The Galapagos Islands:
Total Species Number Prediction Model

ST 412 Project
Khalid Alshammari, Gavin Kim, Natasha Nemyre

Introduction

In island ecology it is generally accepted that an island's species richness is a function of the area of the island. The larger the island, the greater the number of species present. In an effort to establish whether islands in the Galapagos Archipelago subscribe to this general ecological assumption, we conducted a data analysis on a publicly available dataset in the Sleuth library (ex1220). The dataset measured the total number of species along with several other geographic variables, including area, on 30 islands in the Galapagos Archipelago.

Methods & Results

The response variable for our dataset was *Total* number of species. The explanatory variables included the *Area* of the island, the island's *Elevation*, the distance to the *Nearest* island, the distance to the island of *Santa Cruz* (which is the central island in the Galapagos and is also the central hub of human activity), and the *Area of the Nearest* island. To begin, we randomly removing 5 observations in order to reintroduce for predictive purposes after establishing a model. We ran a multiple regression analysis on the remaining 25 observations in order to ascertain which of the variables were statistically significant to the determination of total species number on islands in the Galapagos Archipelago.

To establish the best fit model for our data, we conducted both a forward and backward model selection (Figure 1). Both processes yielded the same model. Our best fit model for the response variable of an island's *Total* species included only the *Elevation* and *Area of the Nearest* island parameters.

We analyzed the dataset for adherence to the assumptions of multiple linear regression. We assumed that the data was randomly collected to satisfy the independence assumption. Our residual

plot initially revealed non-constant variance within the data (Figure 4). After looking at the Box-Cox transformation plot (Figure 5), we concluded that, since the 95% confidence interval did not include λ=0, no log transformation was necessary. Instead, we opted to transform the response variable using a Box-Cox transformation with a λ value of 0.5 (equation: $T(y) = 2 \times [\sqrt{y} - 1]$). The post-transformation residual plot revealed much more constant variance and exhibited no pattern in the data thereby satisfying the linearity assumption (Figure 8). Next, we looked at the Q-Q plot to assess normality and the data appeared to adequately meet the normality assumption (Figure 9). Since we initially ordered the data before deleting the five observations, we faced difficulty in deleting influential points with high leverage. To fix this, we reordered the data after the removal of the five observations. To check for odd influential points, we ran the case statistics (Figure 10). The results showed no high residuals, however, observation 15 appeared influential with high leverage. We removed observation 15 which did affect the parameter estimates (Figure 11). Upon running the assumptions again (Figures 12, 13, 14), we found observation 12 to be influential with high leverage. We deleted observation 12 and ran through analyzed the assumptions again. After transformation and the deletion of observations 12 and 15, all assumptions are satisfied with no odd influential observations (Figures 16, 17, 18). Finally, we used the five initially removed data points for prediction against the actual data. All values were captured successfully within the 95% prediction intervals (Figure 19).

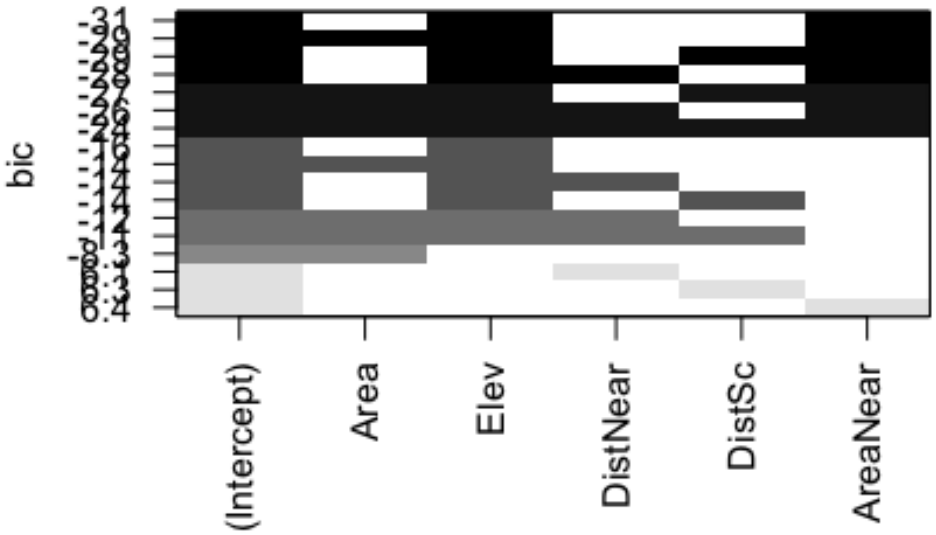
Summary

We created a predictive model for the median total number of species for islands in the Galapagos Archipelago. Our findings show strong evidence that in the Galapagos, the elevation of an island has the most significant impact on the median total number of species found on that island. To achieve smaller prediction intervals with this model, it would be beneficial to use a larger data set for our analysis.

Galapagos Islands: Total Species Prediction Model Appendix

Figure 1. (a) Forward selection plot (b) Backward selection plot

(a)



(b)

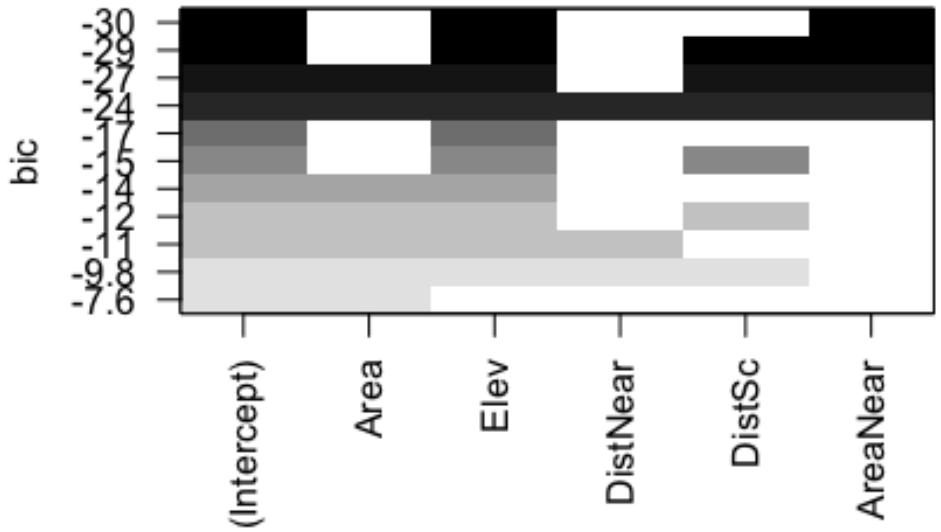


Figure 2. Data matrix

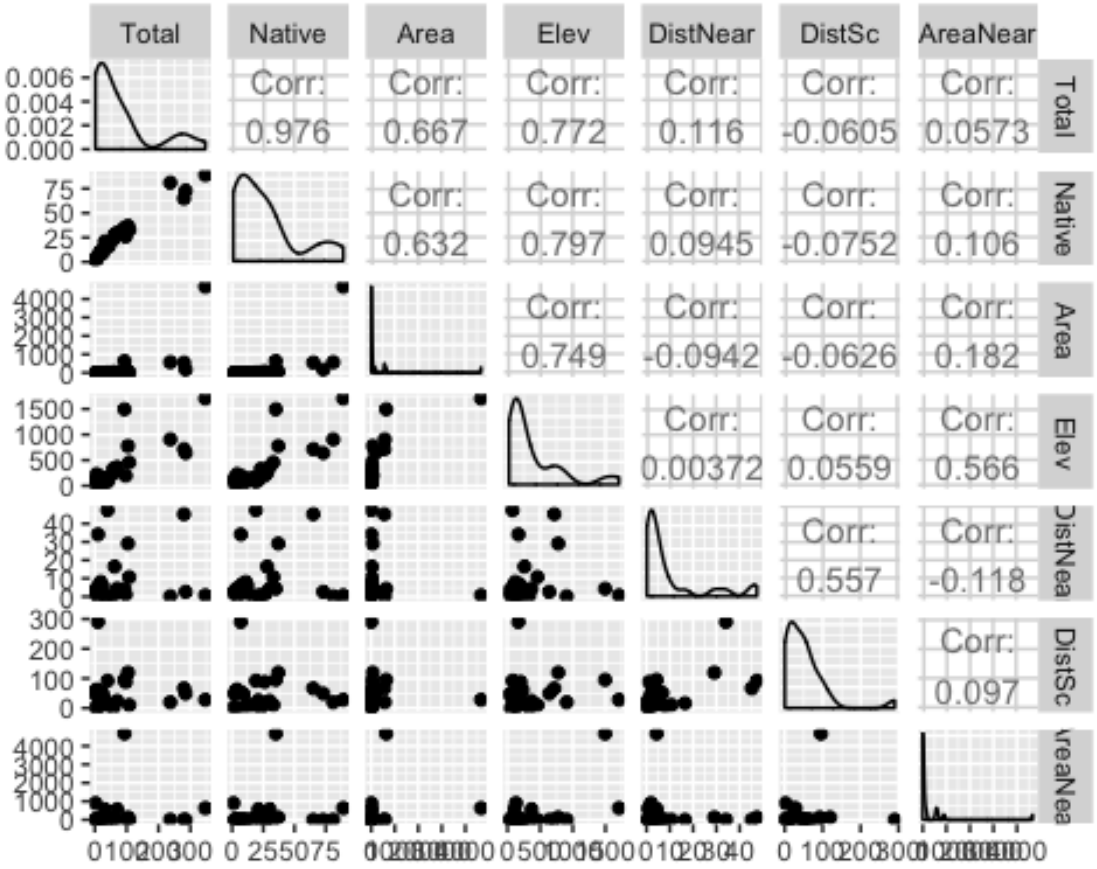


Figure 3. Using the best fit model

```
fit1 = lm(Total ~ Elev+ AreaNear, data = DataTotal)
summary(fit1)

##
## Call:
## lm(formula = Total ~ Elev + AreaNear, data = DataTotal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -83.23 -29.05  -5.67  10.74 123.76
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.11196   12.32645   0.334    0.742
## Elev         0.24552    0.02570   9.554 2.75e-09 ***
## AreaNear    -0.05908    0.01206  -4.900 6.72e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45.95 on 22 degrees of freedom
## Multiple R-squared:  0.8064, Adjusted R-squared:  0.7888
## F-statistic: 45.83 on 2 and 22 DF,  p-value: 1.429e-08
```

```
DataTotal <- fortify(fit1, DataTotal)
```

Figure 4. Checking the best fit graphs: Residual plot showing non-constant variance

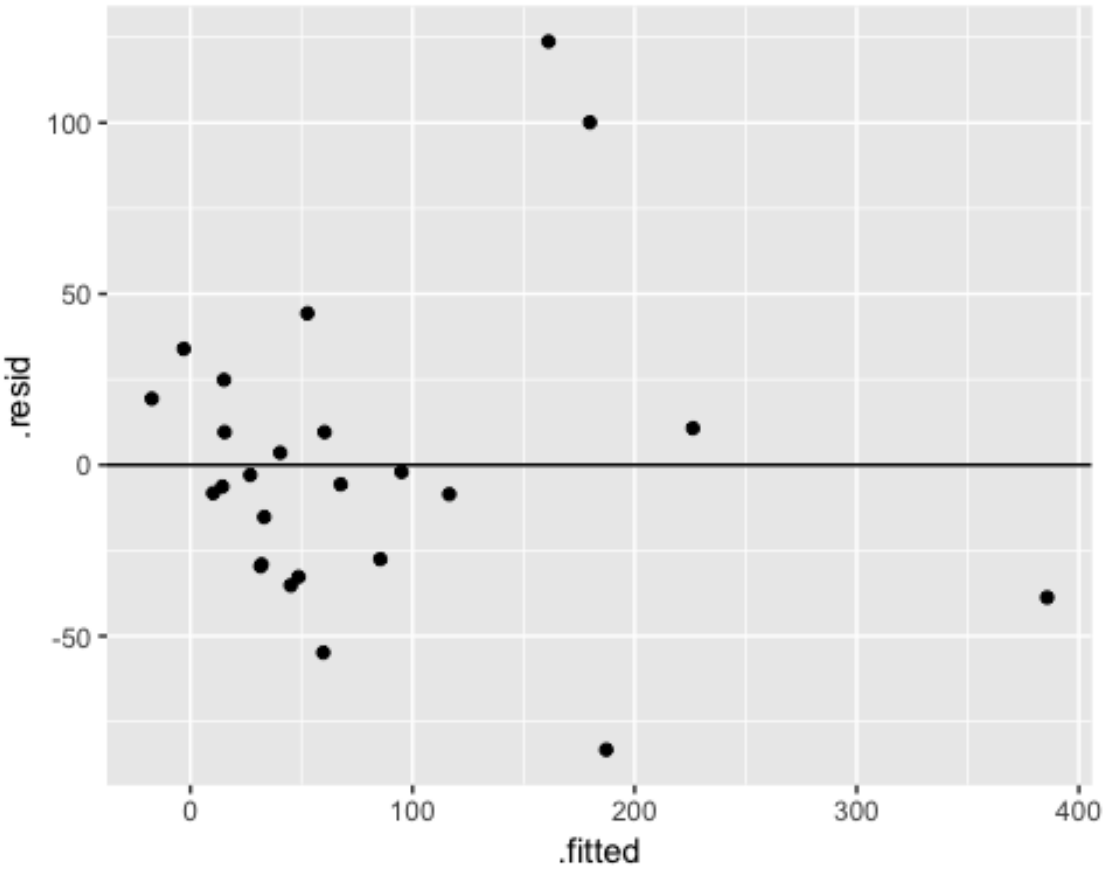


Figure 5. Box-Cox transformation plot: No log transformation necessary since 95% confidence interval does not include a lambda value of zero.

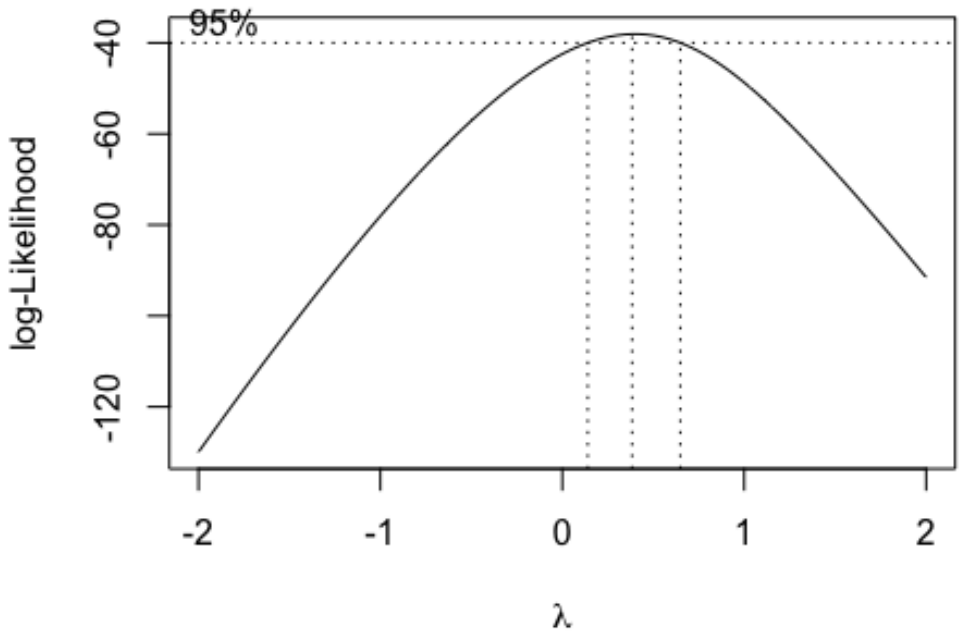


Figure 6. Transformation of the response variable

$$T(y) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

Figure 7. Transformation of response variable using lambda = 0.5

```
DataTotal$Total_Transformed = 2*(sqrt(DataTotal$Total)-1)
fit2 = lm(Total_Transformed ~ Elev+ AreaNear, data = DataTotal)
summary(fit2)

##
## Call:
## lm(formula = Total_Transformed ~ Elev + AreaNear, data = DataTotal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0746 -4.7665  0.4905  2.6539 11.2348
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  5.061514   1.470680   3.442  0.00233 **
## Elev         0.024169   0.003066   7.883 7.53e-08 ***
## AreaNear    -0.005237   0.001439  -3.640  0.00144 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.482 on 22 degrees of freedom
## Multiple R-squared:  0.7416, Adjusted R-squared:  0.7181
## F-statistic: 31.56 on 2 and 22 DF,  p-value: 3.434e-07

DataTotal_Transformed <- fortify(fit2, DataTotal)
```

Figure 8. Checking the graphs after transformation: Residual plot showing a more constant variance and no pattern thereby satisfying the linearity assumption

```
qplot(.fitted, .resid, data = DataTotal_Transformed) + geom_hline(yintercept =
0)
```

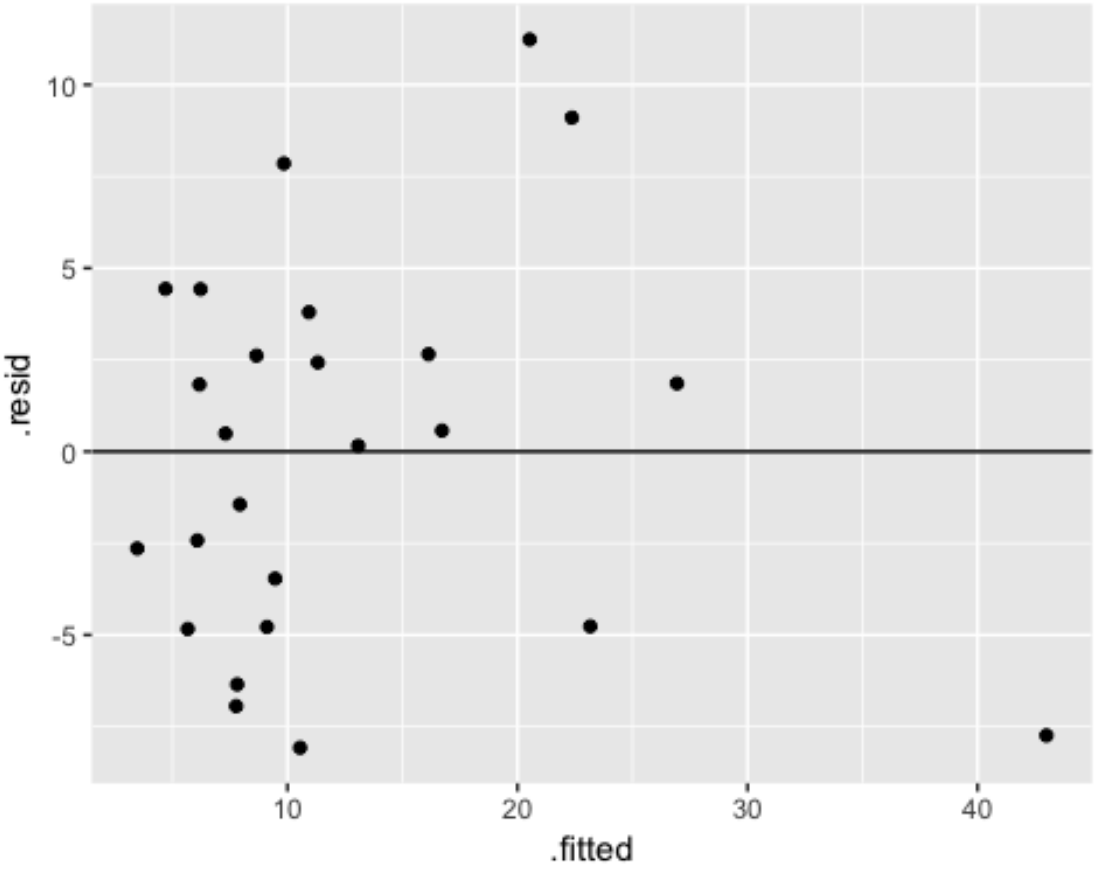
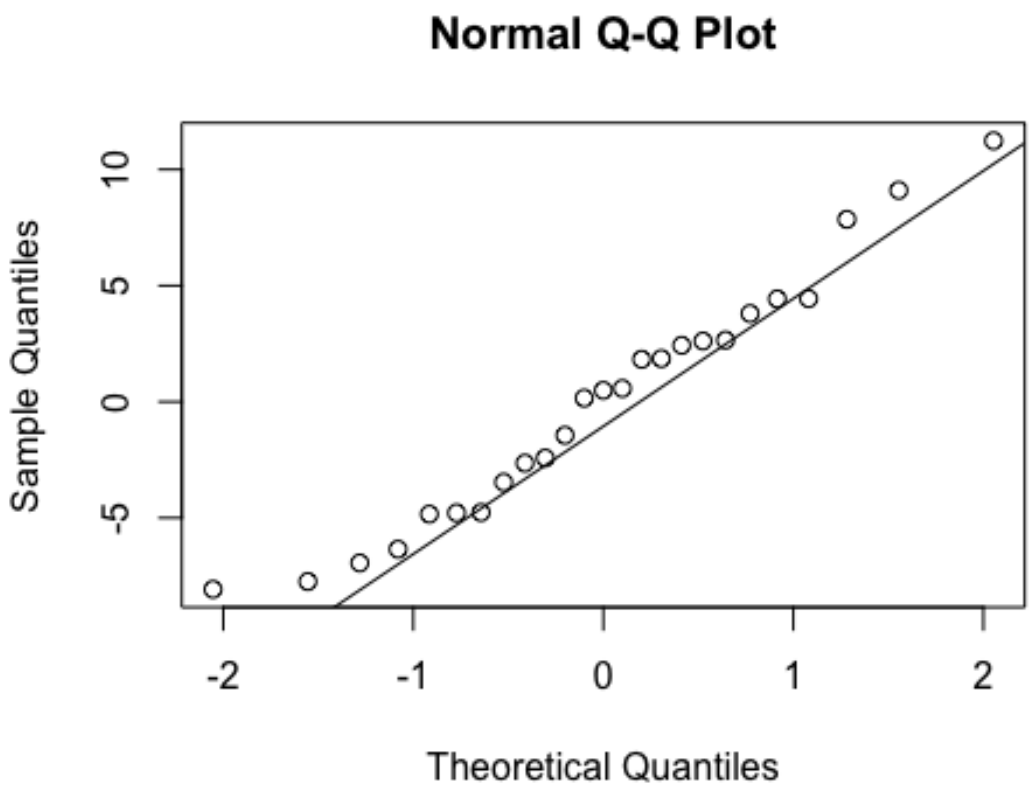


Figure 9. Checking the graphs after transformation: Q-Q plot showing normality in the data

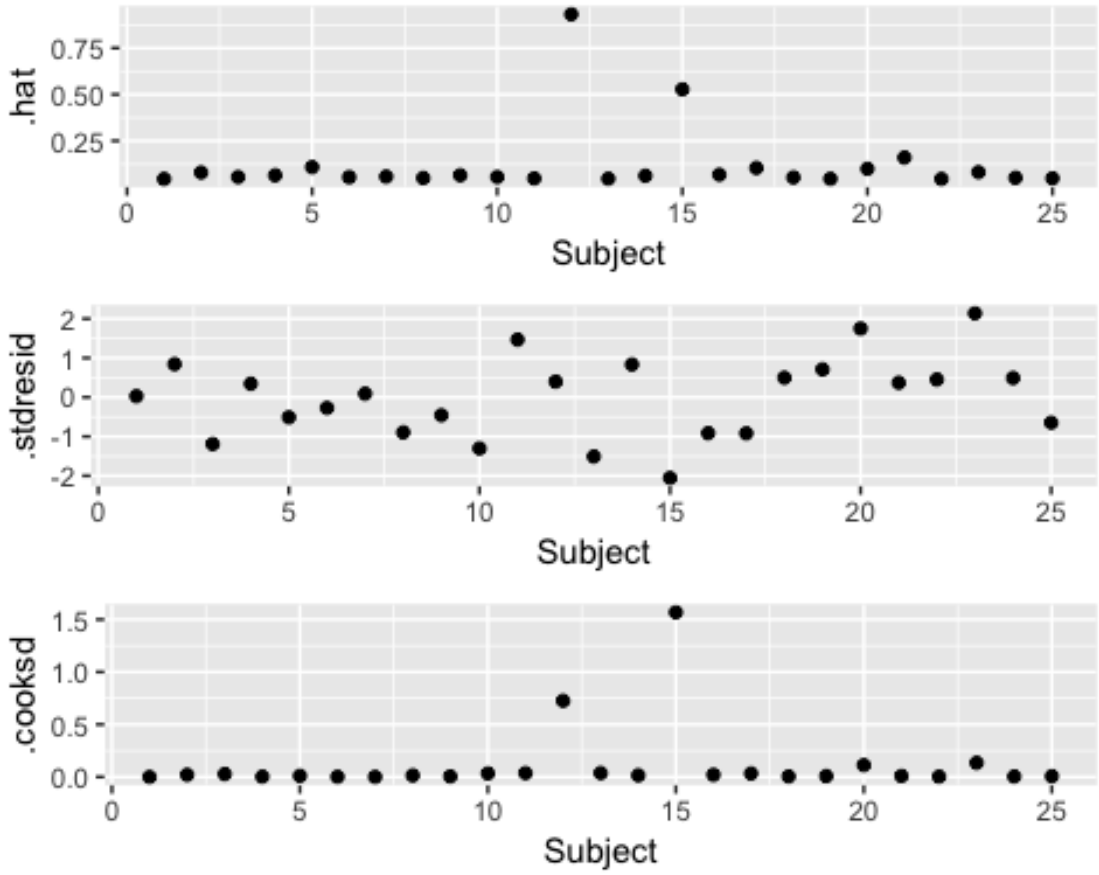
```
qqnorm(DataTotal_Transformed$.resid); qqline(DataTotal_Transformed$.resid)
```




```
Reordering the data
DataTotal_Transformed$Subject = c(1:25)
```

Figure 10. Case statistics on data with transformed response. Observation 15 is influential with high leverage. Observation 12 has high leverage but is not influential. Most data points are between -2 and 2 on the second graph indicating that there are not high residuals.

```
p1 <- qplot(Subject,.hat, data = DataTotal_Transformed)
# Above (2*2)/25 =  0.16 is high leverage
p2 <- qplot(Subject,.stdresid, data = DataTotal_Transformed)
# between [-2,2] is good
p3 <- qplot(Subject,.cooksd, data = DataTotal_Transformed)
# above 1 is high
multiplot(p1,p2,p3,cols=1)
```




```
Deleting observation 15:

DataTotal_Transformed_no_15 <- DataTotal_Transformed[-15, ]
```

Figure 11. Using the transformed data without observation 15

```
fit3 = lm(Total_Transformed ~ Elev+ AreaNear, data = DataTotal_Transformed_no_1
5)
summary(fit3)

##
## Call:
## lm(formula = Total_Transformed ~ Elev + AreaNear, data = DataTotal_Transform
ed_no_15)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2130 -3.4043 -0.7517  3.1347  8.4738
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.751777   1.475135   2.543 0.018915 *
## Elev         0.030530   0.004011   7.612 1.81e-07 ***
## AreaNear    -0.006674   0.001472  -4.534 0.000181 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 5.045 on 21 degrees of freedom
## Multiple R-squared:  0.7359, Adjusted R-squared:  0.7107
## F-statistic: 29.25 on 2 and 21 DF,  p-value: 8.49e-07

DataTotal_Transformed_no_15 <- fortify(fit3, DataTotal_Transformed_no_15)
```

Figure 12. Checking the graphs after transformation and deletion of observation 15: Residual plot

```
qplot(.fitted, .resid, data = DataTotal_Transformed_no_15) + geom_hline(yinterc
ept = 0)
```

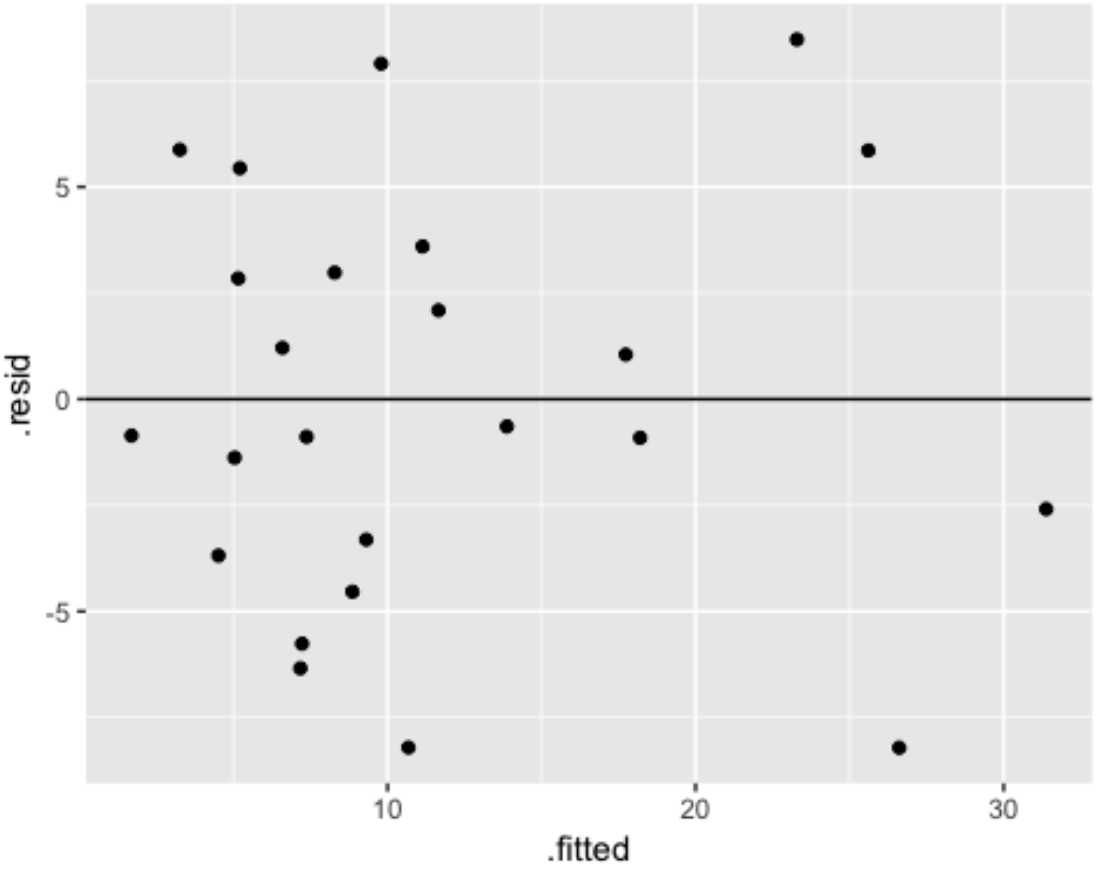
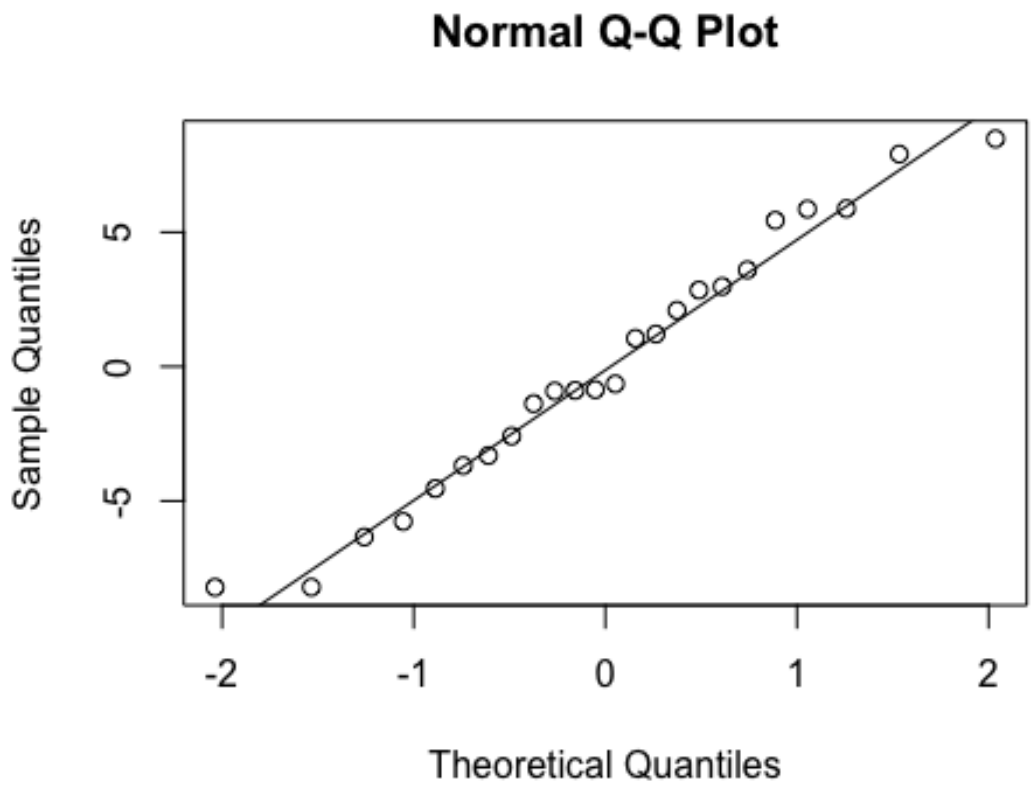


Figure 13. Checking the graphs after transformation and deletion of observation 15: Q-Q plot showing normality

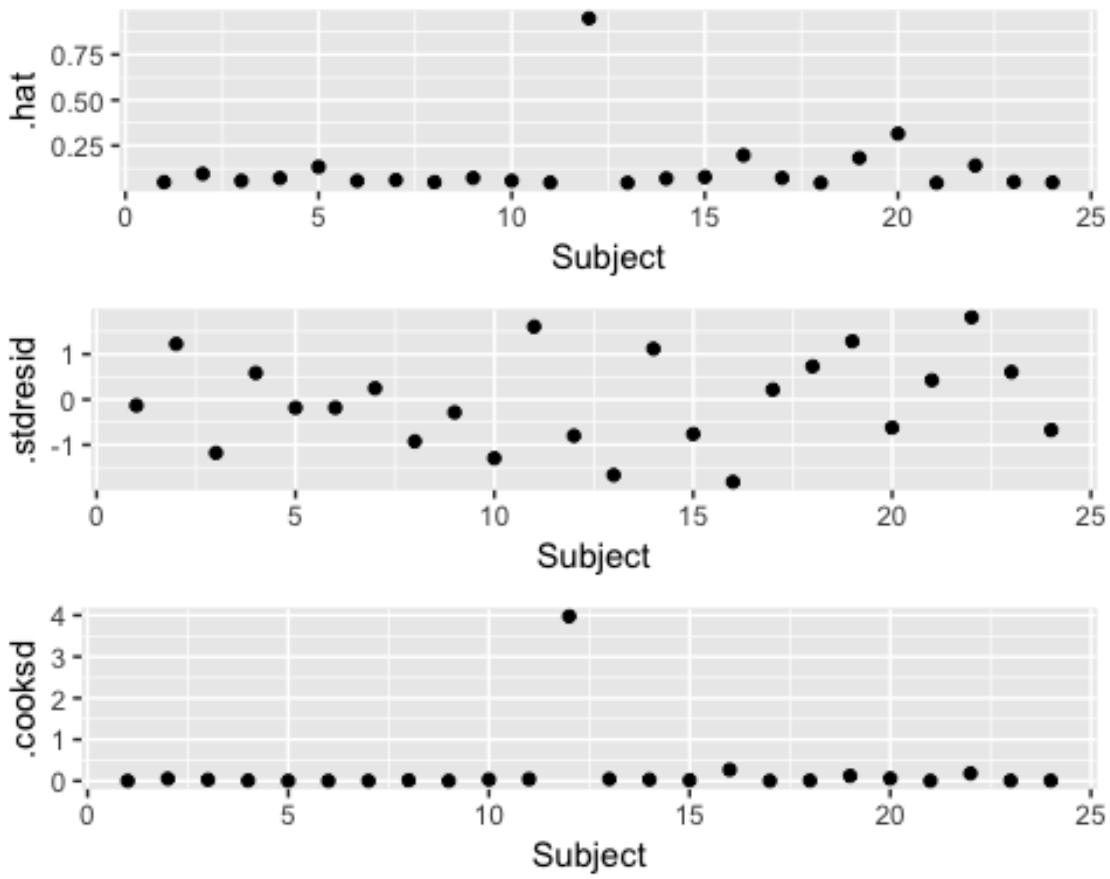
```
qqnorm(DataTotal_Transformed_no_15$.resid);

qqline(DataTotal_Transformed_no_15$.resid)
```




```
Reordering the data
DataTotal_Transformed_no_15$Subject = c(1:24)
```

Figure 14. Case statistics with transformed response and removal of observation 15. Observation 12 is influential so we remove it since it also has high leverage.

```
p1 <- qplot(Subject,.hat, data = DataTotal_Transformed_no_15)
# Above (2*2)/24 =   0.1666667 is high leverage
p2 <- qplot(Subject,.stdresid, data = DataTotal_Transformed_no_15)
# between [-2,2] is good
p3 <- qplot(Subject,.cooksd, data = DataTotal_Transformed_no_15)
# above 1 is high
multiplot(p1,p2,p3,cols=1)
```




```
Deleting observation 12

DataTotal_Transformed_no_12_15 <- DataTotal_Transformed_no_15[-12, ]
```

Figure 15. Using the model without observations 12 and 15

```
fit4 = lm(Total_Transformed ~ Elev+ AreaNear, data = DataTotal_Transformed_no_1
2_15)
summary(fit4)

##
## Call:
## lm(formula = Total_Transformed ~ Elev + AreaNear, data = DataTotal_Transform
ed_no_12_15)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8943 -3.2889 -0.4014  3.3727  8.3846
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.139203   1.676948   1.872   0.0759 .
```

```
## Elev         0.031626   0.004276   7.396 3.84e-07 ***
## AreaNear    -0.003259   0.004557  -0.715   0.4828
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.09 on 20 degrees of freedom
## Multiple R-squared:  0.7397, Adjusted R-squared:  0.7136
## F-statistic: 28.41 on 2 and 20 DF,  p-value: 1.43e-06

DataTotal_Transformed_no_12_15 <- fortify(fit4, DataTotal_Transformed_no_12_15)
```

Figure 16. Checking the graphs after transformation and removing observations 12 and 15: Residual plot

```
qplot(.fitted, .resid, data = DataTotal_Transformed_no_12_15) + geom_hline(yintercept = 0)
```

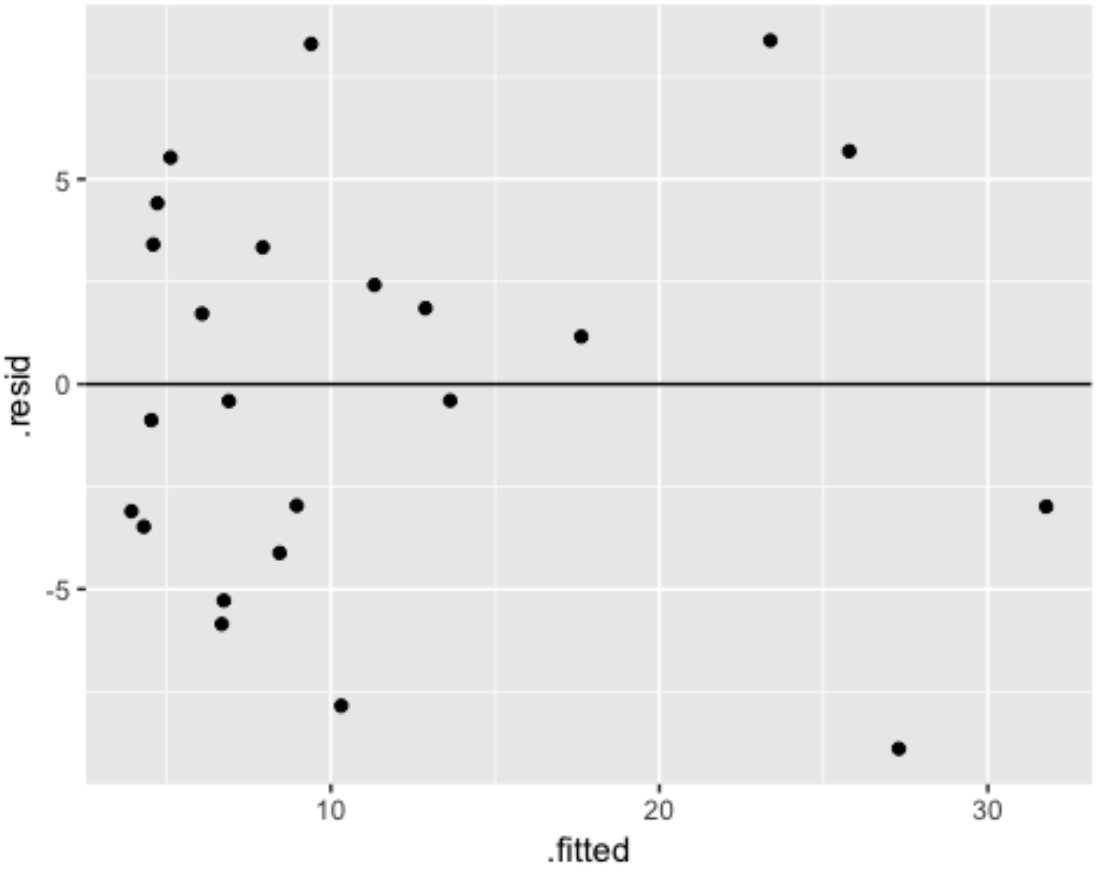
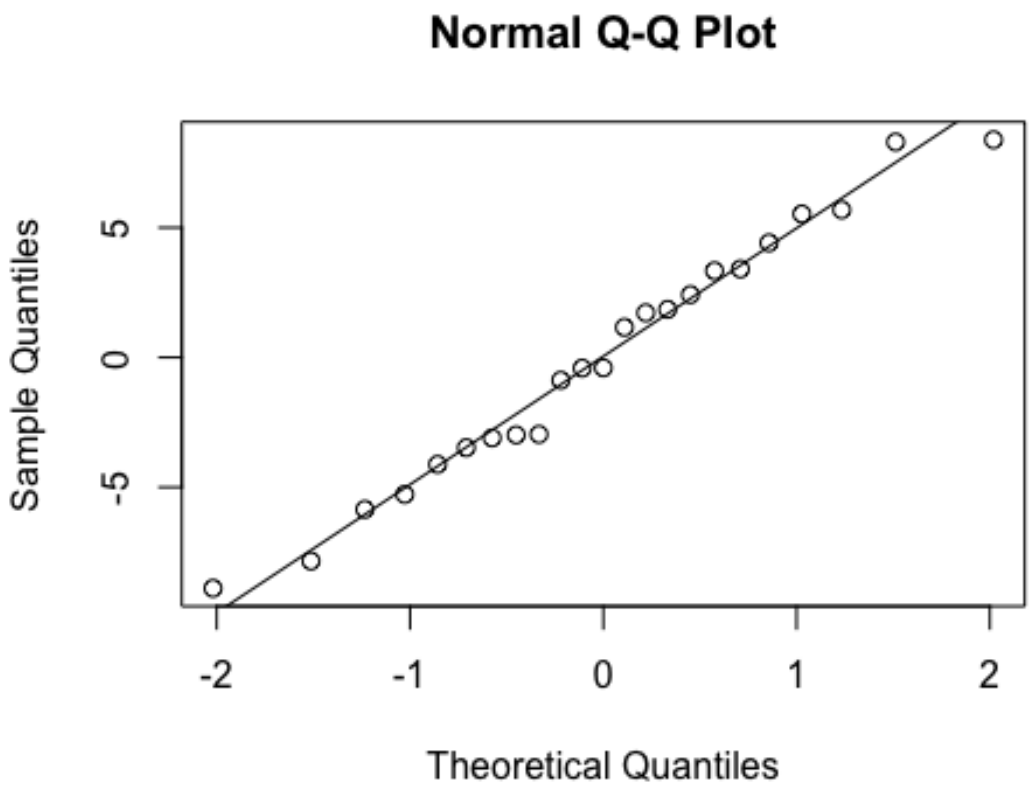


Figure 17. Checking the graphs after transformation and removing observations 12 and 15: Q-Q plot

```
qqnorm(DataTotal_Transformed_no_12_15$.resid);

qqline(DataTotal_Transformed_no_12_15$.resid)
```




```
Reordering the data
DataTotal_Transformed_no_12_15$Subject = c(1:23)
```

Figure 18. Case statistics with transformed response after removing observation 12 and 15. There are some data points with high leverage, however, none of them are influential

```
p1 <- qplot(Subject,.hat, data = DataTotal_Transformed_no_12_15)
# Above (2*2)/23 =  0.173913 is high leverage
p2 <- qplot(Subject,.stdresid, data = DataTotal_Transformed_no_12_15)
```

```
# between [-2,2] is good
p3 <- qplot(Subject,.cooksd, data = DataTotal_Transformed_no_12_15)
# above 1 is high
multiplot(p1,p2,p3,cols=1)
```

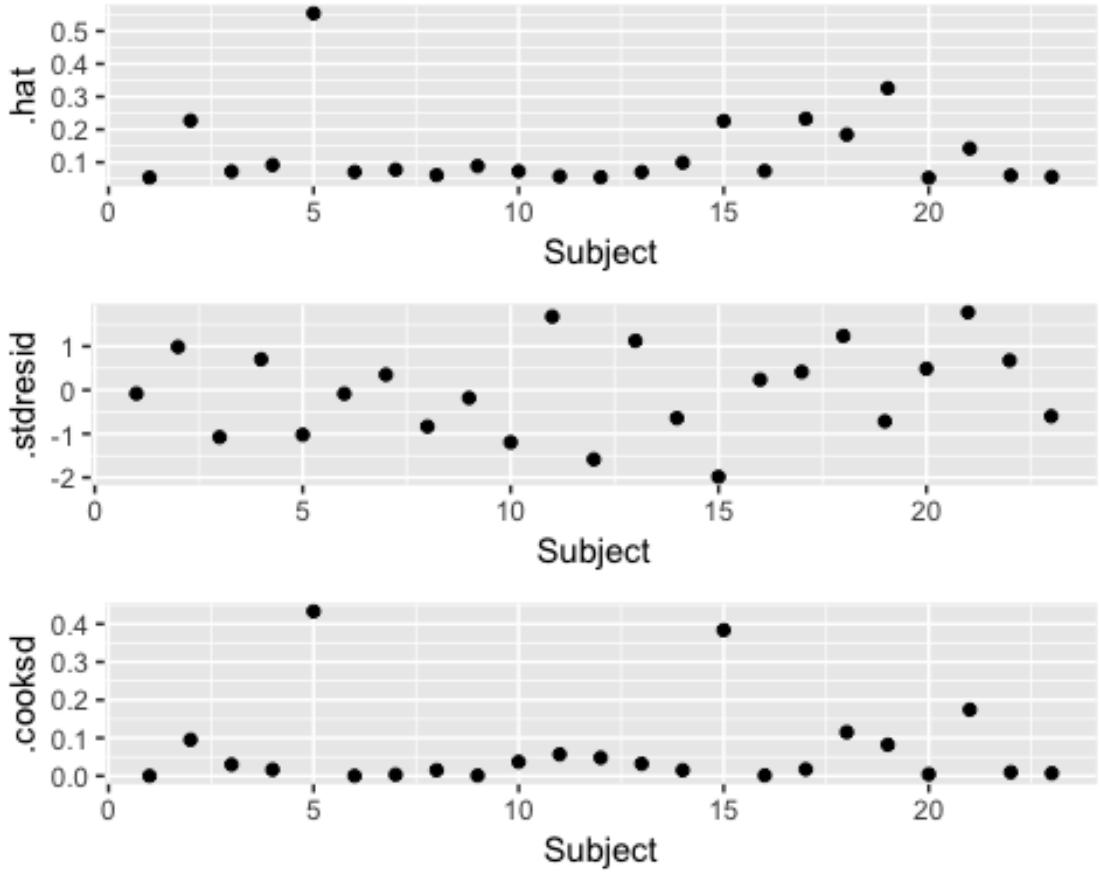


Figure 19. Predicting the five observations using the transformed best fit without 12 and 15. All values are captured within the 95% prediction intervals

```
# Using the deleted data for predection
newdata = ex1220[c(17,13,21,25,30),c(5,8)]
# 5 and 8 are the Elev and AreaNear respectively

# Note1: This is the data before transforming it back
# Note2: Here we are talking about the median not the mean, since we used squar
e root transformation
predict(fit4, newdata, interval = "prediction")
```

```
##          fit        lwr      upr
## 17 13.792723  2.9232934 24.66215
## 13  4.498972 -6.5435633 15.54151
## 21  5.713998 -5.2970174 16.72501
## 25 30.462108 18.4027010 42.52151
## 30 11.132910  0.2420736 22.02375
```

```r
# Transforming the data back and comparing the medians with their intervals wit
h the actual data
((predict(fit4, newdata, interval = "prediction")[,]/2)+1)^2
```

```
##          fit        lwr       upr
## 17  62.35252   6.059705 177.71758
## 13  10.55916   5.160992  76.92613
## 21  14.87644   2.717581  87.65653
## 25 263.44711 104.067552 495.54132
## 30  43.11833   1.256724 144.28510
```

```r
# to compare with original Totals
ex1220[c(17,13,21,25,30),2]
```

```
## [1]  51  58  12 444  21
```