assignment3

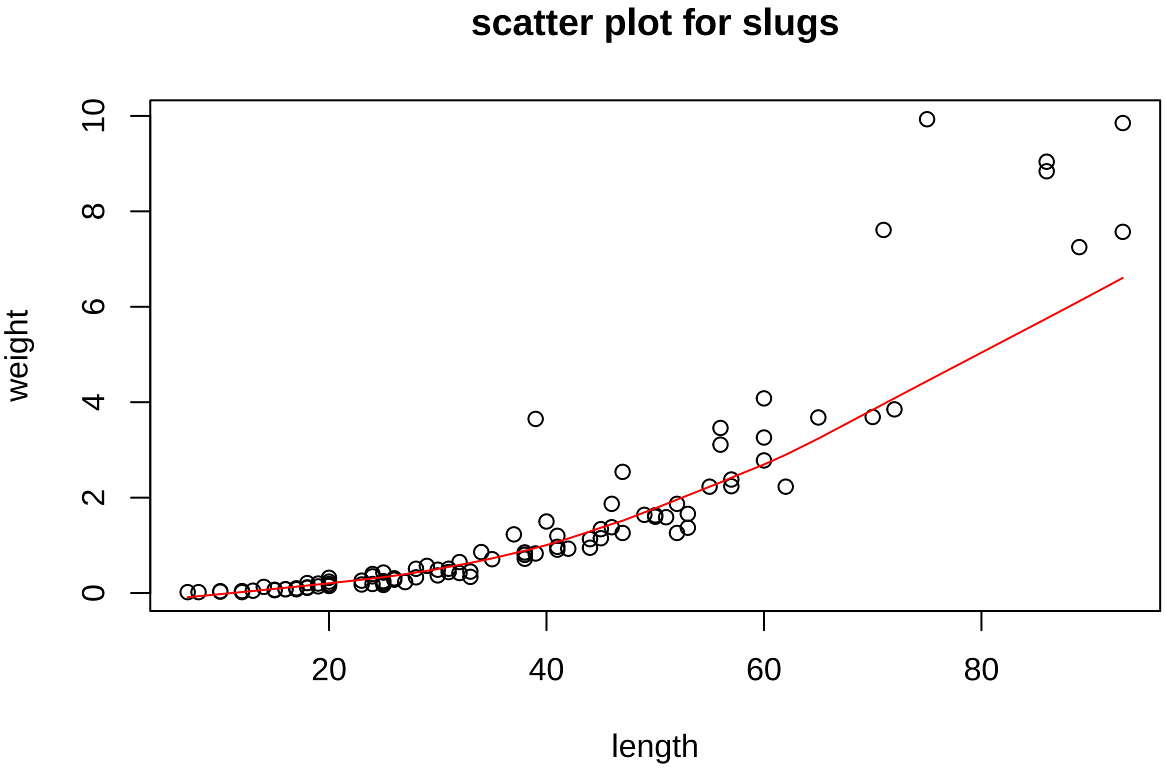
Frank Guo

2024-05-10

Task 1: Plotting (8 Marks)

a) Generate a scatterplot of the data. Ensure that you have an informative title and axis labels in your plot.

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```



b) Describe the relationship between slug lengths and weights with respect to direction and shape (you may add a LOWESS smoother to your plot to help you assess the relationship if you wish).

Answer:

As we can see in the plot,

Direction: Positive (as the x-variable increases, the y-variable tends to increase).

Shape:the relationship between weight and length looks straight. But the variance increases when x-variable increaeses that is Heteroscedasticity.

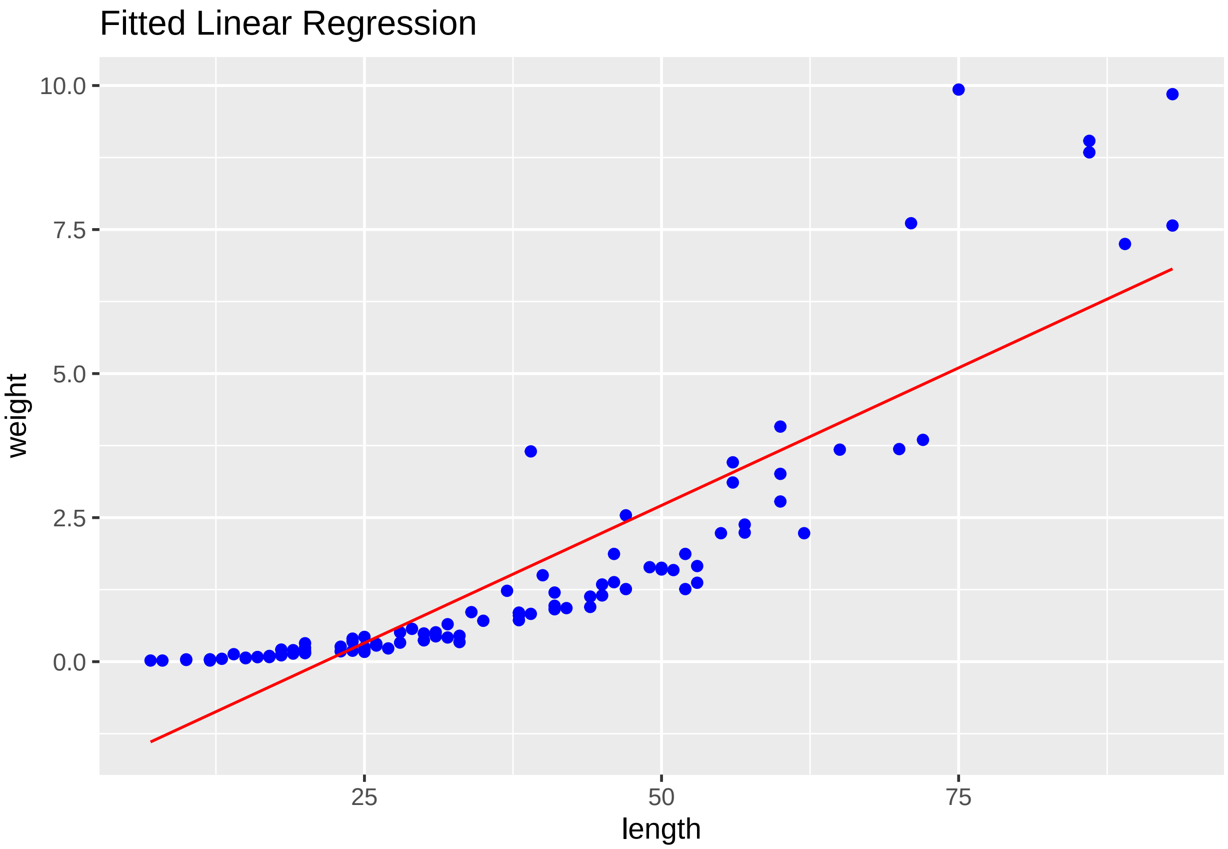
Strength: data points close to lying on a smooth line when x is small, while it strength becomes weaker as x increases.

Task 2: Straight-Line Model (20 Marks)

a) Use the lm() function to fit a straight-line model between weight and length. Use the summary() function to generate the model output and also generate a new scatterplot and plot the straight-line relationship (you may find the abline() function helpful).

Because the plot and abline function in R can't display the intercept correctly, I use the ggplot2 package instead.

```
##
## Call:
## lm(formula = weight ~ length, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6434 -0.8358 -0.1391  0.5209  4.8305
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.061632   0.226257  -9.112 1.02e-14 ***
## length       0.095482   0.005343  17.871  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.083 on 98 degrees of freedom
## Multiple R-squared:  0.7652, Adjusted R-squared:  0.7628
## F-statistic: 319.4 on 1 and 98 DF,  p-value: < 2.2e-16
```



b) State and interpret the straight-line regression equation (in terms of the variables in the dataset) based off the summary output.

Answer:

According the summary, the Equation will be:

```
Estimated Mean weight = -2.061632 + 0.095482 * length.
```

Intercept -2.061632:

The expected (average) value of weight is -2.061632 is when length is 0.

Slope 0.095482:

The expected (average) increase in weight is 0.095482 per unit increase in length.

c) State the value of the R-squared statistic and give the correct interpretation.

Answer:

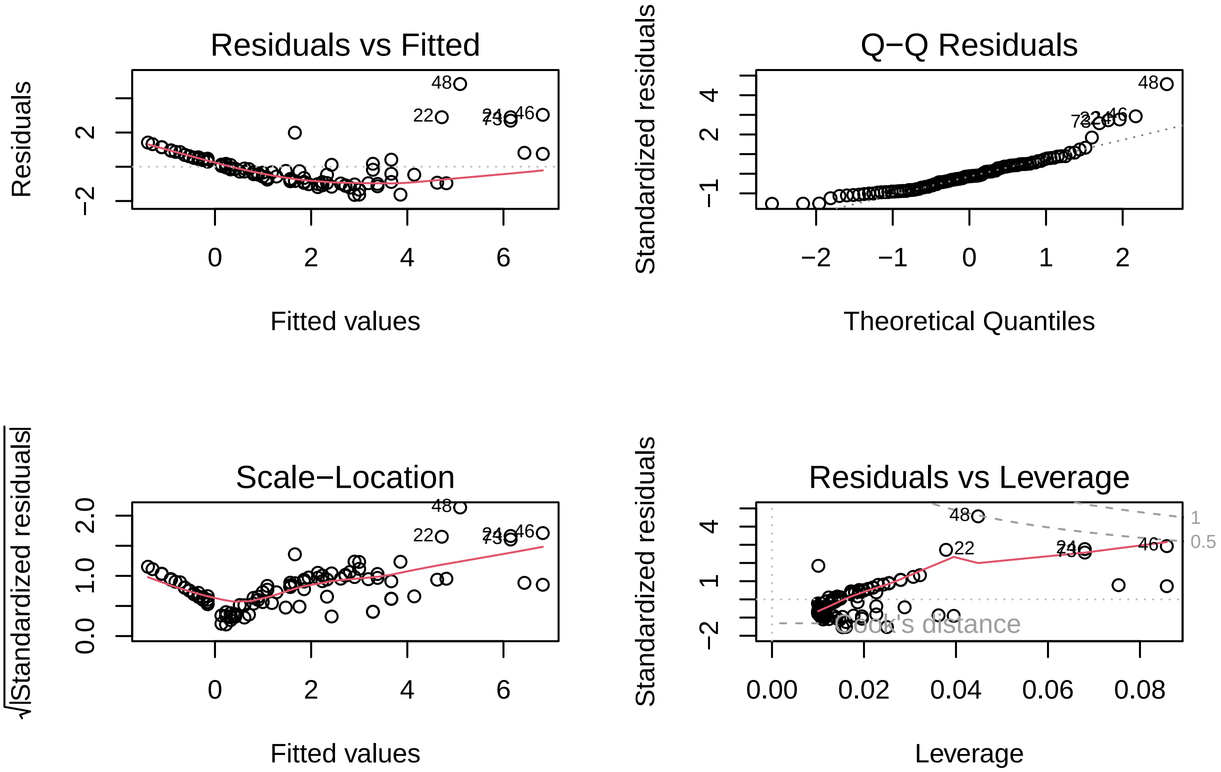
The value of the R-squared is 0.7652. that means 76.52% of variation in the weight is explained by the regression model on the length.

d) From the plot, does it look as though the straight-line model is an appropriate model for the data? Explain why or why not.

Answer:

From the plot of lm fit. it is not a good model for the data. All the points having length less than 21.6 will be predicted as negative on the mean weight. And the residual of points on the right side,looks very big.

e) Assess the error assumptions in the regression model. Paste your outputs in the report. Is the straight-line model appropriate for modelling the relationship?



```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lm_fit)
## W = 0.888, p-value = 4.02e-07
```

Answer:

Regarding the error assumptions in the regression model:

all the errors are independent(not obvious patten in errors).

The errors come from a common Normal distribution with mean 0 and standard deviations.

I use residual plots to check the assumptions:

In the Residuals VS Fitted plot, residuals look not constant, it looks biger in the right. and there is a perfect curve pattern from left to right.Also we can find the non-linear pattern in the scale-location plot.

I use shapiro test to check whether the residuals plausibly come from a Normal distribution, got p-value = 4.02e-07, we have strong evidence against the assumption the residuals were sampled from a Normal distribution!

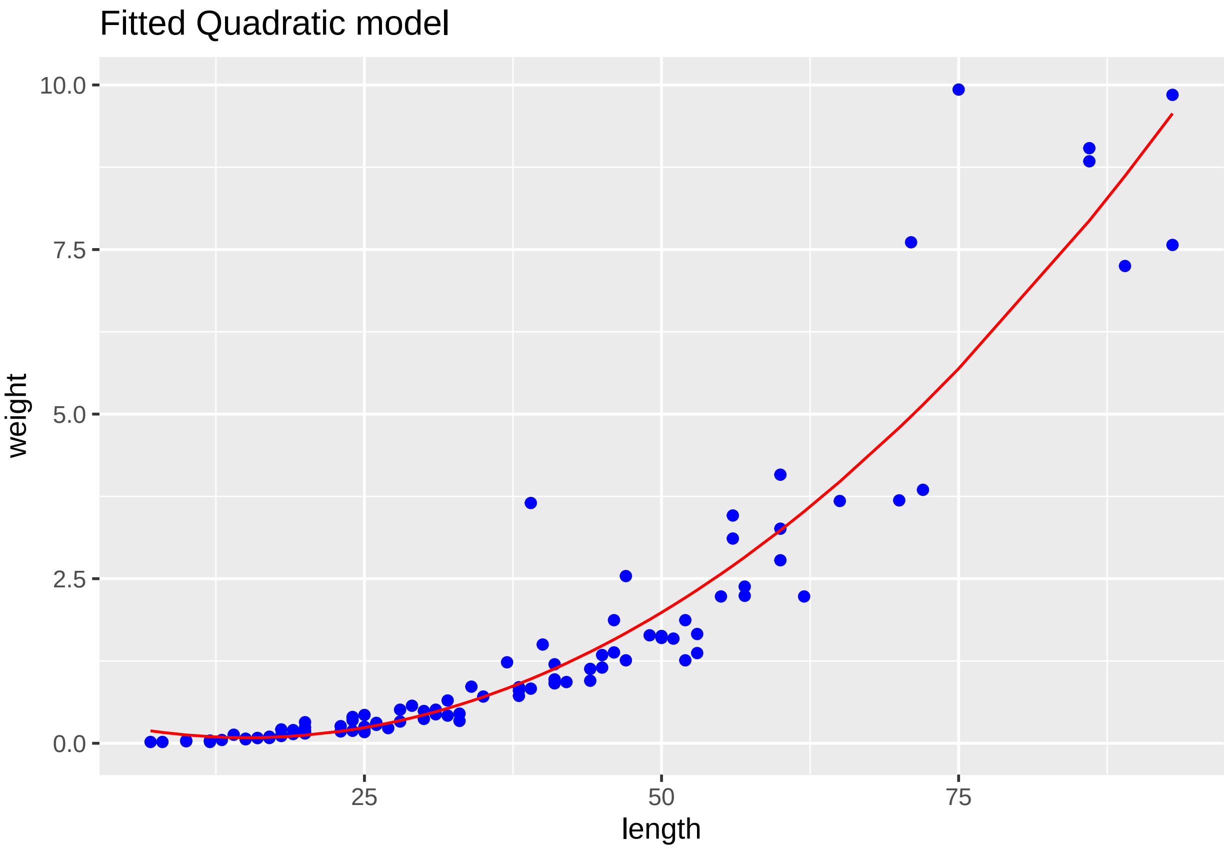
To summrise, I don't think straight-line model is appropriate for modeling the relationship.

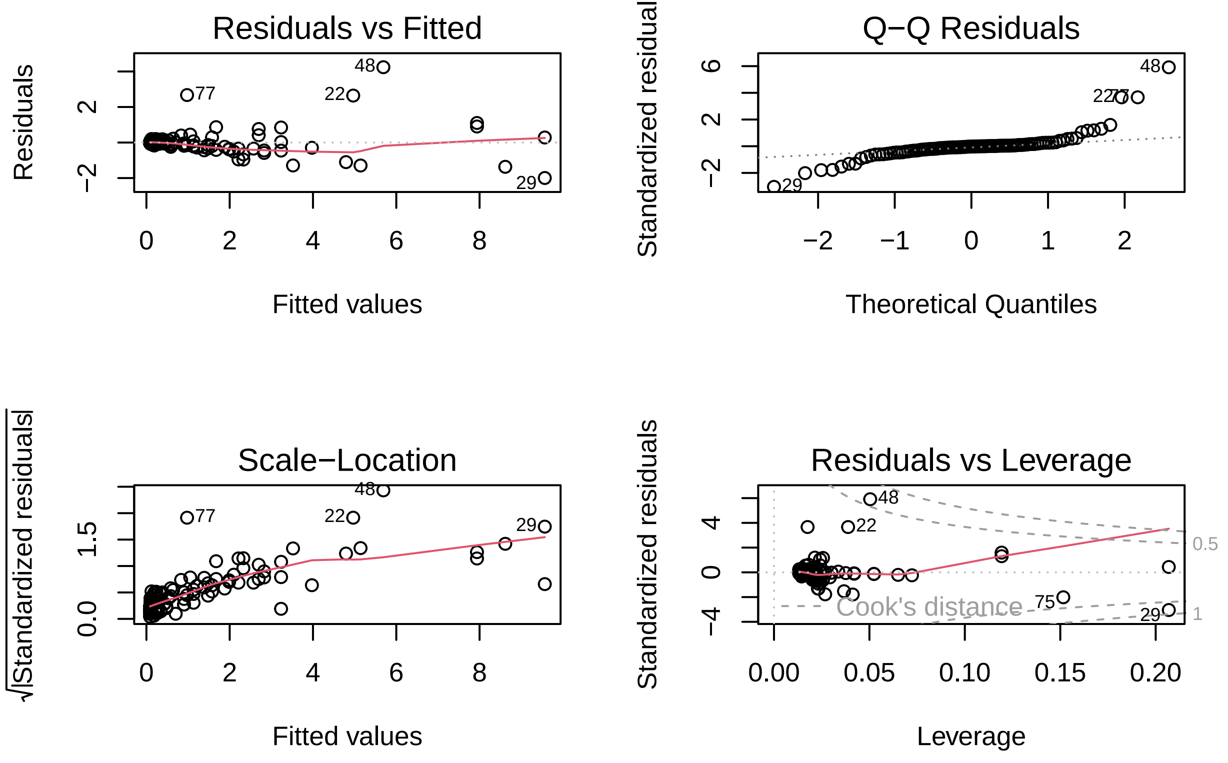
Task 3: Quadratic Model (20 Marks)

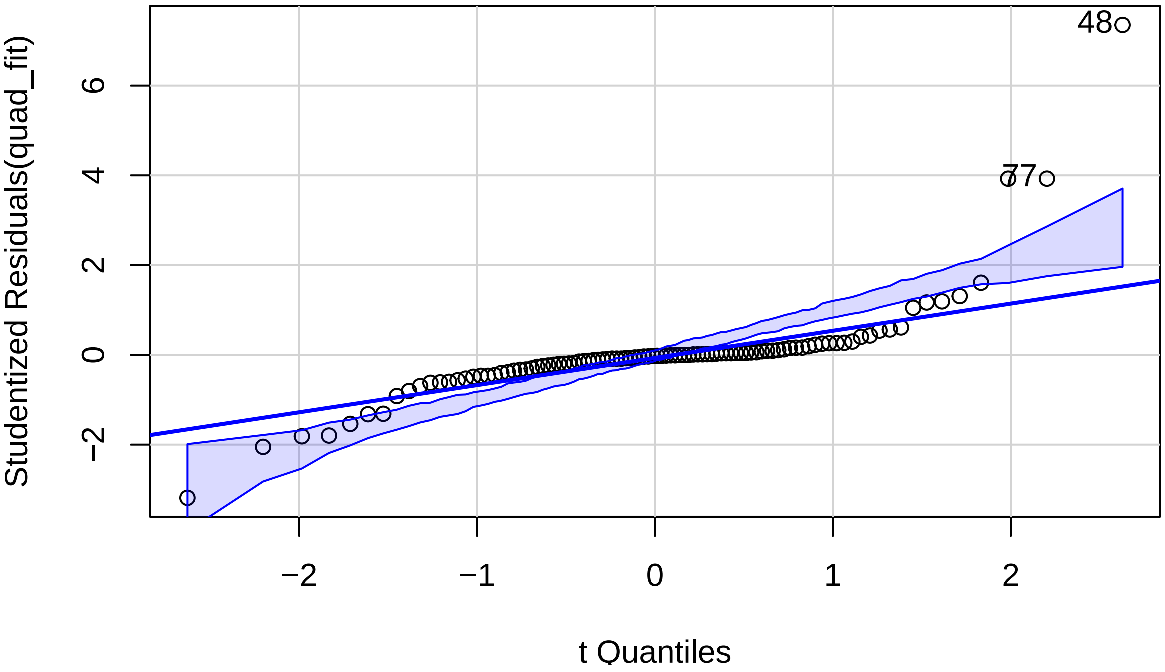
a) Use the lm() function to fit a quadratic model between weight and length. Use the summary() function to generate the model output and also generate a new scatterplot and plot the quadratic relationship.

Answer:

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

##
## Call:
## lm(formula = weight ~ length + I(length^2), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9963 -0.2025 -0.0173  0.0701  4.2399
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4433687  0.2795337   1.586 0.115971
## length      -0.0472923  0.0137915  -3.429 0.000891 ***
## I(length^2)  0.0015633  0.0001457  10.731  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7361 on 97 degrees of freedom
## Multiple R-squared:  0.8926, Adjusted R-squared:  0.8904
## F-statistic: 403.3 on 2 and 97 DF,  p-value: < 2.2e-16
```


Fitted Quadratic model
weight
10.0
7.5
5.0
2.5
0.0
25
50
75
length


Residuals vs Fitted
Residuals
2
−2
48
22
77
29
0
2
4
6
8
Fitted values
Q−Q Residuals
Standardized residuals
6
2
−2
48
2277
29
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
1.5
0.0
48
22
77
29
0
2
4
6
8
Fitted values
Residuals vs Leverage
Standardized residuals
4
0
−4
48
22
75
29
0.5
1
Cook's distance
0.00
0.05
0.10
0.15
0.20
Leverage



```
## [1] 48 77
```

b) State and interpret the quadratic regression equation (in terms of the variables in the dataset) based off the summary output.

Answer:

The quadratic regression equation is:

Estimated Mean weight = 0.4433687 - -0.0472923 × length + 0.0015633 × length^2.

While the intercept is 0.4433687 here, just for mathematics use, not meaningful.

c) From the plot, does it look as though the quadratic model is an appropriate model for the data? Explain why or why not.

Answer:

From the plot, it looks fit the data much better than the straight line. All the meaningful data is in the meaningful range of fitted model compared to the straight line. The points spread symmetrically on the both side of the fitting line. Trend is correct.Looks like an appropriate model for the data.

d) Use the Adjusted R-squared and residual standard error to compare the fits of the quadratic model vs the straight-line model. What can you conclude?

Answer:

For the straight-line model:

Adjusted R-squared: 0.7628.

For the quadratic model:

Adjusted R-squared: 0.8904

The Adjusted R-squared value of the quadratic model is much larger than the straight-line model's. That means the quadratic model fits the data better than the straight-line model.

e) Assess the error assumptions in the regression model. Paste your outputs in the report. Is the quadratic model appropriate in describing the relationship?

Answer:

Regarding the error assumptions:

In the Residuals VS Fitted plot: Scatter is constant except for a couple of points (possible outliers). the smoothing line is quite flat.

In the QQ Plot and Normality test:

The residuals do not look like they come from a normal distribution.

Scale-location plot.

Smoothing line is shows positive trend, maybe exist a pattern that the variance increased with weight. Our fitting is not perfect.

Residuals vs. leverage:

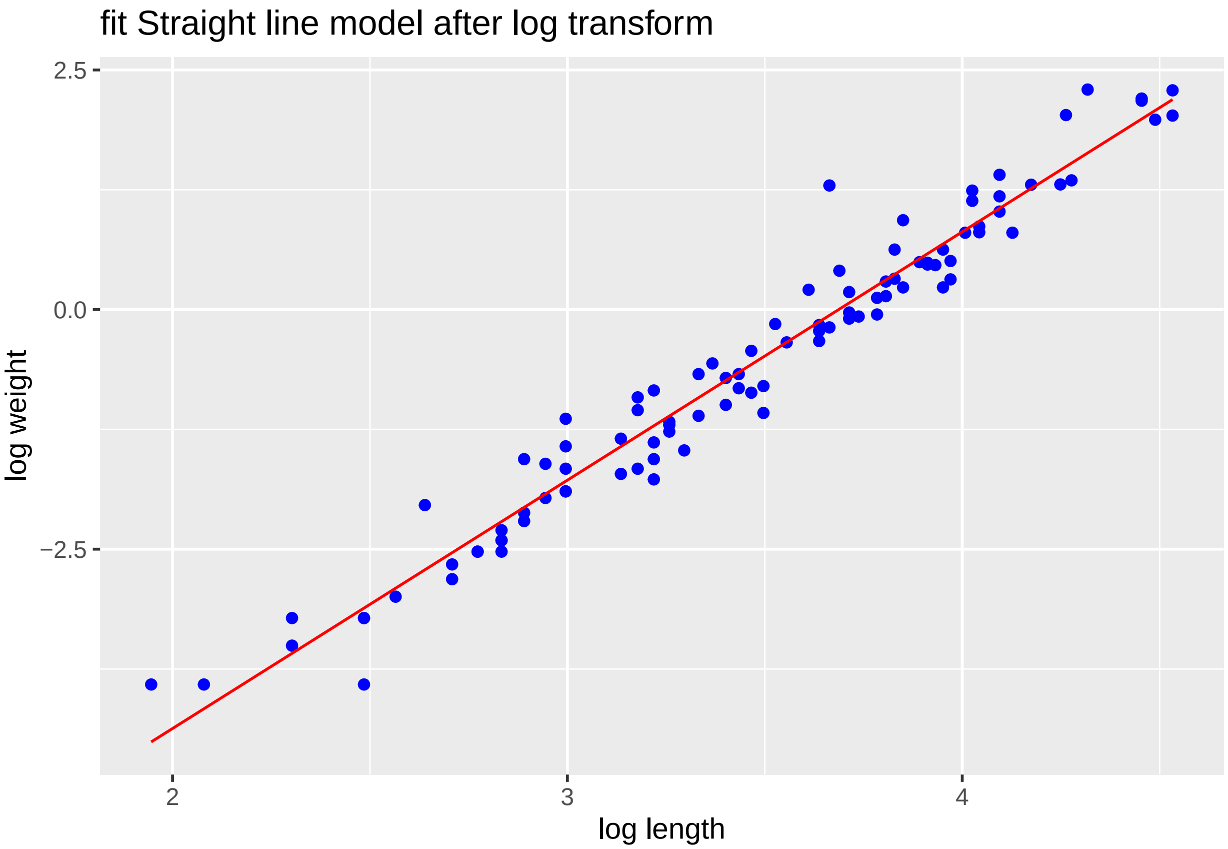
There are some points like 48,29,75 has big residual like outliers which have high influence to the model, but infact, the data is meaningful.

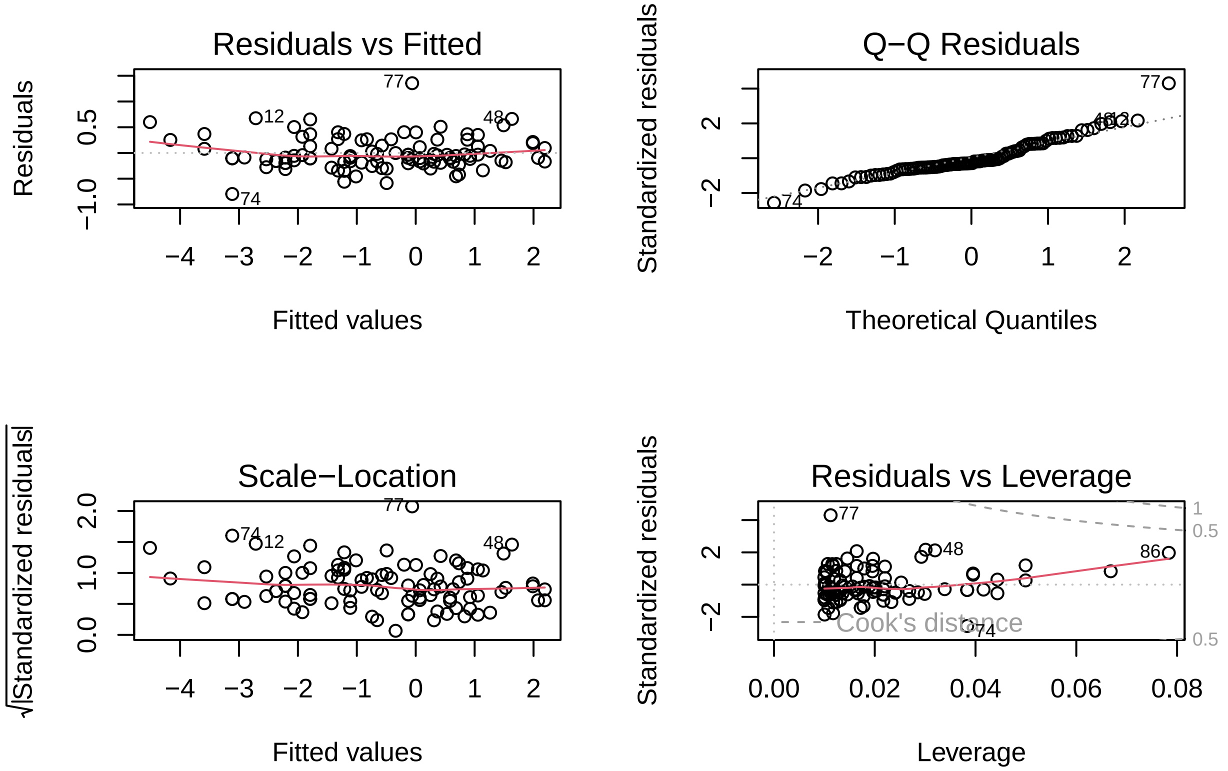
In summary, thought the model looks fitting the data well. still can't explain the heteroscedasticity of the error. It is not very appropriate in describing the relationship.

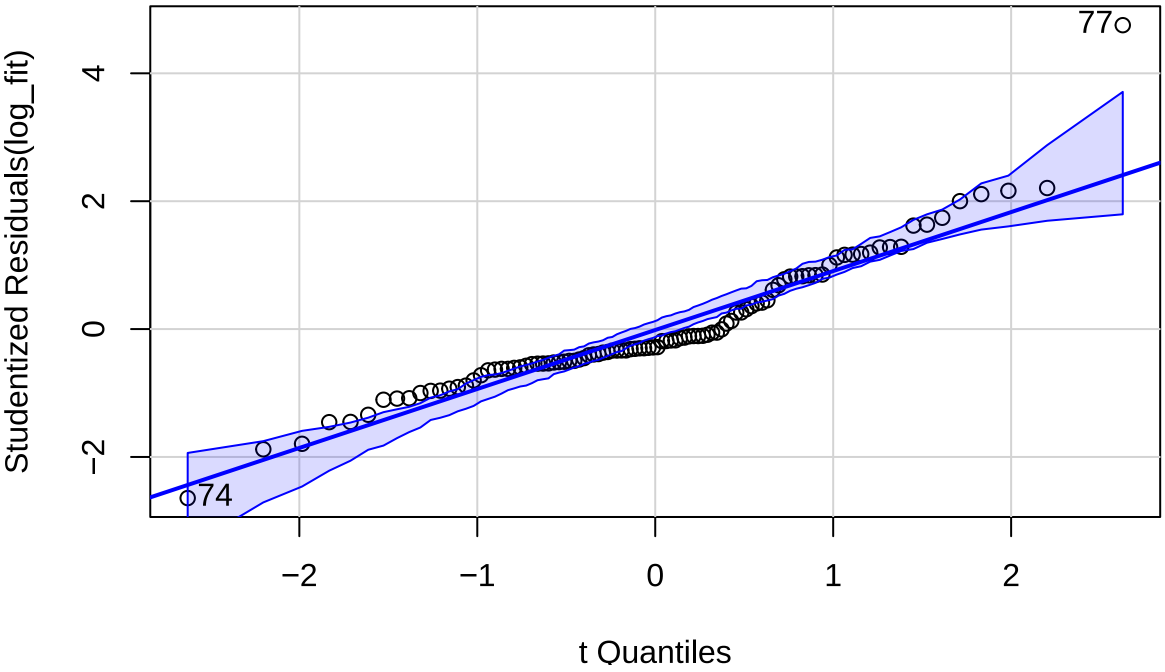
Task 4: Variable Transformation and Inference (32 Marks)

It was decided by the researchers to transform both variables into (natural) log-scale. Use the log() function in R to create two new variables called log weight and log length. ### a) Use the lm() function to fit a straight-line model between log weight and log length. Use the summary() function to generate the model output and also generate a new scatterplot and plot the straight-line relationship. ###Answer:

```
##
## Call:
## lm(formula = logweight ~ loglength, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7972 -0.1713 -0.0893  0.1966  1.3555
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.55359    0.19196  -49.77   <2e-16 ***
## loglength    2.59115    0.05472   47.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3168 on 98 degrees of freedom
## Multiple R-squared:  0.9581, Adjusted R-squared:  0.9577
## F-statistic:  2242 on 1 and 98 DF,  p-value: < 2.2e-16
```


fit Straight line model after log transform
log weight
2.5
0.0
−2.5
2
3
4
log length


Residuals vs Fitted
Residuals
0.5
−1.0
77
12
48
74
−4
−3
−2
−1
0
1
2
Fitted values
Q−Q Residuals
Standardized residuals
2
−2
77
48
12
74
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
2.0
1.0
0.0
77
74
12
48
−4
−3
−2
−1
0
1
2
Fitted values
Residuals vs Leverage
Standardized residuals
2
−2
77
48
86
74
Cook's distance
1
0.5
0.5
0.00
0.02
0.04
0.06
0.08
Leverage



```
## [1] 74 77
```

b) State and interpret the regression equation (in terms of the variables in the dataset) based of the summary output. Also provide the equation in the original scale (by back transforming).

Answer:

The Equation is : mean Log(weight) = -9.55359 + 2.59115 * log(length)

let's do the transform: median weight = exp(-9.55359 + 2.59115 * log(length)) = exp(-9.55359)* exp(2.59115 * log(length))

if the log weight is assumed to be normal distribute, then mean(log weight) = median(log weight), then mean weight = = exp(-9.55359)* exp(2.59115 * log(length))

c) Use the Adjusted R-squared and residual standard error to compare the fit of this model against the quadratic model and the straight-line model. What can you conclude?

For the straight-line model:

Adjusted R-squared: 0.7628.

For the quadratic model:

Adjusted R-squared: 0.8904

After log transformed: Adjusted R-squared: 0.9577

So we conlude that the model after log transformation is fit the data better comparing with the others.

d) Assess the error assumptions in the regression model. Is this model appropriate in describing the relationship?

Answer:

Regarding the error assumptions:

The Residuals VS Fitted plot: Scatter is constant, The smoothing line is quite flat.

The Scale-location plot: Scatter is constant,The smoothing line is quite flat.

In the QQ Plot and Normality test:

The residuals looks like from a normal distribution.Considering of the sample size is 100, it is good.

So this model is appropriate in describing the relationship.

e) Interpret the p-values for the coefficients, and the p-value for the regression

Answer:

The p-values for the coefficients using T-test to test if each coefficient is significantly different from 0.In these case, all the p-values of coefficients<0.05. They are all significantly different from 0.

The p-value for the regression perform ANOVA on the whole model to determine if the overall model is statistically significant. In this case, p-value: < 2.2e-16, the overall model is statistically significant.

f) Use the predict() function with the argument interval ="confidence" to estimate the weight of a slug with length = 10, and interpret the 95% confidence interval. (Hint: predictions of weight will be given in log form, so use the exp() function to transform back into actual units).

Answer:

A slug with length = 10 is expected to (i.e. on an average) has the weight of 0.02767468.

We are 95% confident that this mean weight will be between 0.02404703 and 0.03184958.

```r
exp(predict(log_fit,newdata = data.frame(loglength=log(10)),interval = "confidence"))
```

```
##          fit        lwr        upr
## 1 0.02767468 0.02404703 0.03184958
```

g) Repeat (f) but use the argument interval = "predict" instead. Interpret the 95% prediction interval.

Answer:

A slug with length = 10 is expected to (i.e. on an average) have the weight of 0.02767468.

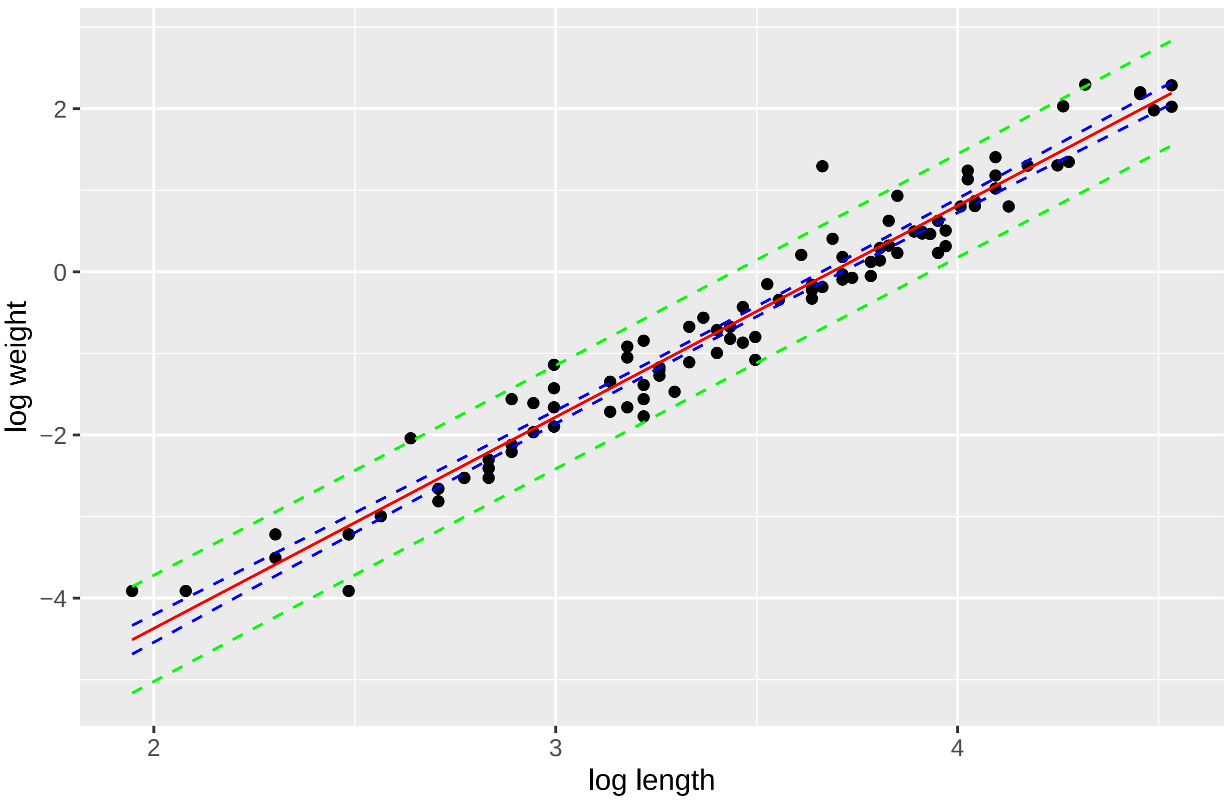
With probability 0.95, a slug with length = 10 will have the weight between 0.0145307 and 0.05270826 .

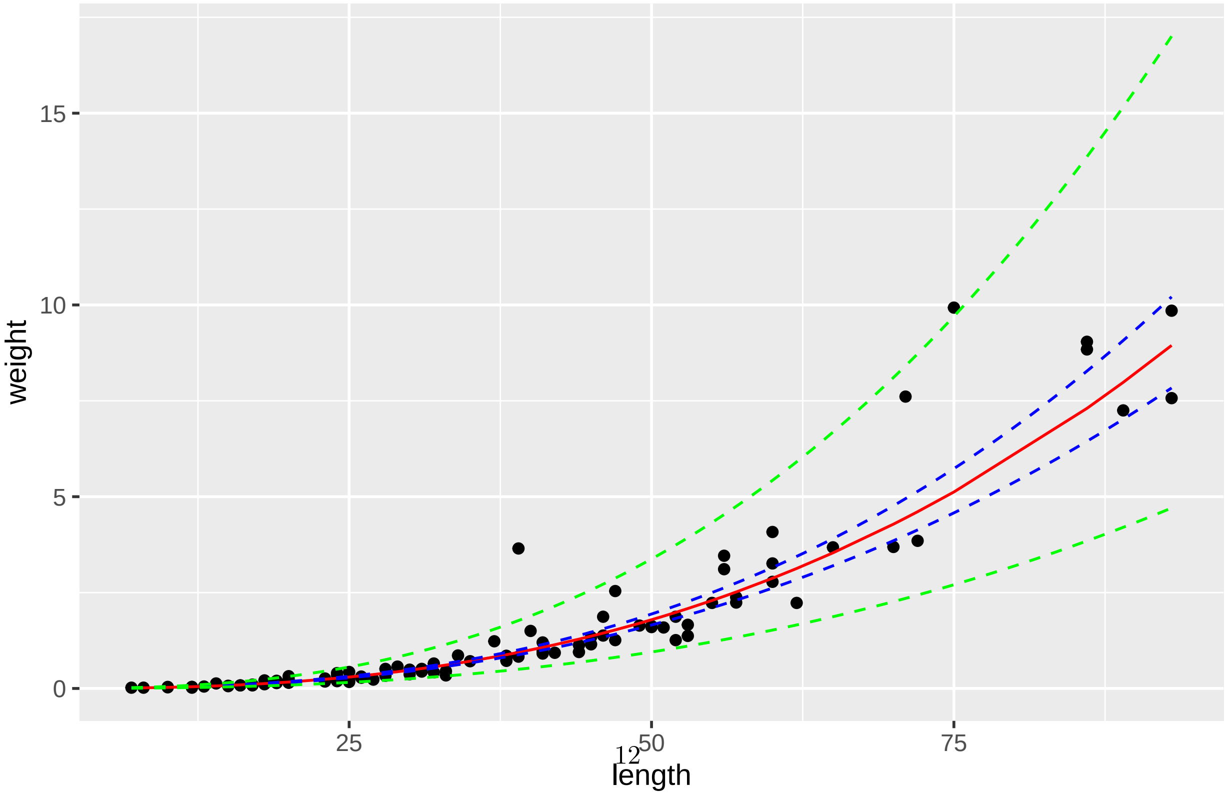
```r
exp(predict(log_fit,newdata = data.frame(loglength=log(10)),interval = "prediction"))
```

```
##          fit       lwr        upr
## 1 0.02767468 0.0145307 0.05270826
```

h) To the scatterplot of data overlayed with the regression line, plot both the confidence interval and the prediction interval in the log scale. Produce another plot in the original scale by back-transforming everything. So, produce two plots in total.

Answer:


Regression Model(log) with Confidence and Prediction Intervals


back−transformed Model with Confidence and Prediction Intervals