MA 575: HW3

Benjamin Draves

September 26, 2017

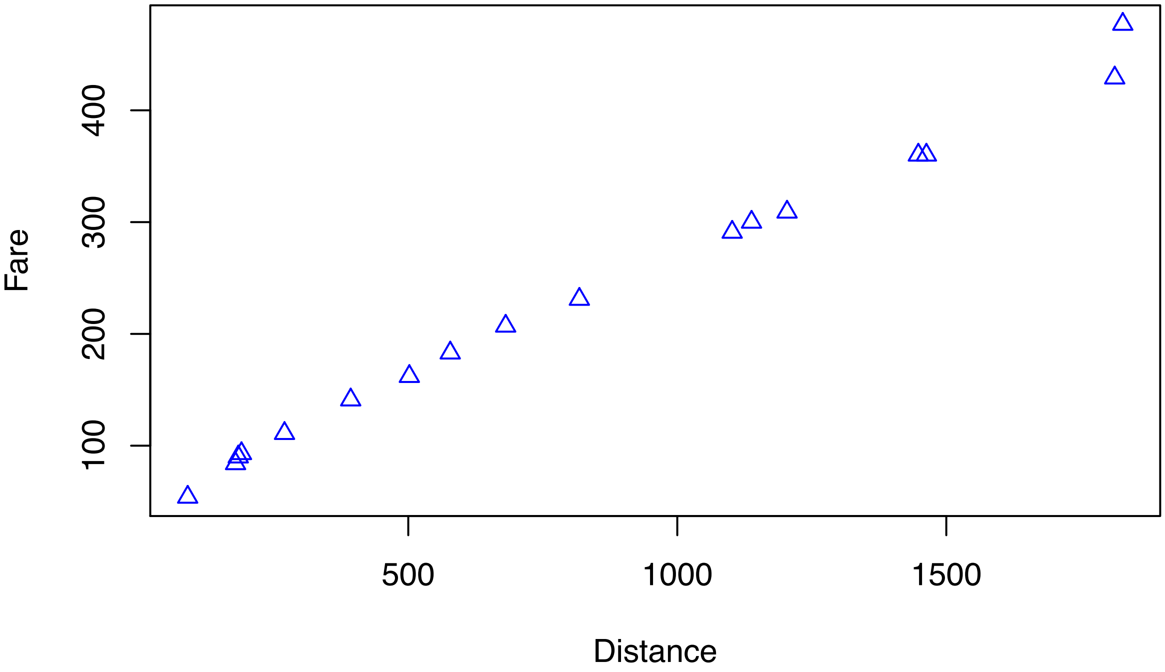
Exercise 3.1

We first reconstruct the model proposed in the problem.

```
#read in data
dat = read.table("~/Desktop/Fall 2017/MA 575/book_data/airfares.txt", header = TRUE)

#take a peak
head(dat)
```

```
##   City Fare Distance
## 1    1  360     1463
## 2    2  360     1448
## 3    3  207      681
## 4    4  111      270
## 5    5   93      190
## 6    6  141      393
```

```
#plot bivariate relationship
plot(dat$Distance, dat$Fare, ylab = "Fare", xlab = "Distance", pch = 2, col = "blue")
```

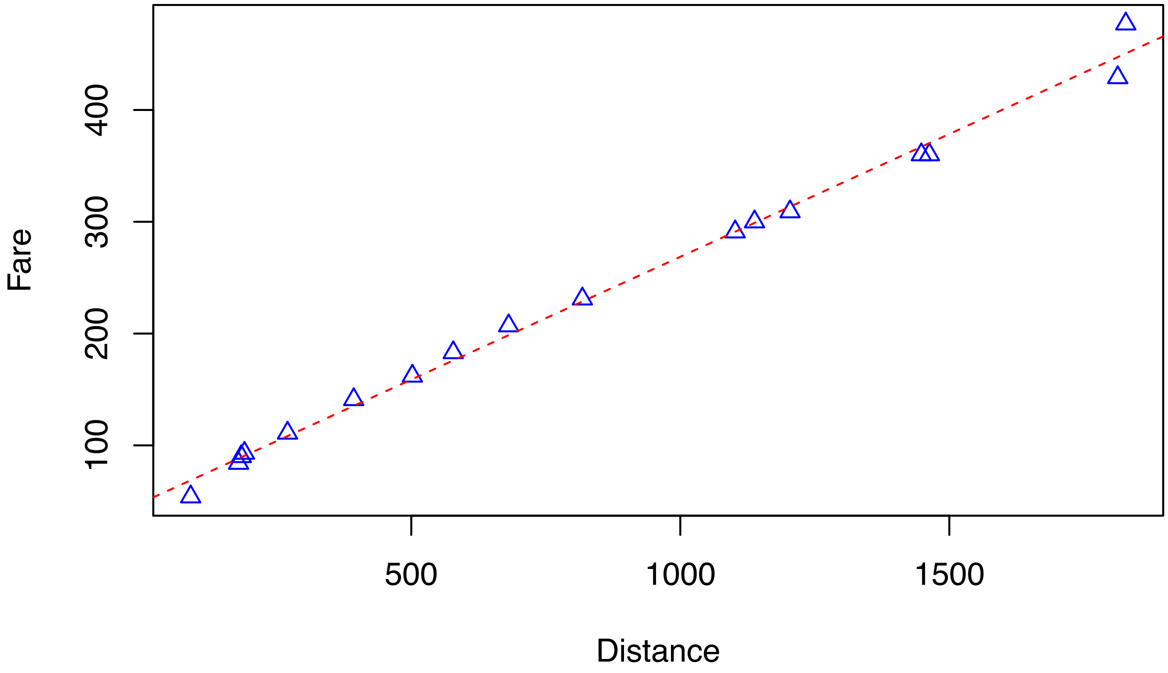


```
#build regression model
model = lm(Fare~Distance, data = dat)

#check out summary statistics
summary(model)
```

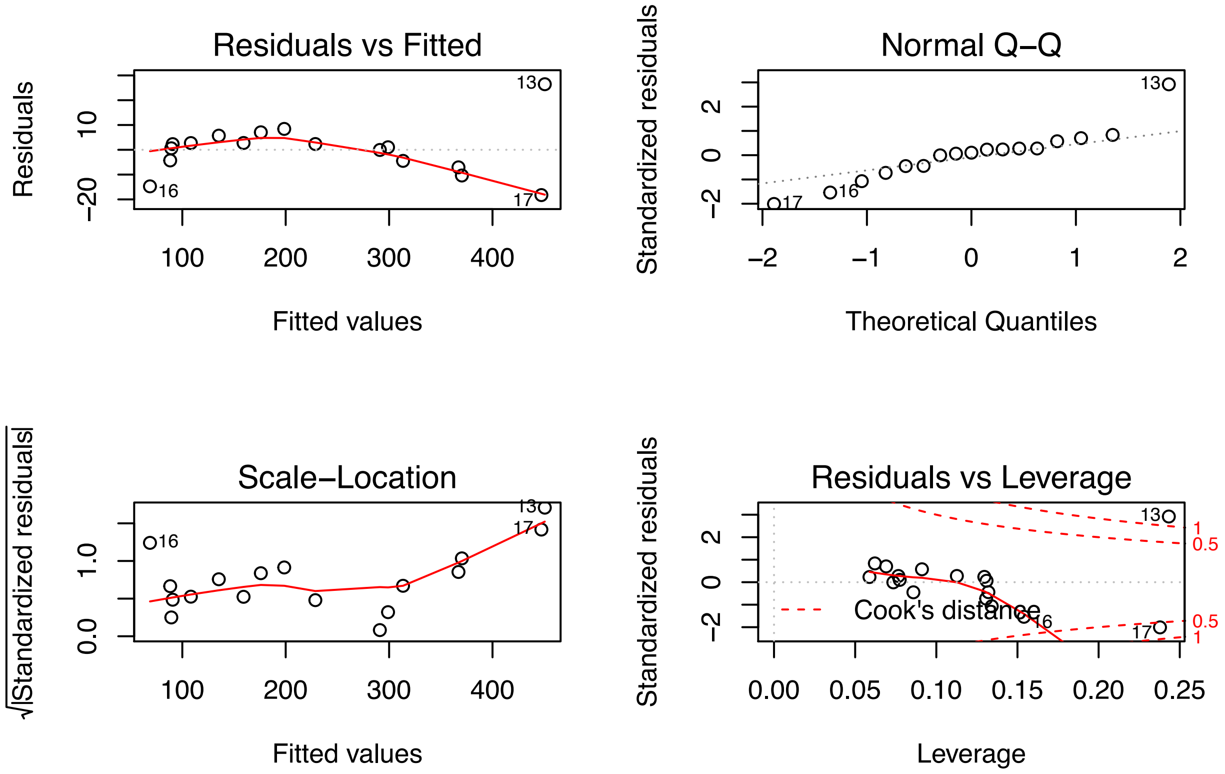
```
##
```

```
## Call:
## lm(formula = Fare ~ Distance, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.265  -4.475   1.024   2.745  26.440
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48.971770   4.405493   11.12 1.22e-08 ***
## Distance     0.219687   0.004421   49.69  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.41 on 15 degrees of freedom
## Multiple R-squared:  0.994,  Adjusted R-squared:  0.9936
## F-statistic:  2469 on 1 and 15 DF,  p-value: < 2.2e-16
```

```
#Check out the fit
plot(dat$Distance, dat$Fare, ylab = "Fare", xlab = "Distance", pch = 2, col = "blue")
abline(model, lty = 2, col = "red")
```



```
#take a look at R's built in residual plots
par(mfrow = c(2,2))
plot(model)
```



a)

The business analyst claims that the relationship is strongly signficant. We cannot make this claim, because this model violates our assumption that we have constant variance. The residual plot clearly indicates that as distance increases, so does our variance. Hence, any hypothesis testing on $\widehat{\beta}_1$ cannot be reliable. The model, however, shows that a *linear* relationship between *Distance* and *Fare* explains 99.4% of the variance. The residuals have a quadratic behavior, so more variance is likely explained by a model containing a quadratic component. This model clearly does not allow us to properly infer the strength of the relationship between these two variables. Moreover, our confidence statements about predictions rely on constant variance of the residuls which we do not have here.

b)

Here, we see the residuals demostrate a quadratic behavior (with the exception of one high leverage point). There are clearly two leverage points in this model, corresponding to point 13 and point 17. Both have large leverage as we see in the residuals vs leverage point. 17 appears to follow the quadratic relationship of the other data point, so if we added this quadratic term in the model, 17 could be a "good" leverage point. 13 is clearly a bad leverage point, with or without the quadratic fit. I would revisit this point to see if there was something that made it different than the other X-values (e.g. international destination). Thus, the linear model assumption is probably not best for this data. Instead a quadratic model should be fit to the data, with possible adding a dummy variable for international vs. domestic flights.

Exercise 3.2

As we showed in class, if we believe the true underlying model is a quadratic model, then $\hat{e}_i \approx \beta_2 x_i^2 + e_i$. Notice that this assumption does not rely on the distribution of Y. Hence if the $\hat{e}_i$ show some quadratic behavior, adding in a quadratic term may help fit the true model. So the statement is *true.*

Depending on the distribution of the residuals, however, to address nonconstant variance may require a Y transformation. Recall under the assumption that the e_i are normal, then $Y|X$ is also normal. Hence to address the nonconstant variance, we may need to go further than just adding in a quadratic term.

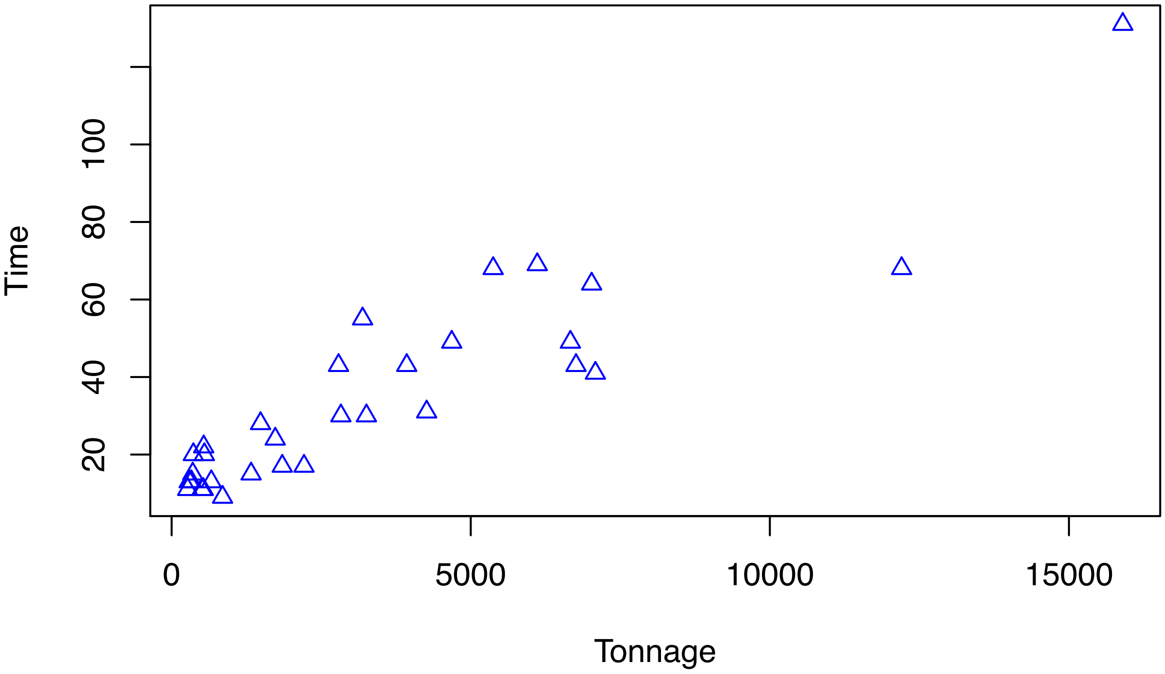
Exercise 3.4

First we will reconstruct the models found in both examples.

```
#read in data
dat = read.table("~/Desktop/Fall 2017/MA 575/book_data/glakes.txt", header = TRUE)

#take a peak
head(dat)
```

```
##   Case Tonnage Time
## 1    1    2213   17
## 2    2    3256   30
## 3    3   12203   68
## 4    4    7021   64
## 5    5     529   11
## 6    6    3192   55
```

```
#plot bivariate relationship
plot(dat$Tonnage, dat$Time, ylab = "Time", xlab = "Tonnage", pch = 2, col = "blue")
```

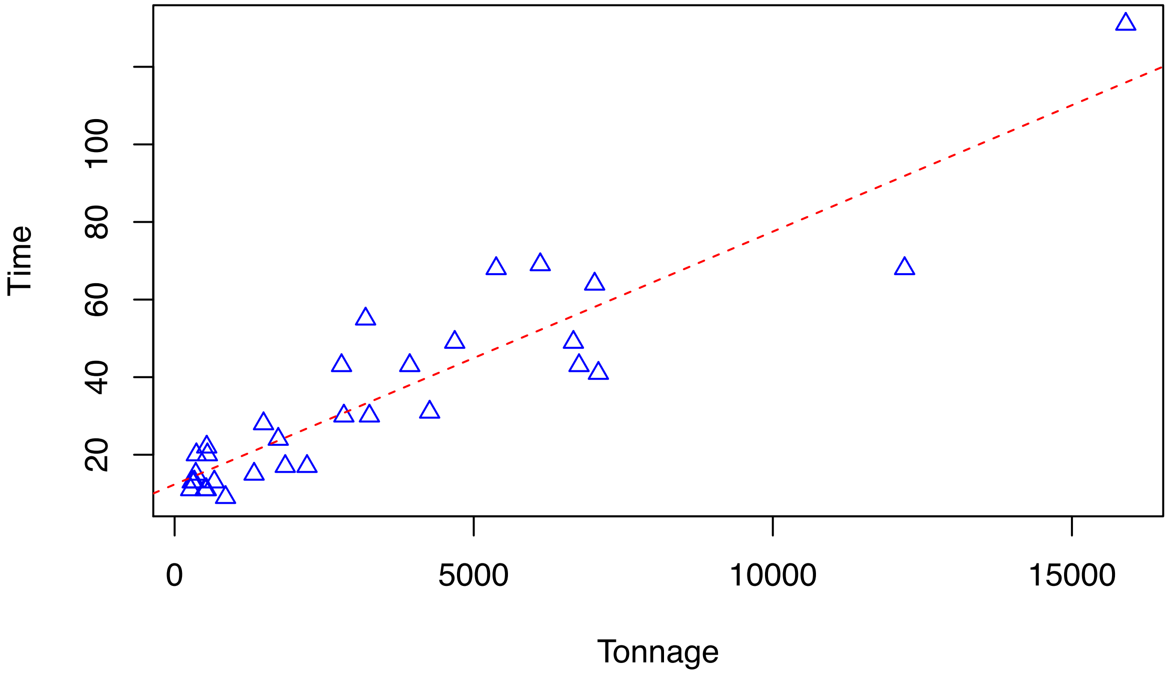


```
#build regression model
model = lm(Time~Tonnage, data = dat)

#check out summary statistics
summary(model)
```

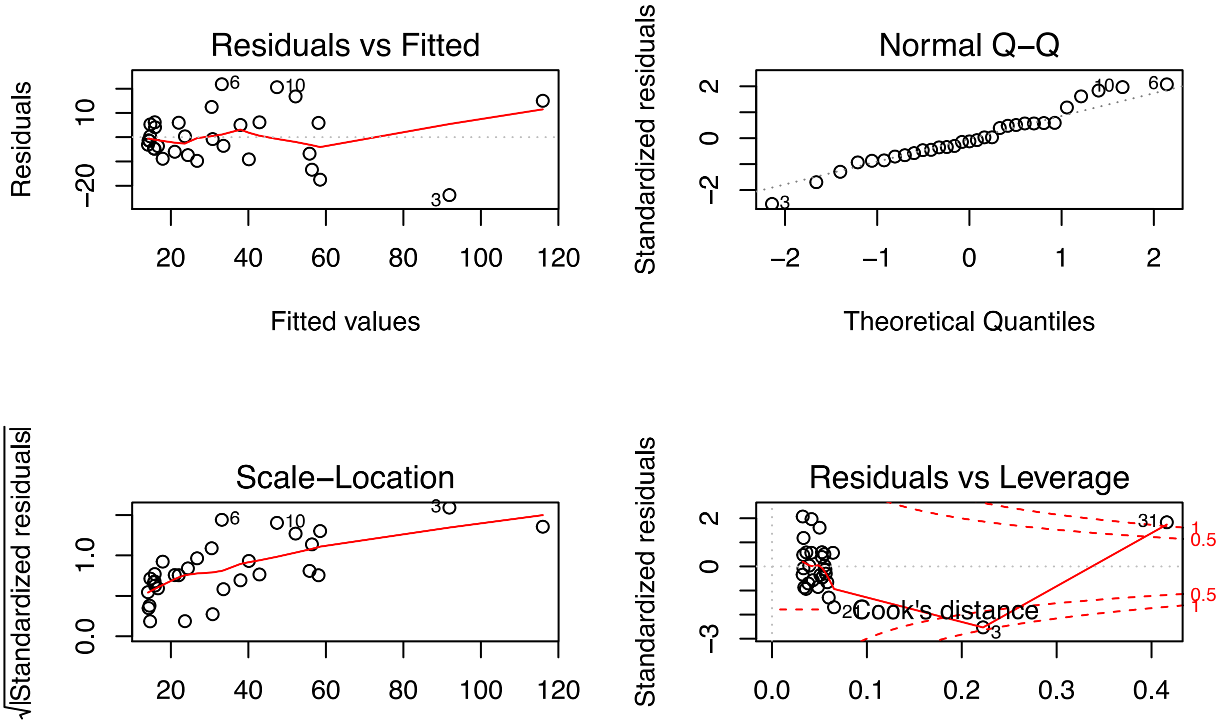
```
##
## Call:
## lm(formula = Time ~ Tonnage, data = dat)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.882  -6.397  -1.261   5.931  21.850
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.344707   2.642633   4.671 6.32e-05 ***
## Tonnage      0.006518   0.000531  12.275 5.22e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.7 on 29 degrees of freedom
## Multiple R-squared:  0.8386, Adjusted R-squared:  0.833
## F-statistic: 150.7 on 1 and 29 DF,  p-value: 5.218e-13
```

```
#Check out the fit
plot(dat$Tonnage, dat$Time, ylab = "Time", xlab = "Tonnage", pch = 2, col = "blue")
abline(model, lty = 2, col = "red")
```



```
#take a look at R's built in residual plots
par(mfrow = c(2,2))
plot(model)
```



a)

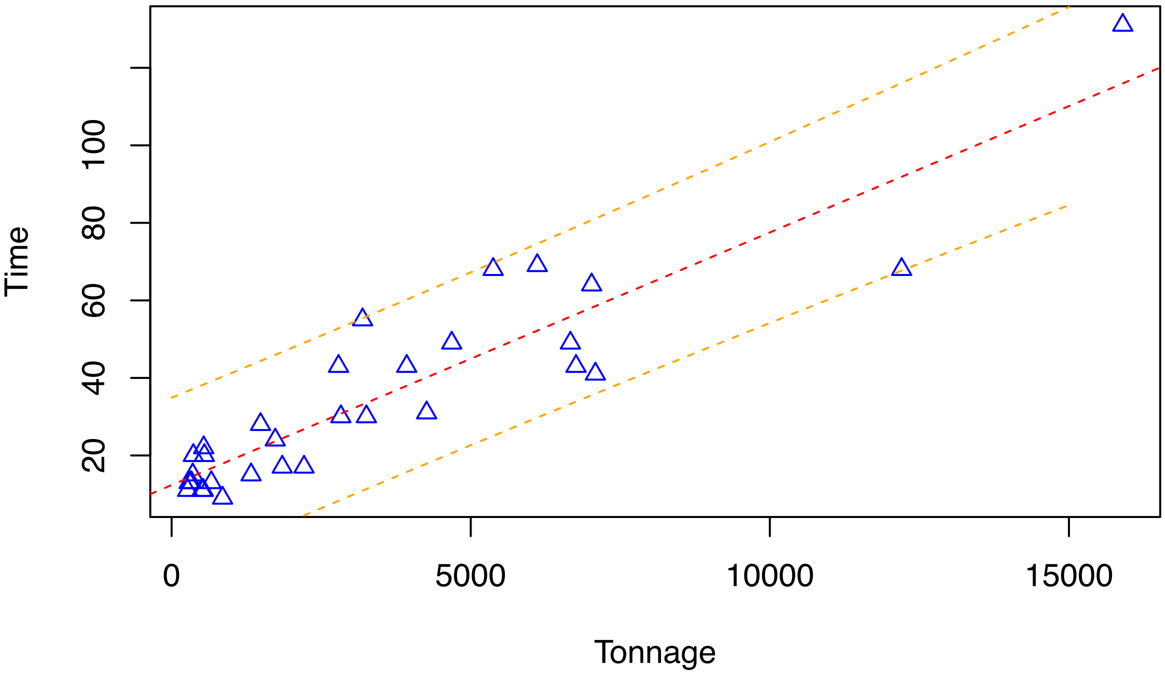
While the data does roughly follow a linear trend, we see that the regression model's assumptions are violated. That is as *Tonnage* increases so does the variance. Moreover, there are two very high leverage points both with very high influence. So while a linear model is probably the correct choice, some adjustments need to be made to make this model valid.

b)

Recall we constructed our prediction intervals off the assumption that there was constant variance. But in this case, as *Tonnage* increases, so does the variance. Since the prediction interval will assume constant variance, there is no way for the interval to accommodate the growing variance for $Tonnage = 10,000$. For this reason, we expect the interval will be too short.

To test this argument, we can plot the prediction intervals as follows

```r
#Check out the fit
plot(dat$Tonnage, dat$Time, ylab = "Time", xlab = "Tonnage", pch = 2, col = "blue")
abline(model, lty = 2, col = "red")

pred = predict(model, newdata = data.frame(Tonnage = seq(0,15000, 1000)), interval = "predict")

lines(seq(0,15000, 1000), pred[,2], col="orange", lty=2)
lines(seq(0,15000, 1000), pred[,3], col="orange", lty=2)
```

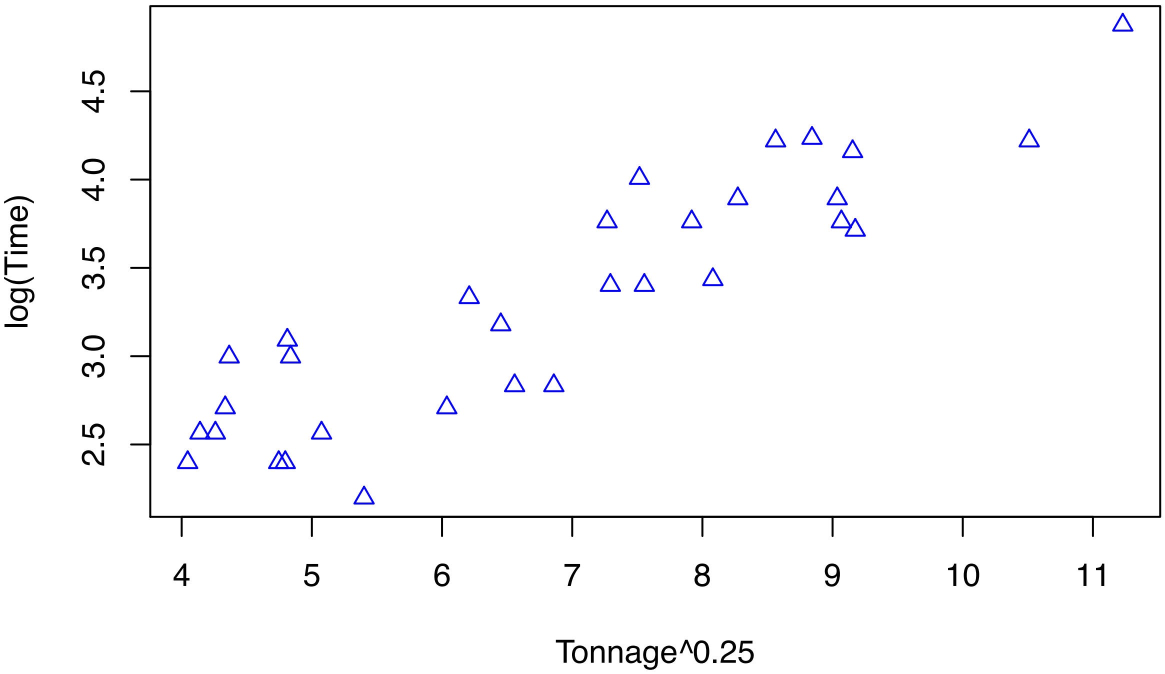


Notice as *Tonnage* increases, the observed values begin to inch closer to the prediction intervals, depicted by the orange lines. This corresponds to the prediction interval not taking into account the increasing variance.

c)

We now fit the second model in this problem.

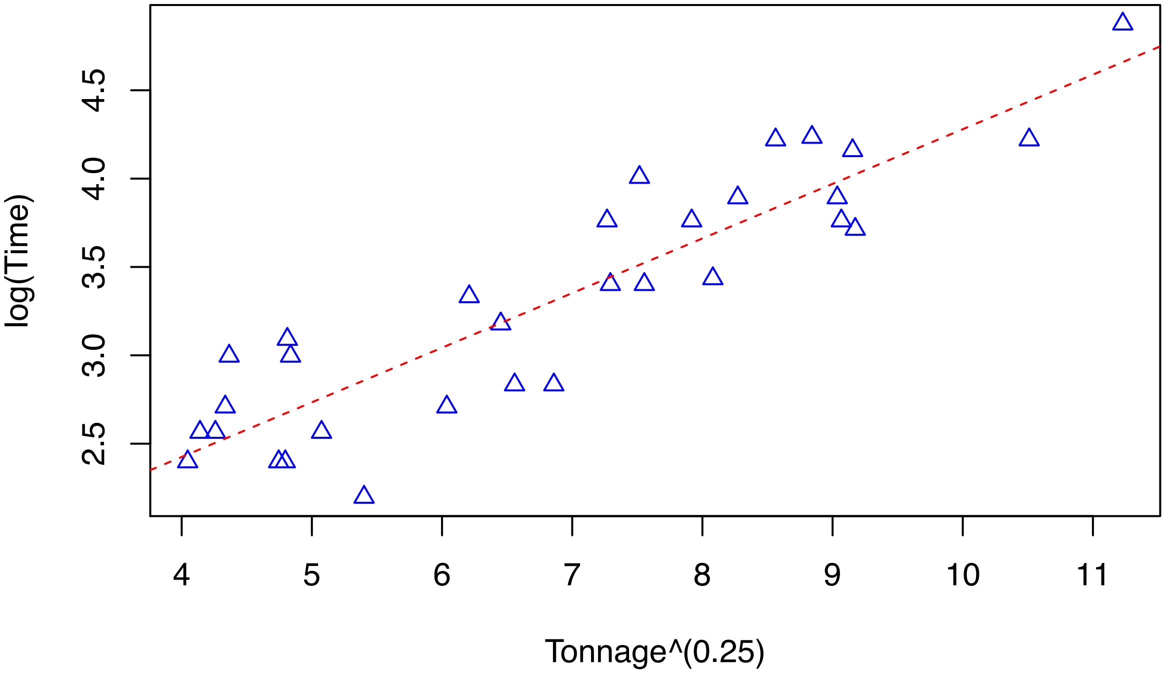
```
#plot bivariate relationship
plot(dat$Tonnage^(0.25), log(dat$Time), ylab = "log(Time)", xlab = "Tonnage^0.25", pch = 2, col = "blue
```



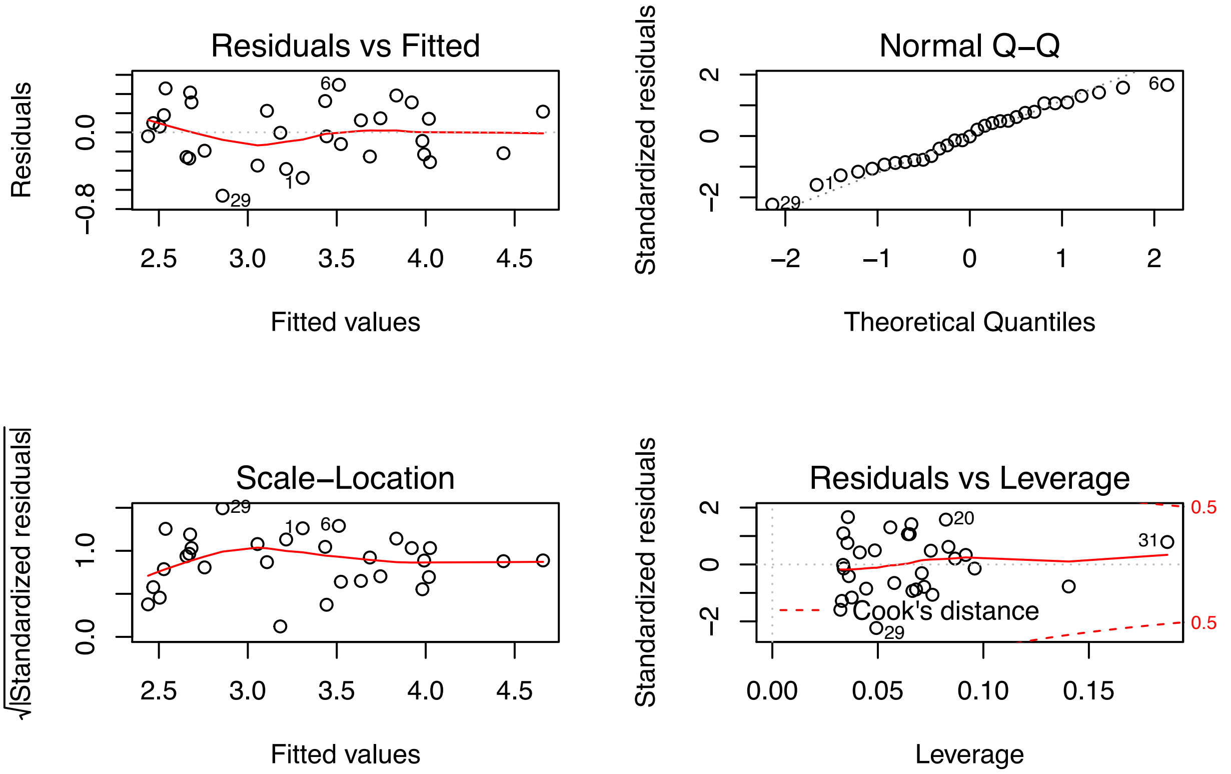
```r
#add fourth root variable
dat$Tonnage2 = dat$Tonnage^(1/4)

#build regression model
model = lm(log(Time)~Tonnage2, data = dat)

#check out summary statistics
summary(model)
```

```
##
## Call:
## lm(formula = log(Time) ~ Tonnage2, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6607 -0.2410 -0.0044  0.2203  0.4956
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.18842    0.19468   6.105  1.2e-06 ***
## Tonnage2     0.30910    0.02728  11.332  3.6e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3034 on 29 degrees of freedom
## Multiple R-squared:  0.8158, Adjusted R-squared:  0.8094
## F-statistic: 128.4 on 1 and 29 DF,  p-value: 3.599e-12
```

```r
#Check out the fit
plot(dat$Tonnage^(0.25), log(dat$Time), ylab = "log(Time)", xlab = "Tonnage^(0.25)", pch = 2, col = "blu
abline(model, lty = 2, col = "red")
```



```r
#take a look at R's built in residual plots
par(mfrow = c(2,2))
plot(model)
```



There appears to be a large improvement for this model over the previous model. Our residuals show that the model has (relatively) constant variance and the trend is linear. Moreover, the leverage points in this model are “good” leverage points since they follow the trend suggested by the rest of the data. Since we have these features, the prediction intervals will be valid for time.

d)

There could be some normality issues with the *Tonnage* variable towards the tail (see Normal Q-Q plot). This could be fine tuned by changing the power transformation of X. I would not recommend this, however, due to possible overfitting.

Ben Draves MA 575: HW3 September 25, 2017

1. (a) Recall that our motivation for defining the leverage was to see the effect of each point (x_i, y_i) in our sample on a predicted value $\hat{y}_i$. Thus, expanding our estimate of $\hat{y}_i$ we have

$$\begin{aligned}
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\
&= \overline{y} - \hat{\beta}_1\overline{x} + \hat{\beta}_1 x_i \\
&= \overline{y} + \hat{\beta}(x_i - \overline{x}) \\
&= \overline{y} + \sum_{j=1}^{n} c_j y_j (x_i - \overline{x}) \\
&= \frac{1}{n}\sum_{j=1}^{n} y_j + (x_i - \overline{x})\sum_{j=1}^{n}\frac{(x_j - \overline{x})}{SXX} y_j \\
&= \frac{1}{n}\sum_{j=1}^{n} y_j + (x_i - \overline{x})\sum_{j=1}^{n}\frac{(x_j - \overline{x})}{SXX} y_j \\
&= \sum_{j=1}^{n}\left(\frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{SXX}\right) y_j
\end{aligned}$$

From here, we define $h_{ij} = \frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{SXX}$ and the leverage of point i as $h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{SXX}$. This is the (marginal) effect of point y_i on the predicted value $\hat{y}_i$.

(b) High leverage points can be identified by looking at which points maximize h_{ii}. $\frac{1}{n}$ and SXX are constants in the data set, so we look for points that have large $(x_i - \overline{x})^2$. In a plot, we can find these by looking at those x values are that are far from the mean of x. That is, points with high leverage will be those that are absolutely "far" from the sample mean, $\overline{x}$.

(c) Suppose we are given a data set $\{(x_i, y_i) : i = 1, 2, \dots, n\}$. Then, having $\overline{x}$ and SXX fixed, if there exists a point x_i with $h_{ii} = 1$ then

$$\begin{aligned}
1 &= \frac{1}{n} + \frac{(x_i - \overline{x})^2}{SXX} \\
\frac{n-1}{n} SXX &= (x_i - \overline{x})^2 \\
\sqrt{\frac{n-1}{n} SXX} &= x_i - \overline{x} \\
\sqrt{\frac{n-1}{n} SXX} + \overline{x} &= x_i
\end{aligned}$$