ProblemSet5

Aaron Graybill

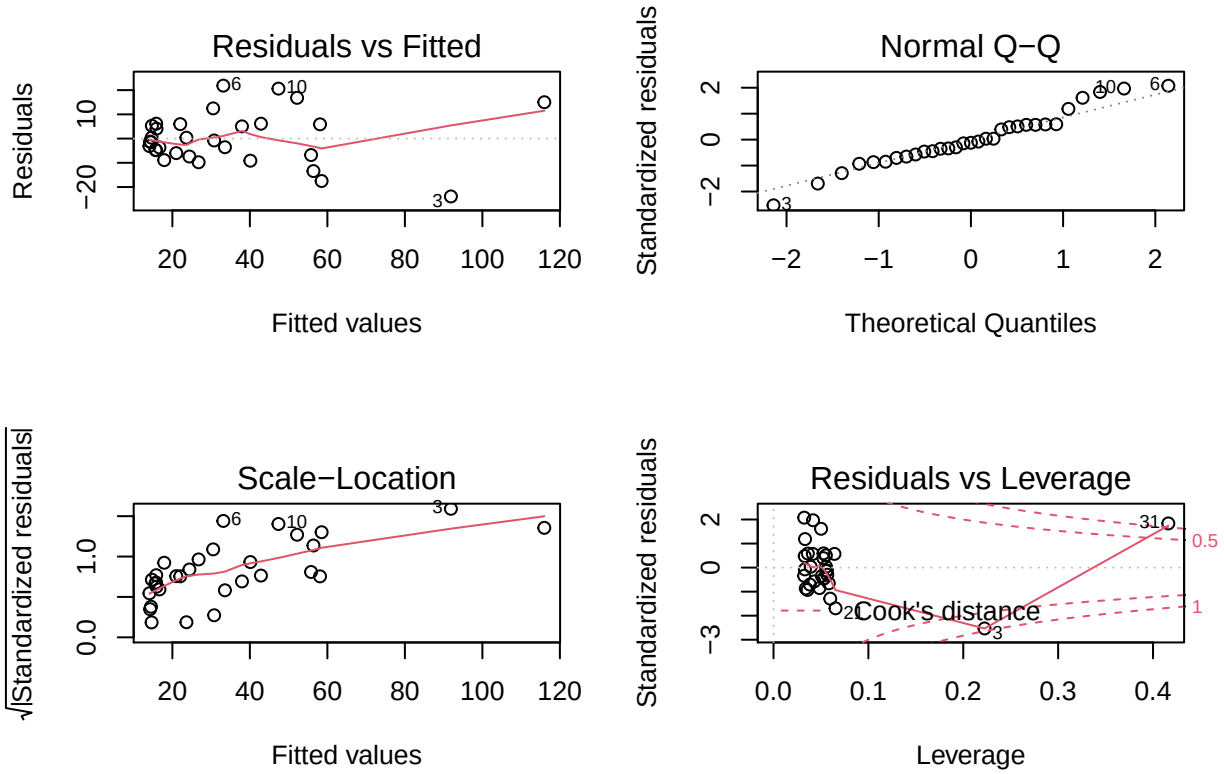
3/22/2021

Homework #5: Chapter 3: #4, #6, #7, #8

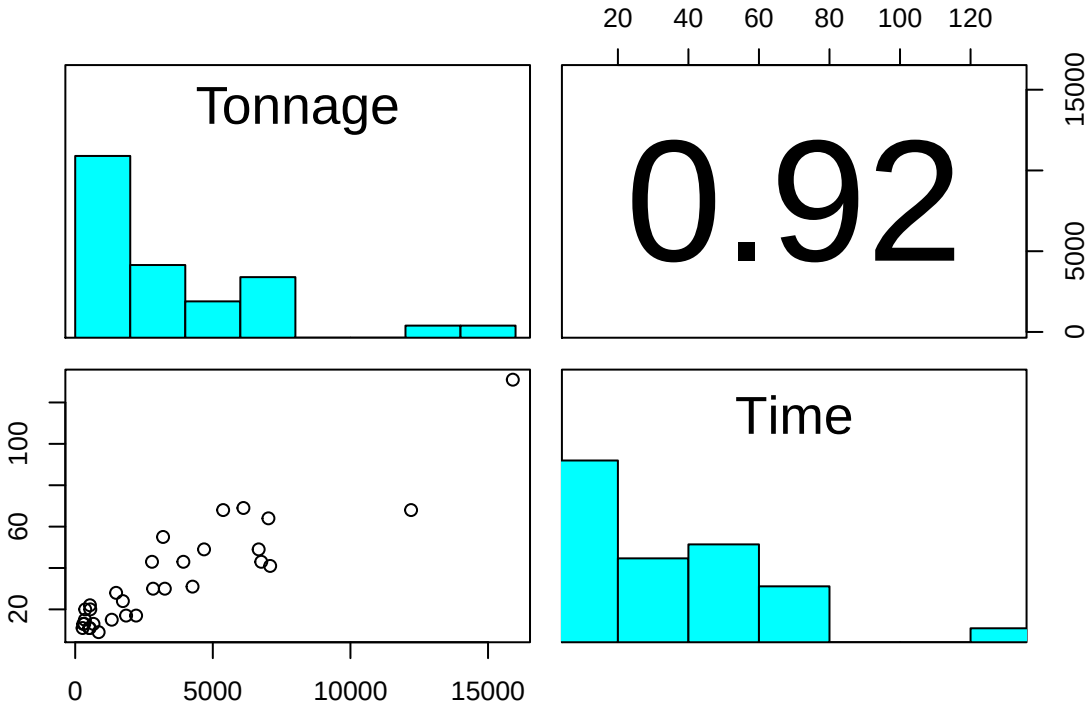
Problem 4

a.

```
## here() starts at /Users/aarongraybill/Documents/Haverford Stuff/Math/Math286
```



It doesn't look terrible, but there is clearly room for improvement. The scale location plot inidicates that the model performs poorly for high values of the dependent variable. Maybe we should investigate the distributions



to see if there is an issue there:

The above diagram makes it quite clear that the distribution of both variables is problematic. The data is clustered in th bottom left of the scatter plot, so we are skew right on both variables.

b.

I belive the interval would be small and understate the variance in the prediction intervals. I don't have much explanation for this other than something like Jensen's inequality that since we must do a concave function on the data. Wait let me try this. Let $f(x)$ be the optimal transformation. It will be concave by the distribution of the data. Let $d(x) = P_2(x) - P(x)$ where the P_is are the the upper and lower of the prediction interval. Jensen's inequality guarantees that for f concave $f(d(x)) \leq d(f(x))$. Yeah something like that. It's obviously not a proof, but something to that effect.

c.

The new model is a drastic improvement, all of the graphs look close to ideal if the assumptions were satisfied. THe residual scale plot in particular now seems to be constant across the predicted values of the dependent variable. Moreover, the density plots look much more normal.

d.

Honestly, I'm not seeing any shortcomings that could easily be overcome, the distributions could be more normal, but no power transform could fully fix the distribution function is not concave (it goes down and then back up), so we will always have some non-normalities.

Problem 6.

x is given to be very skewed which makes estimating the ideal $g(\cdot)$ quite challenging or it might make the process a biased estimator, I'm not certain.

Problem 7.

Suppose, $E(Y) = \mu$ and $Var(Y) = \mu^2$. Taking a Taylor expansion of Y around $E[Y]$ gives:

$$f(Y) = f(E[Y]) + f'(E[Y])(Y - E[Y]) + \dots$$

Now to find the variance of $f(Y)$ let's consider just those first two terms:

$$Var(f(y)) \approx Var\left(f(E[Y]) + f'(E[Y])(Y - E[Y])\right) = f'(E[Y])^2 Var(Y) = f'(\mu)^2\mu^2$$

Ooh! And now we have to solve a (rather simple) differential equation to find the f satisfying $f'(\mu)^2\mu^2 = c$ where c is some constant. Solving for f' gives:

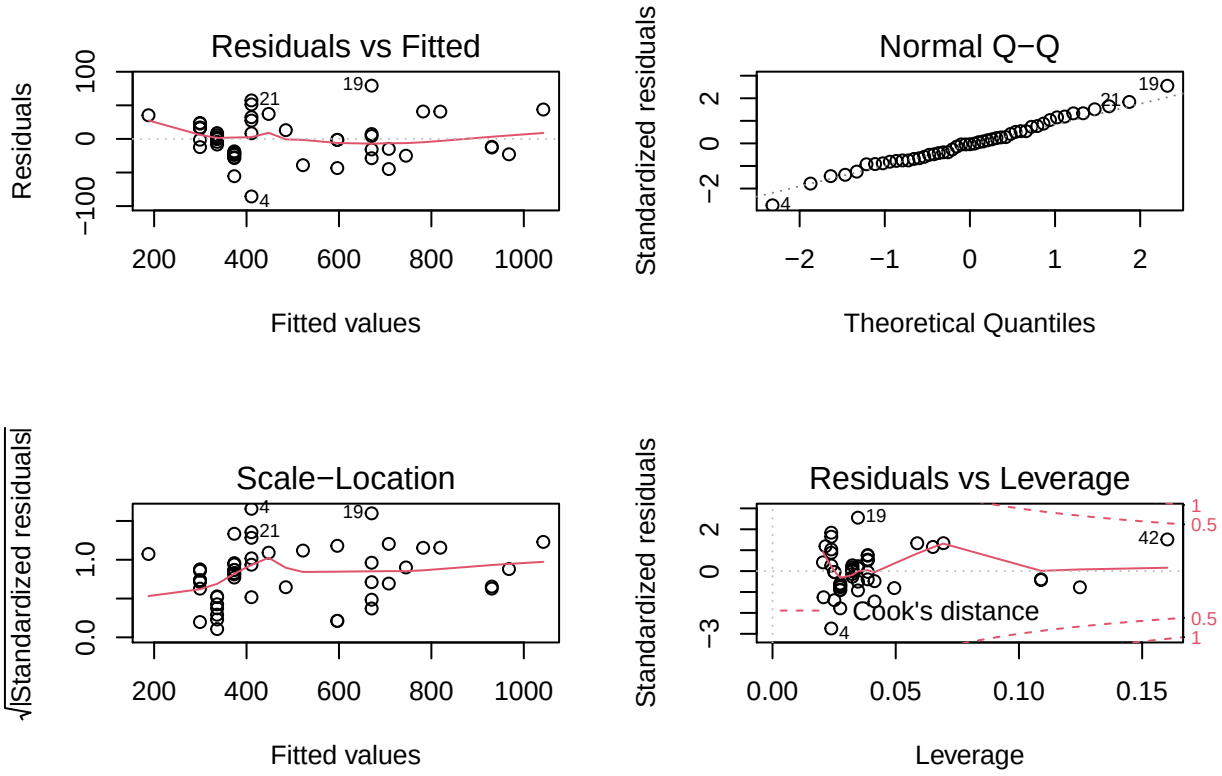
$$\sqrt{f'(\mu)^2} = \sqrt{c} \implies f'(\mu) = \frac{\sqrt{c}}{\mu}$$

And what is the function that has the property that its derivative is proportional to 1 over its input? Well that's exactly ln. The natural log will suffice to normalize variance.

Problem 8 Part 1.

a.

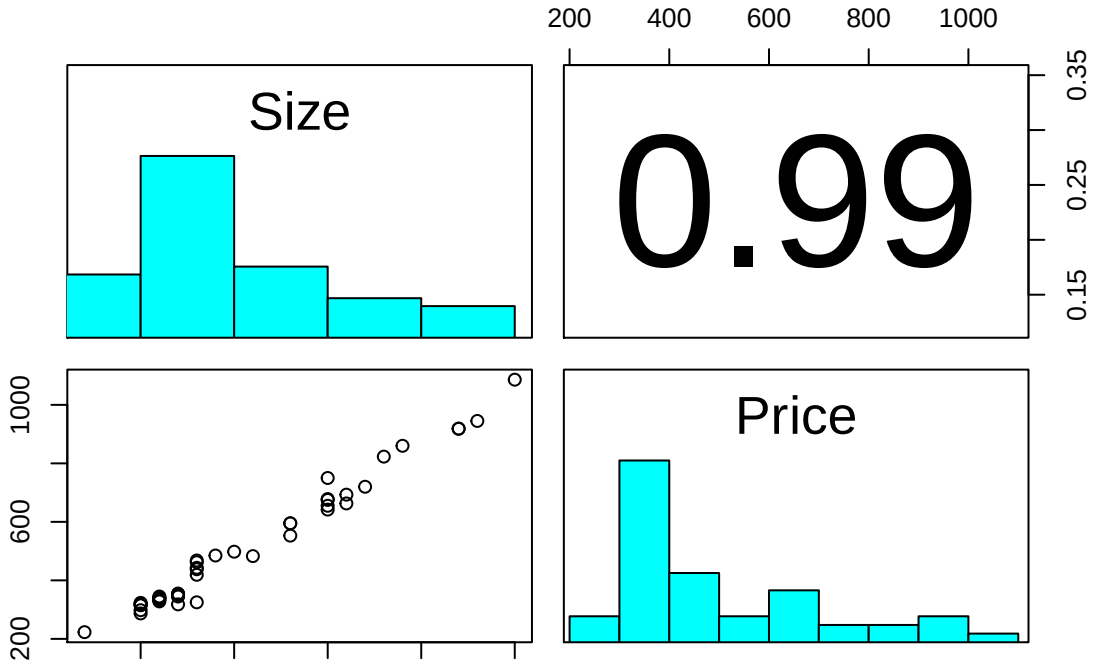
```
##
## Call:
## lm(formula = Price ~ Size, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.654 -21.503  -1.203  16.797  79.295
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -258.05      16.94  -15.23   <2e-16 ***
## Size         3715.02      80.41   46.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.6 on 47 degrees of freedom
## Multiple R-squared:  0.9785, Adjusted R-squared:  0.978
## F-statistic:  2135 on 1 and 47 DF,  p-value: < 2.2e-16
```



I'm being asked to provide justificatio for this model but it also tells me what model to run, so I'm a bit constrained! That being said the simple linear regression with no transformations is always a good place to start. Furthermore, looking at the scatter plot below shows that the relationship seems fairly linear.

b.

Absent another point of reference, this regression seems relatively okay. The residuals vs fitted values plot seems to have a bit of an upward bow, but the $Q-Q$ plot looks good. The scale location plot seems to be upwards sloping. These are some weaknesses, we can look at the distribution of the variables to see if they are problematic



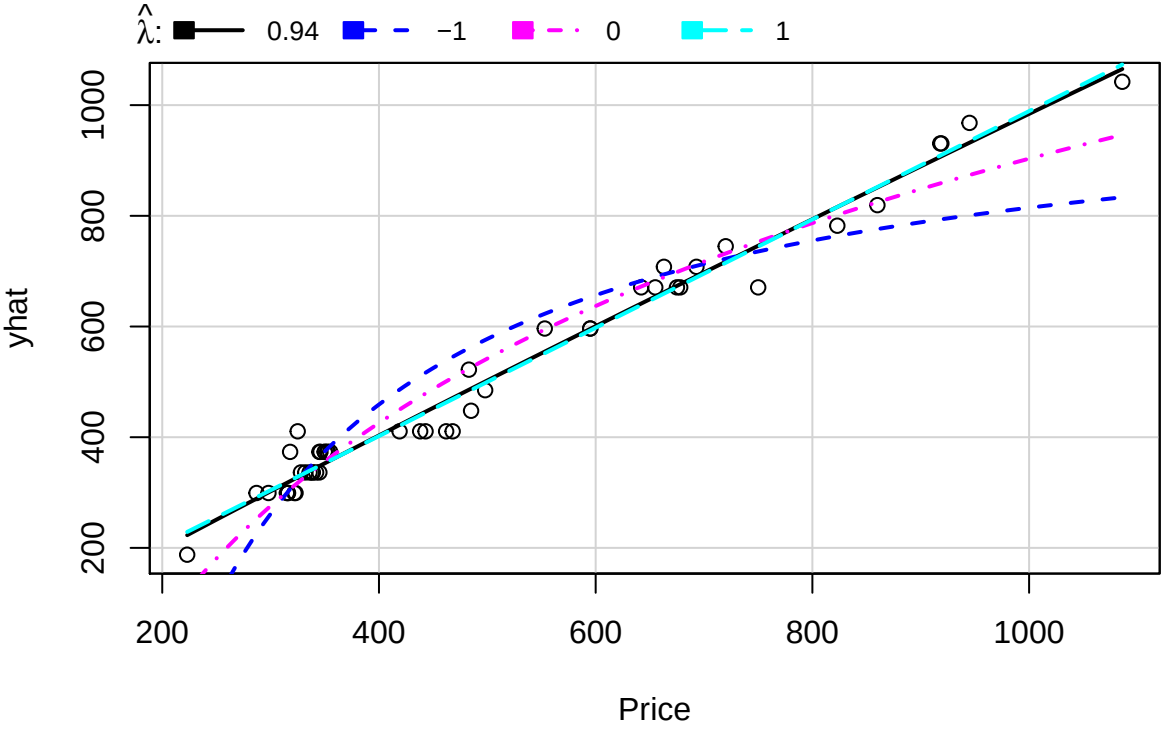
as below:

Both distributions seem a little skew right, so the normality of errors assumption is likely not satisfied.

Problem 8 Part 2

a.

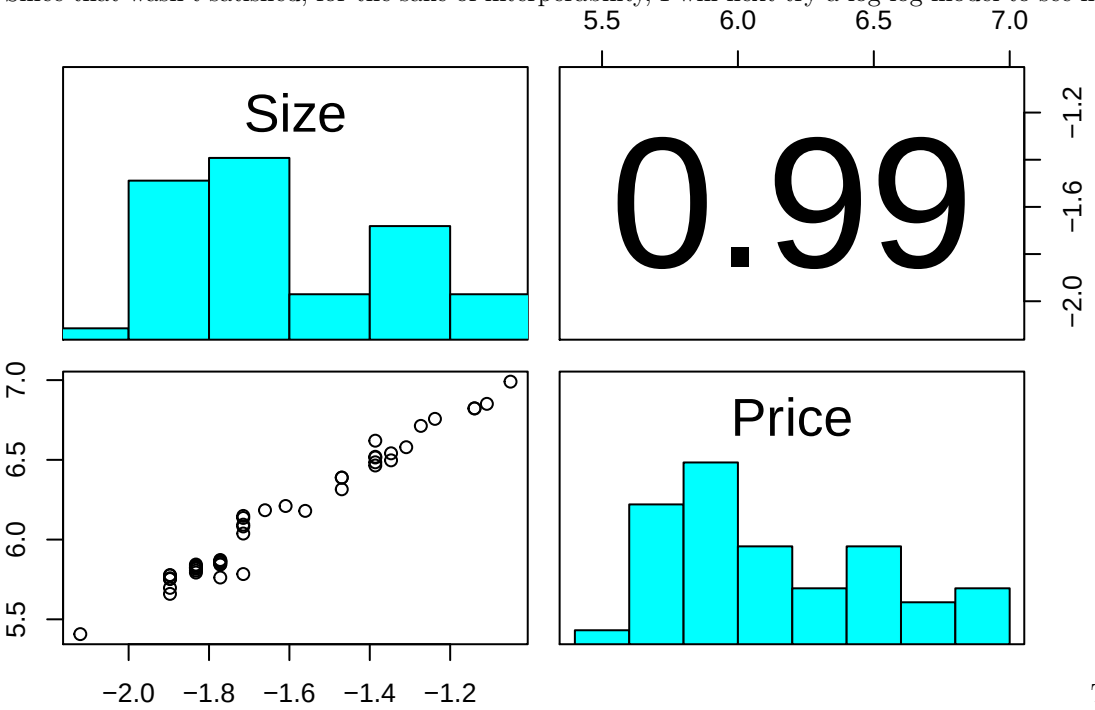
I first test an inverse response plot to see if there is an obvious transformation. There is not.




```
##       lambda       RSS
## 1  0.9376257  45670.12
```

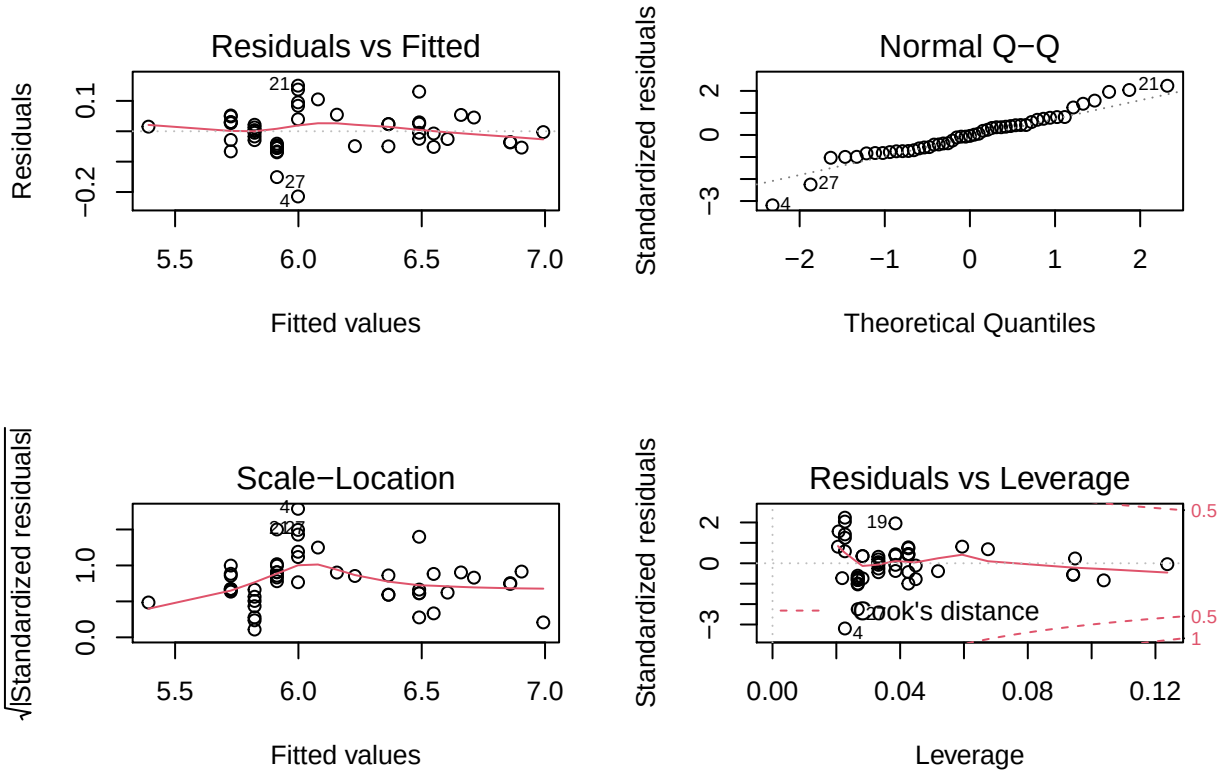
```
## 2 -1.0000000 272143.61
## 3  0.0000000 101071.53
## 4  1.0000000  45918.17
```

Since that wasn't satisfied, for the sake of interperability, I will next try a log-log model to see how that looks:



That seems to have evened out the distributions a little and since this is an easy interpretation, this seems like a suitable model to run. This is not a count variable, so we don't need to use square root transformations. I have no inclination for why percentage changes in size should induce percentage changes in size, but nor does a strictly linear model seem any more appropriate, so let's proceed.

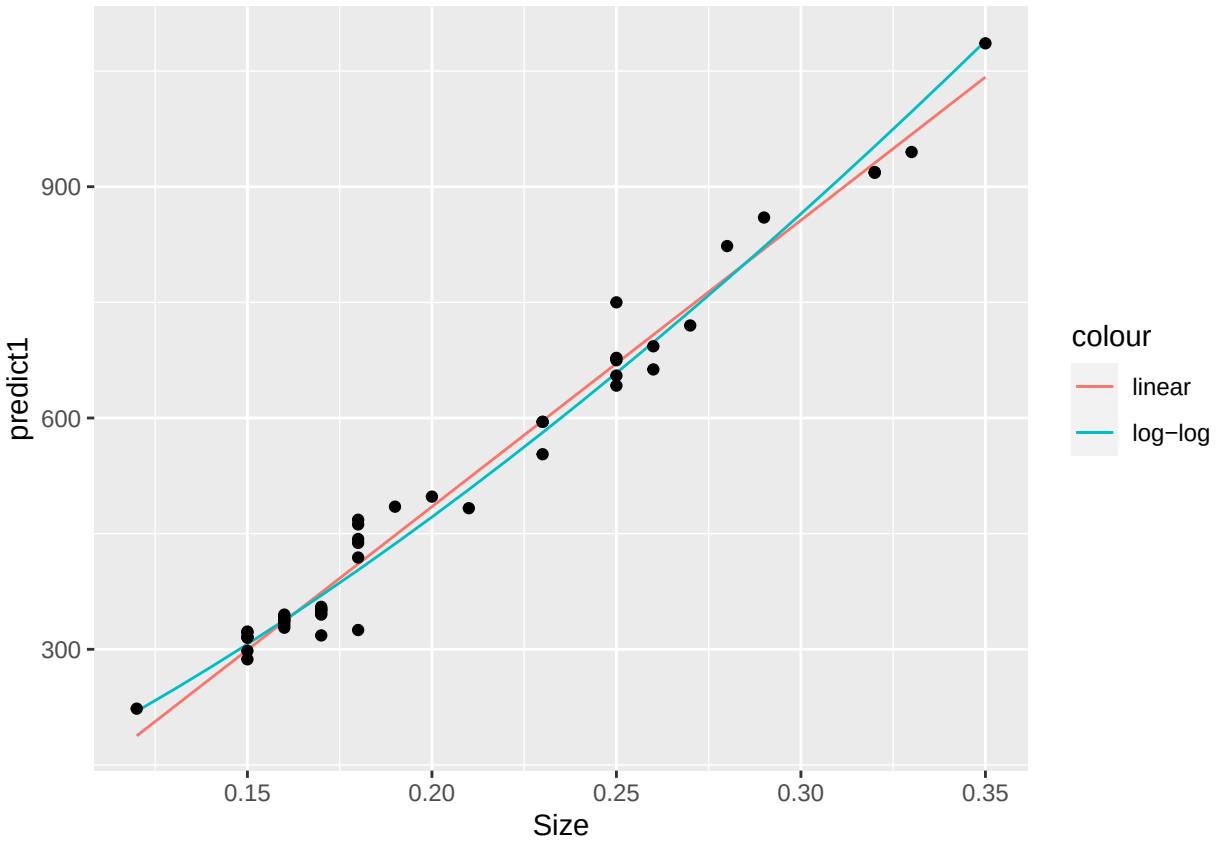
```
##
## Call:
## lm(formula = log_Price ~ log_Size, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21460 -0.04646 -0.00274  0.03001  0.15005
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.56317    0.06221  137.65   <2e-16 ***
## log_Size     1.49566    0.03772   39.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06796 on 47 degrees of freedom
## Multiple R-squared:  0.971,  Adjusted R-squared:  0.9704
## F-statistic:  1572 on 1 and 47 DF,  p-value: < 2.2e-16
```



b. The quantile plots actually look worse, but the other plots look better. The residuals have no discernable relation to the fitted values and nor do the standardized residuals. There are still some rather leveraged points, but otherwise the plots look okay.

Problem 8 Part 3

The two outputs have their pluses and minuses to compared to one another. With no immediately discernable winner, I will plot the two and inspect which one looks better visually.



Okay well all that work was basically for naught because we have so few datapoints it's difficult to elucidate which model is better. I will say that the log-log model is better because it seems more homoskedastic.