707ash945

2024-05-14

TASK 1: Plotting

Q1.a)

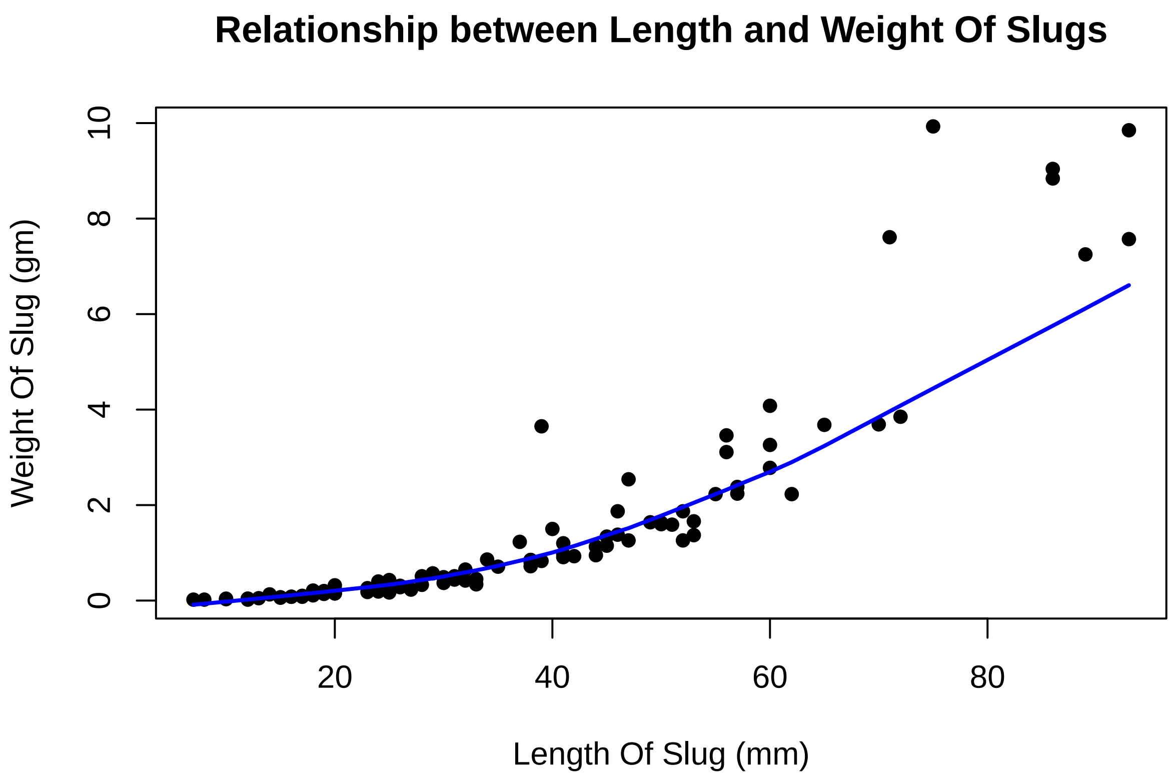
Solution:

```
x11()
slugs.df = read.csv("slugs.csv", header = TRUE)
with(slugs.df,  plot(length, weight, xlab="Length Of Slug (mm)",pch =16, ylab="Weight Of Slug (gm)", mai
```

Q1.b)

Solution:

```
with(slugs.df, {
  plot(length, weight, xlab = "Length Of Slug (mm)", pch = 16, ylab = "Weight Of Slug (gm)", main = "Rel
  lines(lowess(length, weight), col = "blue", lwd = 2)
})
```



The relationship with respect to direction is Positive The relationship with respect to shape can be given by using smoothers. Here I have used LOWESS smoother to identify the trend. It shows that the relationship is roughly linear. To quantify the relationship whether the relationship is linear or non linear we can also use Pearson's Correlation Coefficient

```
with(slugs.df, cor(length,weight))
```

```
## [1] 0.8747506
```

So Pearson correlation coefficient of 0.87 indicates a strong linear relationship.

TASK 2: Straight-Line Model

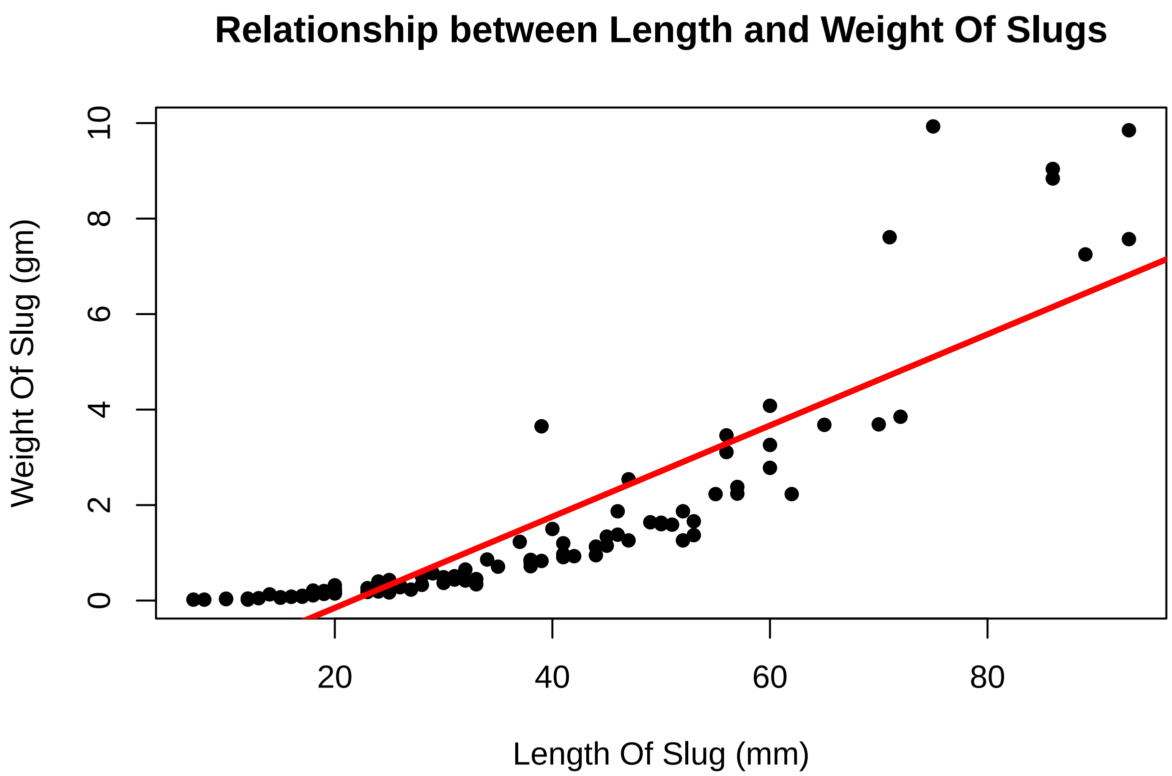
Q2.a)

Solution:

```
slugs.lm = lm(weight ~ length, data = slugs.df)


with(slugs.df,  plot(length, weight, xlab="Length Of Slug (mm)",pch =16, ylab="Weight Of Slug (gm)", ma

abline(slugs.lm, col = "red", lwd = 3)
```



# Q2.b)

Solution:

```
summary(slugs.lm)
```

```
##
## Call:
## lm(formula = weight ~ length, data = slugs.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
```

Straight-line regression equation = Estimated Mean Weight = -2.0616 + 0.0954 * length Here, intercept is

-2.0616 and slope is 0.095482. This is interpreted as follows.

Intercept (-2.0616):The intercept is the value of the dependent variable (estimated weight) when the independent variable (length) is zero. Here we get a negative value of weight which does not have any significant practical use.

Slope (0.095482): For every additional unit increase in length of slug , its estimated weight increases by 0.095482 gms.

Q2.c)

Solution:

Multiple R-squared: 0.7652 It means that 76.52% variation in weight of Slugs can be explained by variability in slug length. In the context of regression analysis, R-squared value represents the proportion of the variance in the dependent variable (in this case, slug weight) that is explained by the independent variable(s) (in this case, slug length) included in the regression model. The remaining 23.48% variability in slug weight could be due to other factors.

Q2.d)

Solution:

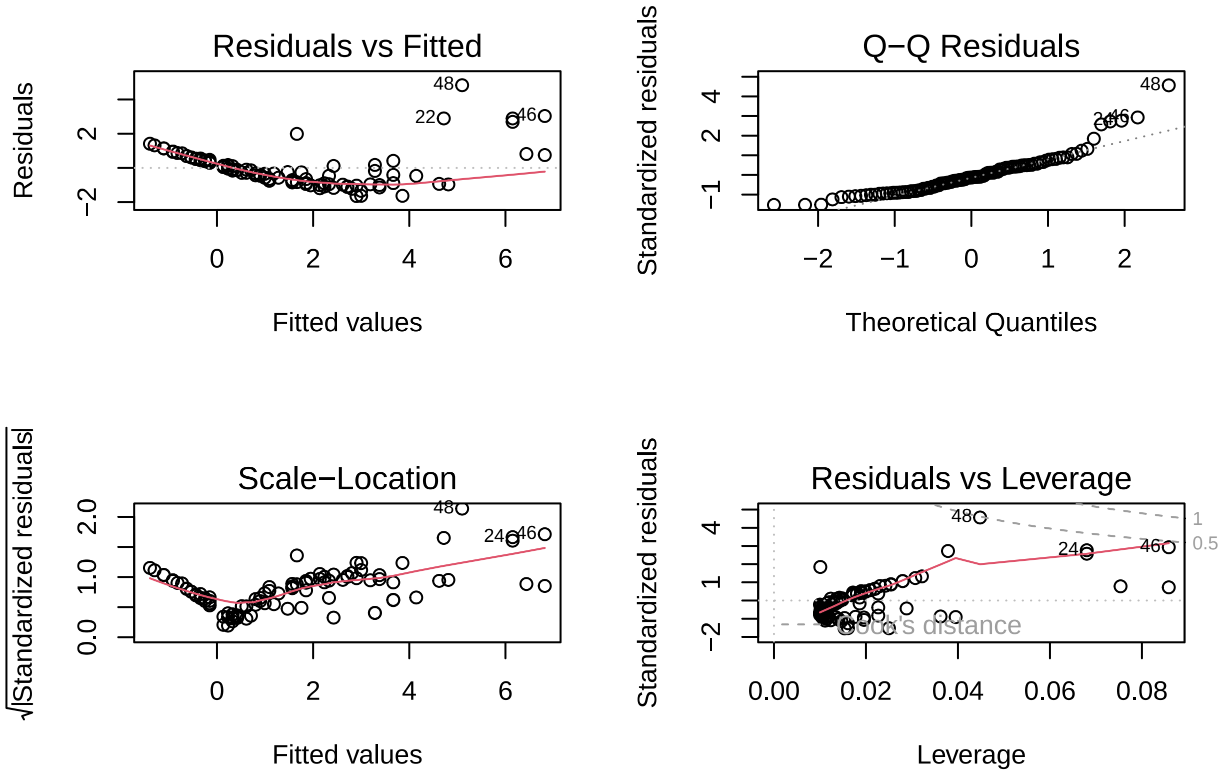
As we can visually see that there are a few outliers in our model which are at extreme ends. So the straight line model may not be the appropriate model for our data. The overall trend in the data does not seem to be captured by the straight line model. Outliers can distort the estimated slope and intercept, leading to misleading conclusions about the relationship between weight and length.

Q2.e)

Solution:

Error Assumptions can be made by looking at the residuals. For checking the assumptions we will create a 4 in 1 residual plot

```r
old.par = par(mfrow = c(2, 2))
plot(slugs.lm)
```



```r
par(old.par)
```

From the plots we can conclude following error assumptions:

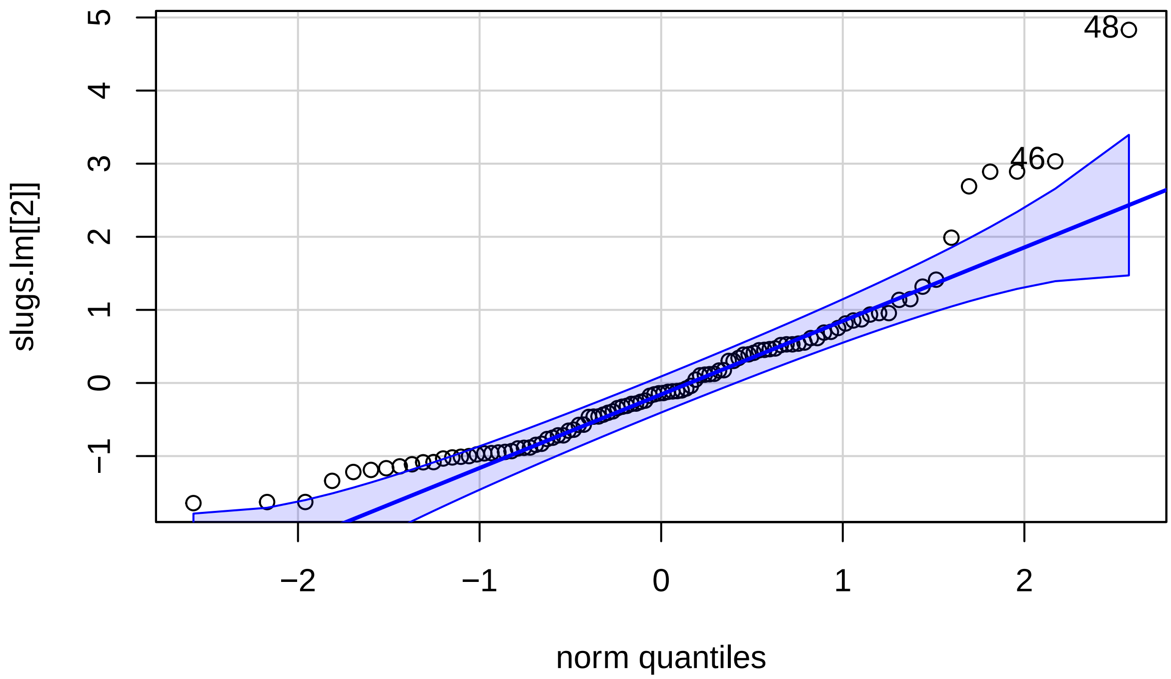
- All the errors are independent. This condition is not satisfied because there is a pattern left in the errors that we could model with the variables we have identified.In both Residual vs Fitted plot and Scale-Location plot we can see some pattern. This means that the errors are not independent.

- The errors come from a common Normal distribution with mean 0 and standard deviation s This condition can be further checked using QQ plot and then Shapiro Test. In Shapiro Test we will test following hypothesis: H0: data is consistent with a Normal distribution, HA: It is not.

```r
library(car)
```

```
## Loading required package: carData
```

```r
qqPlot(slugs.lm[[2]])
```



```
## [1] 48 46
```

```
shapiro.test(residuals(slugs.lm))
```

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

QQPlot of residuals shows that approximately 90% are inside confidence interval In Shapiro test the p-value is very less. Hence we have a strong evidence against the assumption the residuals were sampled from a Normal distribution. Hence we can say that the straight line model may not be appropriate for modelling the relationship between Slugs Weight and Height.

TASK 3:Quadratic Model

Q3.a)

Solution:

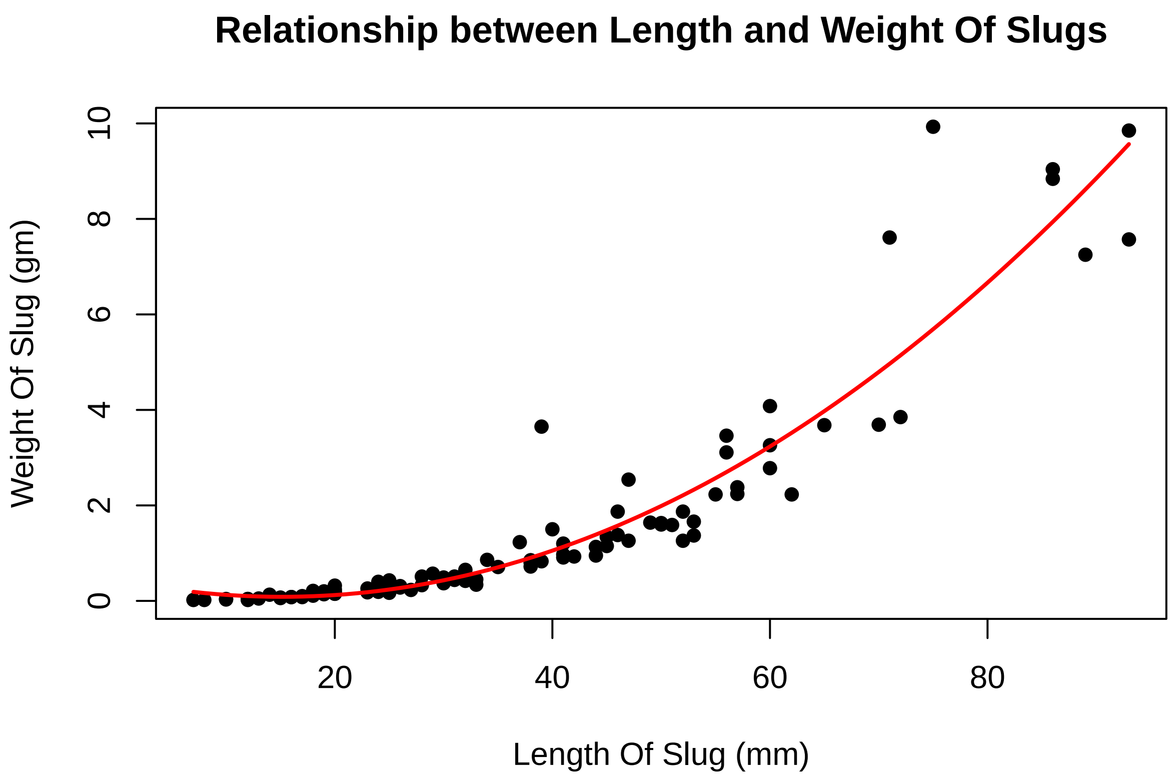
```
slugs.quad.lm = lm(weight ~ length + I(length^2), data = slugs.df)
summary(slugs.quad.lm)
```

```
##
## Call:
## lm(formula = weight ~ length + I(length^2), data = slugs.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
```

```r
#Plot data
with(slugs.df,  plot(length, weight, xlab="Length Of Slug (mm)",pch =16, ylab="Weight Of Slug (gm)", ma

#Plot a quadratic
preds.df=data.frame(length=seq(7,93,1))
quadfit = predict(slugs.quad.lm, preds.df)
lines(seq(7,93,1), quadfit,col= "red", lwd = 2)
```



# Q3.b)

Solution:

Quadratic regression equation:

Estimated mean Weight = 0.4433687 + -0.0472923 * length + 0.0015633 * length^2

Interpretation:

Intercept(0.4433687): Estimated weight of slug when length is zero.

Coefficient of Linear Term (-0.0472923): The linear effect of length on weight is indicated by this. As the value is negative it indicates that as length increases by one unit the mean weight decreases by 0.0472923, all else being equal.

Coefficient of Quadratic Term (0.0015633): This coefficient represents the change in the rate of change of estimated weight with respect to length. In other words, it captures the curvature in the relationship between length and weight. A positive coefficient indicates that the rate of increase in estimated weight slows down as length increases, and may eventually reach a maximum or minimum before changing direction.

Q3.c)

Solution:

The fitted curve closely follows the pattern of the data points. It captures the curvature/trend in the relationship between weight and length variables.So we can say that the quadratic model may be appropriate.

Q3.d)

Solution:

For Quadratic Model: Residual standard error: 0.7361 on 97 degrees of freedom Multiple R-squared: 0.8926, Adjusted R-squared: 0.8904

For Linear Model: Residual standard error: 1.083 on 98 degrees of freedom Multiple R-squared: 0.7652, Adjusted R-squared: 0.7628

Adjusted R-squared: The quadratic model has a higher adjusted R-squared (0.8904) compared to the linear model (0.7628), indicating that the quadratic model explains a larger proportion of the variance in weight.

Residual Standard Error: The quadratic model has a lower residual standard error (0.7361) compared to the linear model (1.083), so we can conclude that the quadratic model provides a better fit to the data, with smaller residuals on average.

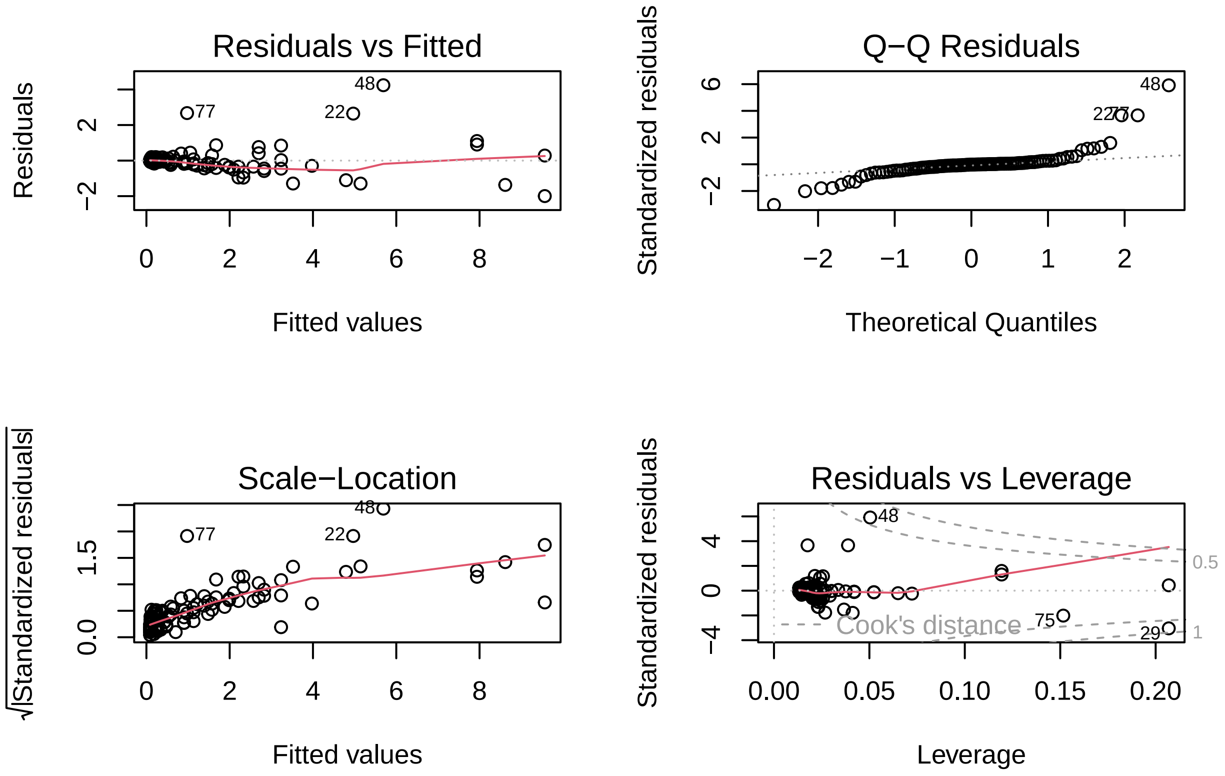
By looking at these data we can conclude that the Quadratic model provides better fit to the data compared to the linear model. The quadratic model captures the relationship between length and weight more accurately.

Q3.e)

Solution:

Error Assumptions can be made by looking at the residuals. For checking the assumptions we will create a 4 in 1 residual plot

```r
old.par = par(mfrow = c(2, 2))
plot(slugs.quad.lm)
```

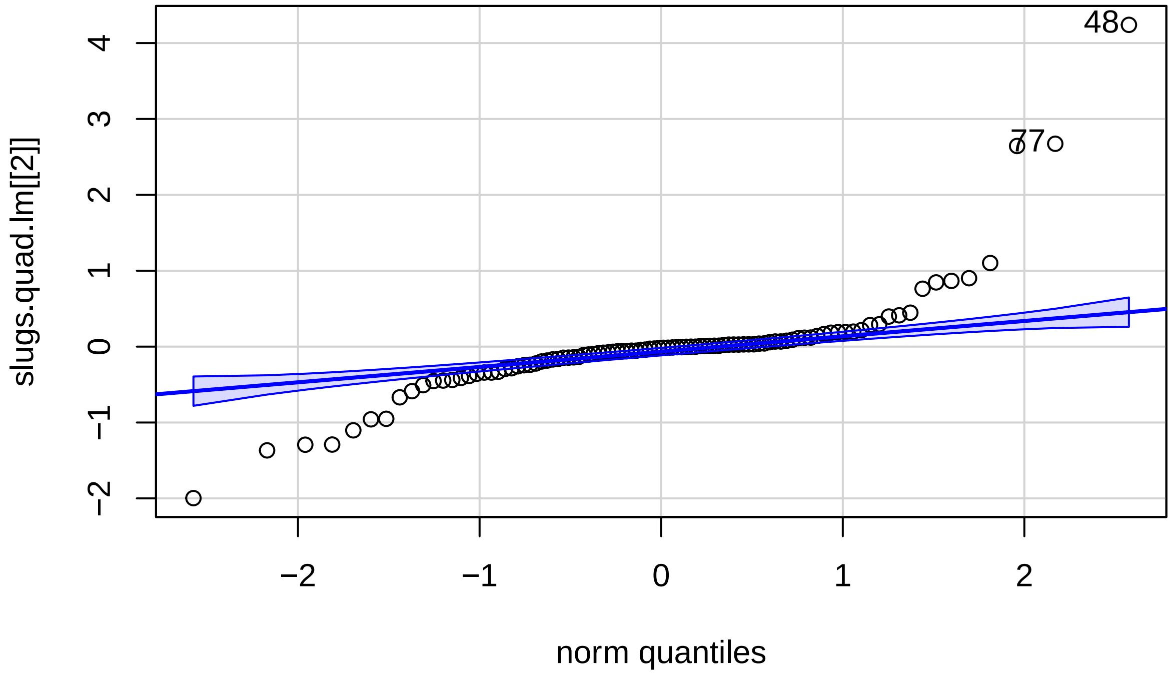


```
par(old.par)
```

From the plots we can conclude following error assumptions:

- All the errors are independent. The condition of independence of errors is not satisfied as there are patterns remaining in the residuals after accounting for the variables included in the model.The Scale-Location plot shows a trend, indicating non independence of errors
- The errors come from a common Normal distribution with mean 0 and standard deviation s This condition can be further checked using QQ plot and then Shapiro Test. In Shapiro Test we will test following hypothesis: H0: data is consistent with a Normal distribution, HA: It is not.

```
library(car)
qqPlot(slugs.quad.lm[[2]])
```



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

In QQPlot of residuals we can see that there are approximately 80% residuals that lie within the confidence interval.It shows clear deviation from normality.

```
shapiro.test(residuals(slugs.quad.lm))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(slugs.quad.lm)
## W = 0.69737, p-value = 4.686e-13
```

In Shapiro test the p-value is very low. Hence we have a strong evidence against the assumption that the residuals were sampled from a Normal distribution.

We can say that the quadratic model may not be appropriate in describing the relationship as the error assumptions are not met.

TASK 4: Variable Transformation and Inference

Q4.a)

Solution:

```
# Create new variables log_weight and log_length
slugs.df$log_weight <- log(slugs.df$weight)
```

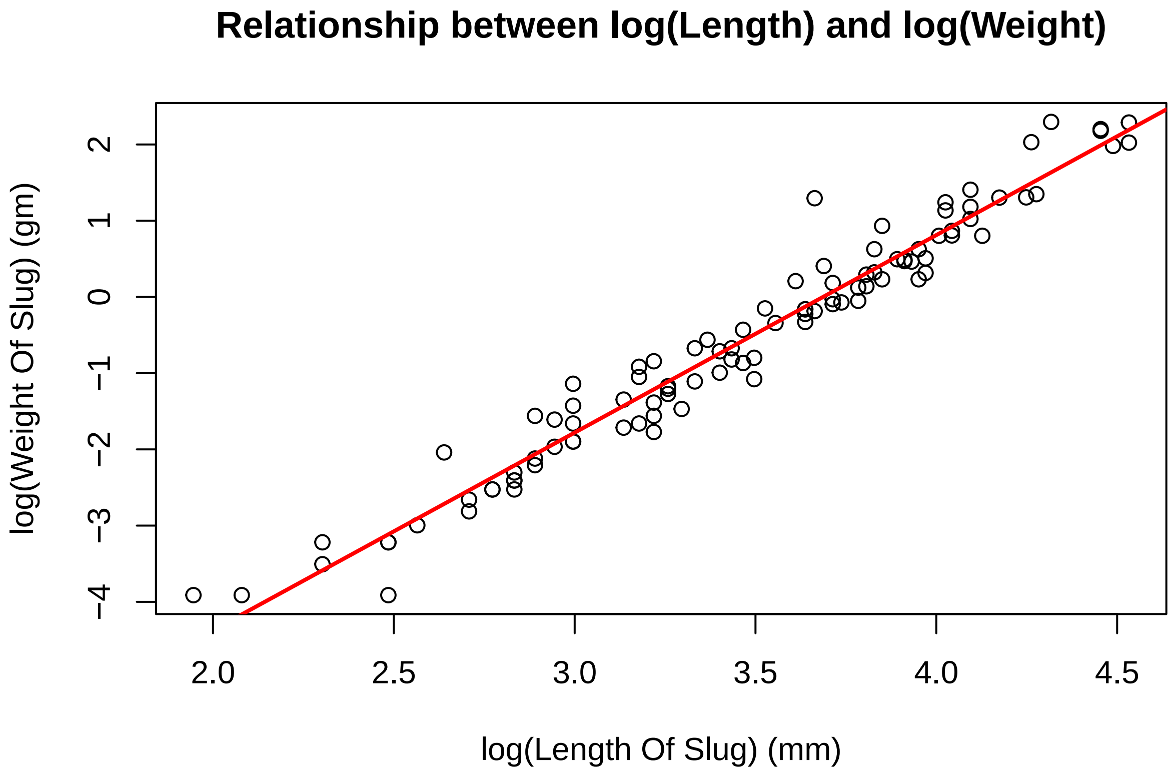
```r
slugs.df$log_length <- log(slugs.df$length)

# Fit a straight-line model
slugs.log.lm <- lm(log_weight ~ log_length, data = slugs.df)

# Summary of the model
summary(slugs.log.lm)
```

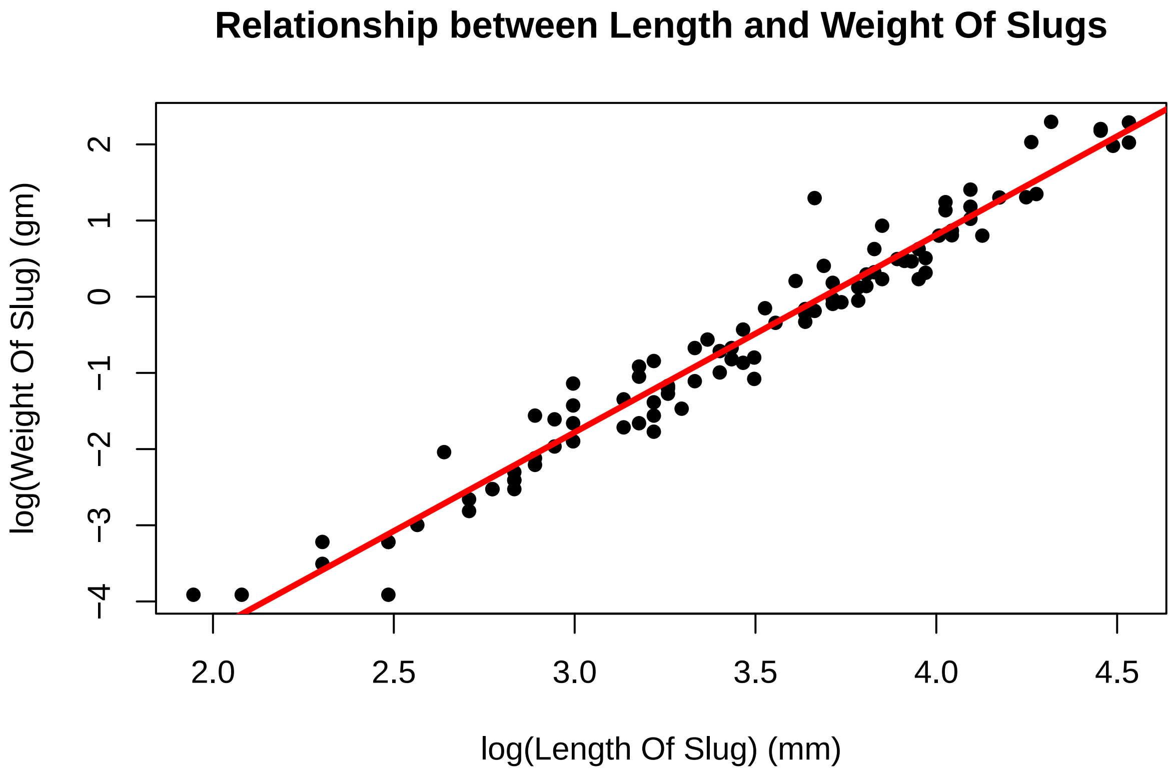
```
##
## Call:
## lm(formula = log_weight ~ log_length, data = slugs.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7972 -0.1713 -0.0893  0.1966  1.3555
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.55359    0.19196  -49.77   <2e-16 ***
## log_length   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
```

```r
# Scatterplot with straight-line relationship
plot(slugs.df$log_length, slugs.df$log_weight, xlab = "log(Length Of Slug) (mm)", ylab = "log(Weight Of
abline(slugs.log.lm, col = "red", lwd = 2)
```



```
slugs.df$log_weight <- log(slugs.df$weight)
slugs.df$log_length <- log(slugs.df$length)
slugs.log.lm = lm(log(weight) ~ log(length), data = slugs.df)
with(slugs.df,  plot(log(length), log(weight), xlab="log(Length Of Slug) (mm)",pch =16, ylab="log(Weigh

abline(slugs.log.lm, col = "red", lwd = 3)
```



```r
summary(slugs.log.lm)
```

```
##
## Call:
## lm(formula = log(weight) ~ log(length), data = slugs.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 ***
## log(length)  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
```

Q4.b)

Solution:

mean log(weight) = -9.55359 + 2.59115 * length

median weight = exp(-9.55359 + 2.59115 * length) = exp(-9.55359) * exp(2.59115)^length = 0.000070946 * 13.3451^length

That is, we predict the median weight of slugs to be about 0.000070946 when height is zero, but the median weight increases by a factor of approximately 13.3451 for each unit increase in length.

Q4.c)

Solution:

For log model Residual standard error: 0.3168 on 98 degrees of freedom Multiple R-squared: 0.9581, Adjusted R-squared: 0.9577

For Quadratic Model: Residual standard error: 0.7361 on 97 degrees of freedom Multiple R-squared: 0.8926, Adjusted R-squared: 0.8904

For Linear Model: Residual standard error: 1.083 on 98 degrees of freedom Multiple R-squared: 0.7652, Adjusted R-squared: 0.7628

Adjusted R-squared: The log model has the highest adjusted R-squared (0.9577), followed by the Quadratic model (0.8904), and then the linear model (0.7628). Therefore, the log model explains the largest proportion of the variance in the dependent variable, followed by the quadratic model and then the linear model.

Residual Standard Error: The log model has the lowest residual standard error (0.3168), followed by the quadratic model (0.7361), and then the linear model (1.083). Therefore, the log model has the smallest spread of residuals around the fitted values, indicating a better fit compared to the other models.

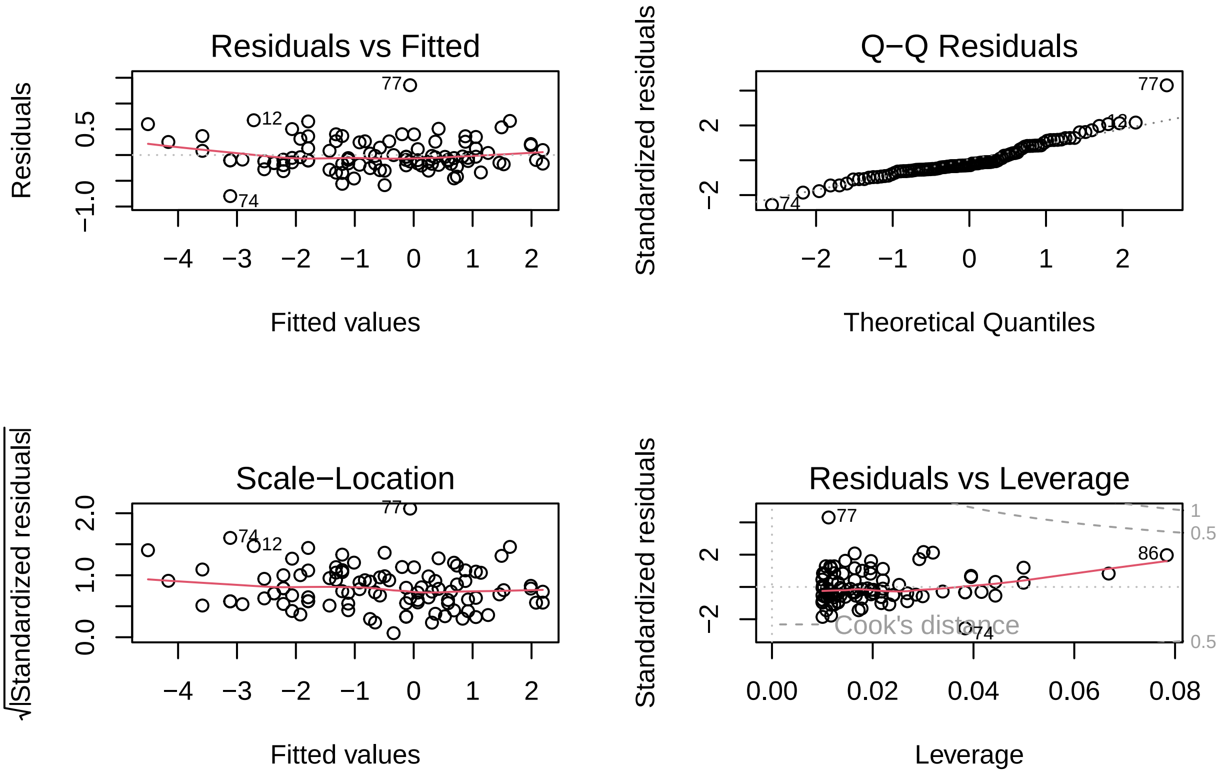
Conclusion: We can conclude that the log model provides the best overall fit among the three models, as it has both a higher adjusted R-squared and a lower residual standard error compared to the quadratic and linear models.

Q4.d)

Solution:

Error Assumptions can be made by looking at the residuals. For checking the assumptions we will create a 4 in 1 residual plot

```r
old.par = par(mfrow = c(2, 2))
plot(slugs.log.lm)
```



```
old.par
```

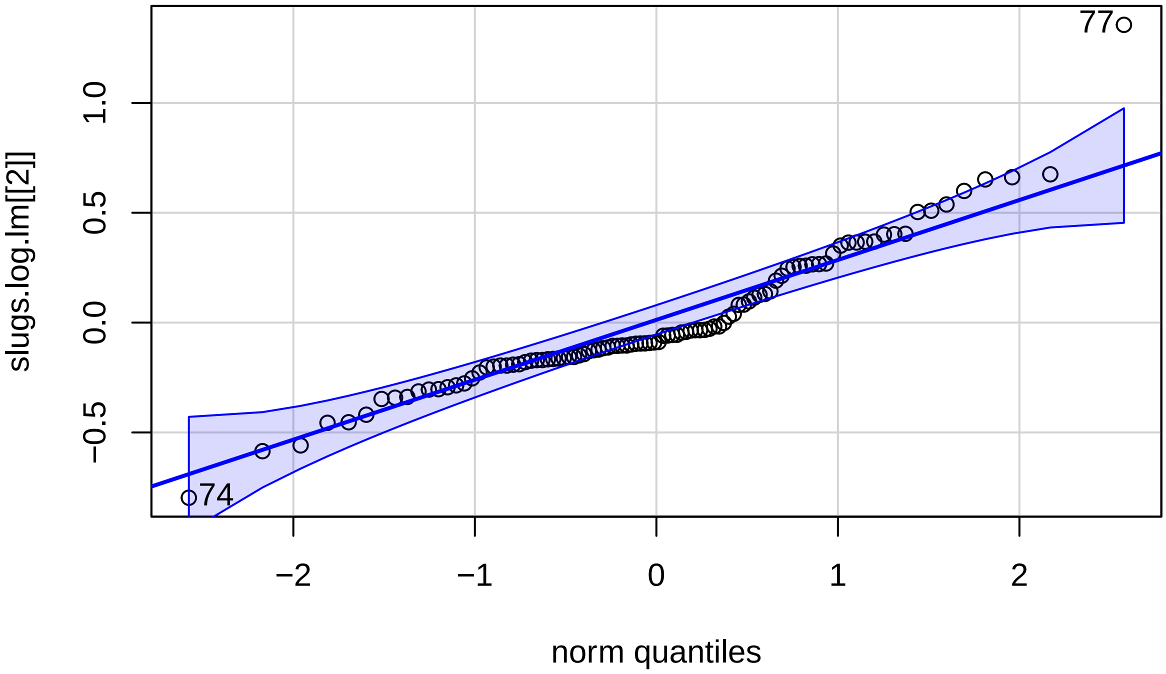
```
## $mfrow
## [1] 1 1
```

From the plots we can conclude following error assumptions:

- All the errors are independent. The condition of independence of errors is satisfied as the points seem to be randomly scattered.The Scale-Location plot also show no precise trends and it shows somewhat constant variance of the points,indicating independence of errors.

- The errors come from a common Normal distribution with mean 0 and standard deviation s This condition can be further checked using QQ plot and then Shapiro Test. In Shapiro Test we will test following hypothesis: H0: data is consistent with a Normal distribution, HA: It is not.

```
qqPlot(slugs.log.lm[[2]])
```



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

```
shapiro.test(residuals(slugs.log.lm))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(slugs.log.lm)
## W = 0.9331, p-value = 7.494e-05
```

In QQPlot of residuals we can see that there are approximately 95% residuals that lie within the confidence interval.It shows normality. In Shapiro Test we can see that the p-value is below 0.05 hence we will reject the null hypothesis that the data is consistent with normal distribution.

Hence this model can be considered better than the linear and quadratic model as the error assumptions are somewhat met by this model.

Q4.e)

solution:

P-values of coefficient:

Coefficients: Pr(>|t|)
(Intercept) <2e-16 ***log(length) <2e-16***

The p-value measures the significance of each coefficient in the model. Here the null hypothesis is that the intercept coefficient as well as the length coefficient are zero. and alternate hypothesis that the coefficients are different than zero Here as we can see the p-value is significantly small hence indicating that both coefficients are highly significant predictors of log(weight)

P-value for regression model:

F-statistic: 2242 on 1 and 98 DF, p-value: < 2.2e-16 The p-value associated with the F-statistic tests the overall significance of the regression model.The null hypothesis for the F-test is that all coefficients in the model are equal to zero (i.e., none of the predictor variables have a significant effect on the response variable).In this output, the p-value for the F-statistic is extremely small (< 2.2e-16), indicating that the regression model as a whole is highly significant.

Q4.f)

Solution:

```
exp(predict(slugs.log.lm, newdata = data.frame(length = 10), interval = "confidence"))
```

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

Here we are 95% confident that median weight of a slug with length 10mm would be somewhere between 0.0.02404703gms and 0.03184958gms

Q4.g)

Solution:

```
exp(predict(slugs.log.lm, newdata = data.frame(length = 10), interval = "prediction"))
```

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

Here with 0.95 probability , the weight of slug with length 10mm would be somewhere between 0.0145307gms and 0.05270826gms

Q4.h)

Solution:

“‘