliers: There are outliers in the plot.

3

(a)

We can see that the correlation coefficient value is 0.924 (significant).

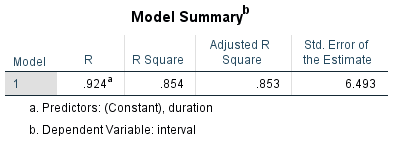


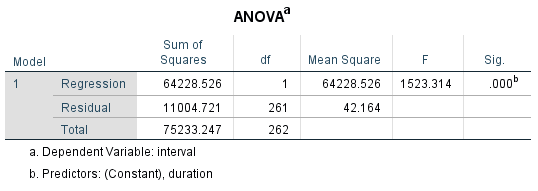
(b)

The sign and magnitude of the coefficient is consistent with the scatterplot we’ve seen in Question 2, as we have a positive linear relationship between the two variables the correlation coefficient value is quite strong.

4

(a)

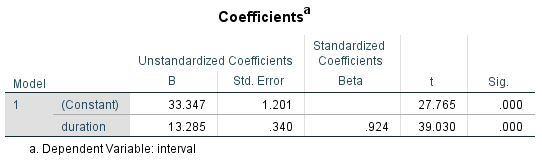


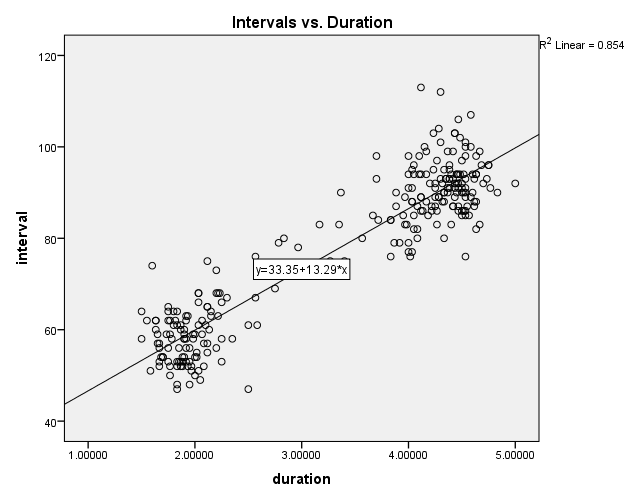


Model Assumptions:

1. Linearity: There is a real linear relationship between the interval and duration.
2. Constant Variance: The spread of the interval about the line is the same at all levels of the duration.
3. Normality: The sub-populations of interval at different values of duration all have a normal distribution.
4. Independence: All intervals are independent of each other.

(b)





Equation of the least-squares line:

Intervals= β0+β1Duration+Ɛ where Ɛ represents error, assumed to be a random variable that follows a normal distribution with mean zero and constant variance for all values of Duration.

µ (Intervals | Durations) = β0+β1Duration+Ɛ or Intervals=33.347+13.285\*Durations

Meaning of the slope:

It means the effect of duration on the interval.

The effect of duration on the interval is measured as the change in the mean interval that is associated with as one-unit increase in that variable holding all other explanatory variables fixed.

Quality of the fit:

There seems to be a linear increasing pattern, and it is quite strong not overly spread out, it can fit a straight line easily by visualizing. So it fit quite good.

Outliers: There are outliers.

(c)

Duration explain 85.4% of the variability in the time between eruptions.

Utility of the linear regression:

Based on the observed data, we can have the estimation and estimated regression equation. Once the model is estimated, we can use it to estimate mean responses or predict single responses at certain values of the explanatory variable.

(d)

µ (Intervals | Durations) = β0+β1Duration+Ɛ =33.347+13.285\*Durations and inserting Durations=2 we have:

Estimate =33.347+13.285\*2=59.917 when Durations=2

95% confidence interval:

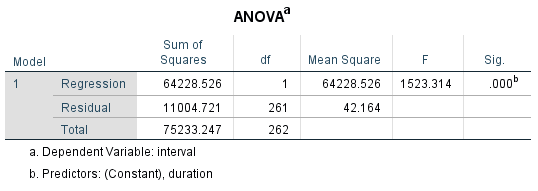
(58.72984, 61.10667)

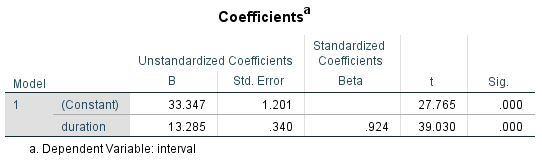
95% prediction interval:

(47.07711, 72.75940)

The prediction interval is wider and nonsensical for this case.

(e)





H0: β1=0 versus Ha: β1≠0

Test statistic: The test statistic follows a t-distribution with 261 degrees of freedom

Test statistic is 39.030

P-value: 0.000

There is convincing evidence that there is a significant linear relationship between the Intervals and Durations by the day.

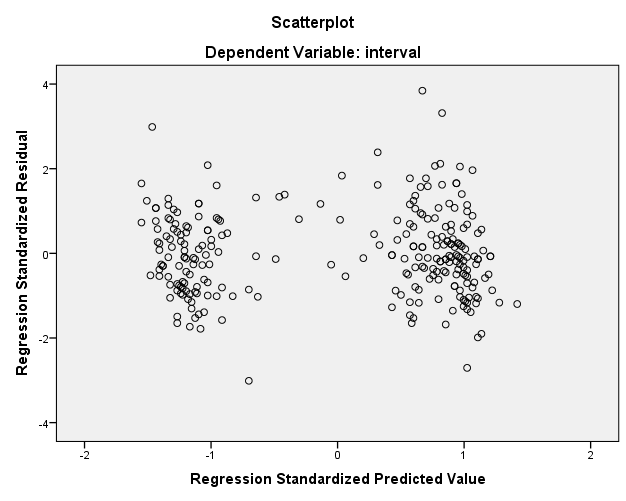
F=1523.314=t2=39.0302 and P-value is exactly the same as with the t-test. The test statistic follows an F-distribution with 1 degree of freedom for the numerator and 261 for the denominator

Sum of squares residuals for each model:

SSR(e) =11004.721

SSR(a) =75233.247

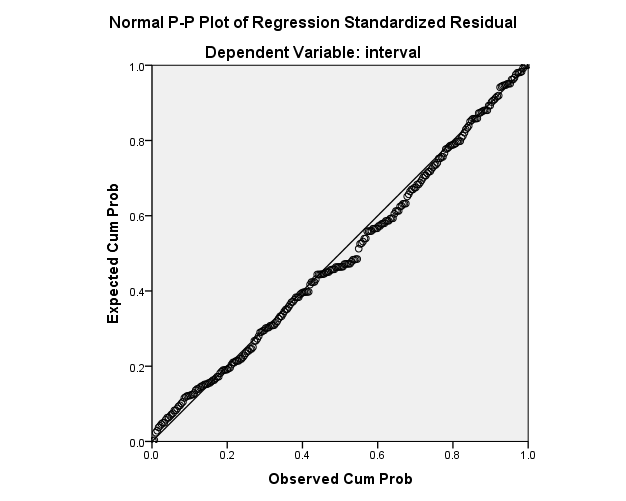
(f)



Pattern: no pattern.

The values of the standardized residuals are greater in absolute value. Thus, variance decreases as the regression predicted values increase when the regression standardized predicted value less than 0, variance increases as the regression predicted values increase when the regression standardized predicted value greater than 0, suggesting severe violation of the assumption of constant variance.

(g)

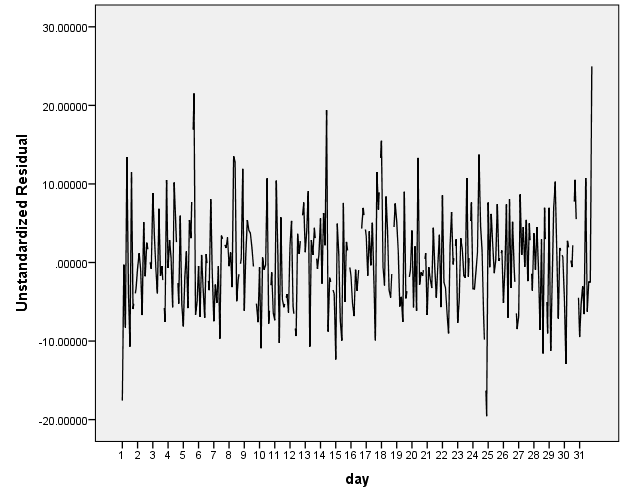


Yes, the assumption of normality for the response variable seem appropriate.

The plots fairly fit the normal distribution.

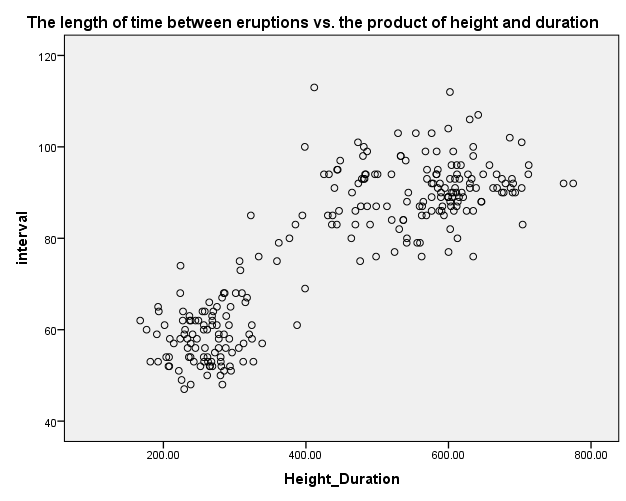
(h)

There does not appear to be a trend in time, so the relationship incorporated in the regression model is consistent over time.



5

(a)



(b)

Pattern: weak positive linear pattern.

Outliers: There are outliers.

(c)

µ (Intervals | Height\*Duration) = β0+β1 (Height\*Duration) + Ɛ

=38.656+0.088\*(Height\*Duration)

R2: 0.738

R2 of Q3 is 0.854, R2 of this new model is 0.738. So R2 of Q3 is more close to 1, which means that the model in Q3 fits the line better than the new model, thus there are less error in using explanatory value to predict response value. So the model in Q3 is better.

