ProblemSet6

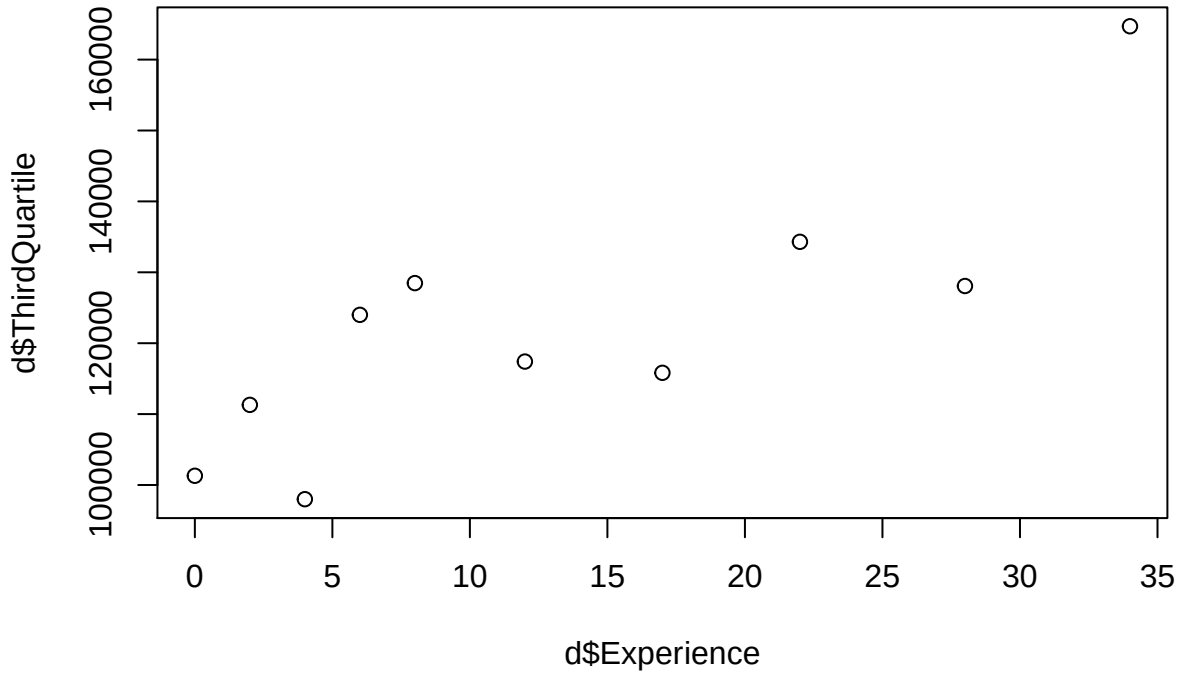
Aaron Graybill

3/30/2021

Problem 1.

This question is asking us to find the estimate of Third Quartile income with six years of experience. I'll report the confidence interval to be safe.

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




```
##
## Call:
## lm(formula = ThirdQuartile ~ Experience, data = d, weights = SampleSize)
##
## Weighted Residuals:
##    Min     1Q Median     3Q    Max
## -67520 -40994   4937  51648  87516
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 104759.0     5752.2   18.21 8.49e-08 ***
## Experience    1172.5      336.9    3.48  0.00832 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 57620 on 8 degrees of freedom
## Multiple R-squared:  0.6022, Adjusted R-squared:  0.5524
## F-statistic: 12.11 on 1 and 8 DF,  p-value: 0.008323
```

```
##        fit      lwr      upr
## 1 111793.8 102013.3 121574.3
```

The predicted salary at six years is $\$102,013.30$.

Problem 2.

The question asks us to minimize first wrt $\hat{\beta}$

$$RSS_w = \sum_{i=1}^{n} w_i \hat{e}_i^2 = \sum_{i=1}^{n} w_i \left(y_i - \hat{\beta} x_i\right)^2$$

There is one normal equation which is:

$$\begin{aligned} 0 &= \sum_{i=1}^{n} w_i \left(-2x_i\right)\left(y_i - \hat{\beta} x_i\right) \\ &= \sum_{i=1}^{n} w_i y_i x_i - \sum_{i=1}^{n} w_i \hat{\beta} x_i^2 \\ \sum_{i=1}^{n} w_i y_i x_i &= \hat{\beta} \sum_{i=1}^{n} w_i x_i^2 \\ \hat{\beta}^* &= \frac{\sum_{i=1}^{n} w_i y_i x_i}{\sum_{i=1}^{n} w_i x_i^2} \end{aligned}$$

what we need now is w_is that ensure $\hat{\beta}^* x_i$ has constant variance. We need $Var(w_i y_i | x_i) = \sigma^2$, but given the model assumptions that reduces to: $w_i^2 x_i^2 \sigma^2 = \sigma^2$ or $w_i = \frac{1}{x_i^2}$. Nice, we can plug that into the expression above to give:

$$\begin{aligned} \hat{\beta}^* &= \frac{\sum_{i=1}^{n} \frac{1}{x_i^2} y_i x_i}{\sum_{i=1}^{n} \frac{1}{x_i^2} x_i^2} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i} \end{aligned}$$

Problem 3.

a.

Weighting is necessary in this case becauese we are using a data aggregate. We do not have one datapoint per house sold, so even if the variance in individual house prices was itself constant in the predictor variables, since we are aggregating those variances would no longer be the same. Think for example if we have a very large number of houses aggregated into one, we would expect the variance of that aggregate and the true value y_i to be much lower than the variance between an individual observation. This comes from something like the central limit theorem.

Weighting by n is appropriate because it takes into account how many datapoints each category has and effectively undos the effect of the aggregation. A quick bit of wikipedia-ing shows that asymptotically the distribution of the sample median for density $f(y)$ is

$$\frac{1}{4nf(m)^2}$$

Which note that when we multiply by n the variance is constant. Although maybe that's not true bc it seems like it should also be proportional to the median, but that's as much as I know.

b.

The regression model is inappropriate because the fitted values against residuals show clear non-linearity. We do not see circular errors, we see all of the smooshed up against the higher values of Y another indication of an issue would be with the first two plots which show that there are much more values in the left half than the right half, so we need a transformation so the data is less smooshed.

c.

I would start by thinking about how each of the variables is measured. Noting that none of them are count variables, I would skip a square root transform. I would then look at histograms of each of the variables to see which violate the normality assumption. From there I would transform those variables with a log transform and then if that did not work I would try an inverse response plot.