# MA 575: HW4

*Benjamin Draves*

*October 3, 2017*

### Exercise 4.1

Here, we build a weighted regression model for third quartile of full professor salaries on years of experience. The weights correspond to $w_i = 1/n_i$ where the $n_i$ correspond to the number of professors sampled for each fixed years of experience value. We fit the model below.

```
#read in data
dat = read.table("~/Desktop/Courses/MA 575/book_data/ProfessorSalaries.txt", header = T)

#take a peak
head(dat)
```

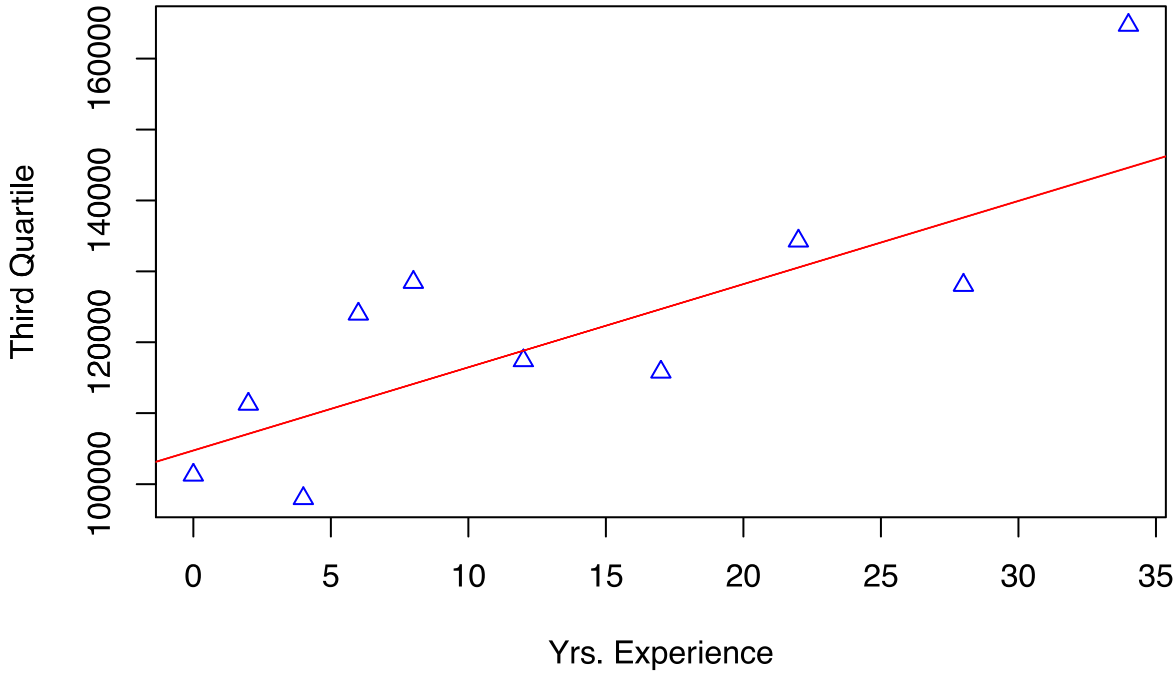
```
##   Experience SampleSize ThirdQuartile
## 1          0         17        101300
## 2          2         33        111303
## 3          4         19         98000
## 4          6         25        124000
## 5          8         18        128475
## 6         12         60        117410
```

```
#look at the bivariate relationship
plot(dat$Experience, dat$ThirdQuartile, main = "", pch = 2, col = "blue", xlab = "Yrs. Experience", ylab

#Define weights for WLS
w = dat$SampleSize

#Build WLS
wlm = lm(ThirdQuartile~Experience, data = dat, weights = w)

#check out fit
plot(dat$Experience, dat$ThirdQuartile, main = "", pch = 2, col = "blue", xlab = "Yrs. Experience", ylab
abline(wlm, col = "red")
```



```
#get fit statistics
summary(wlm)
```

```
##
## Call:
## lm(formula = ThirdQuartile ~ Experience, data = dat, weights = w)
##
## 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
```

Hence, the expected third quartile salary of a sixth year full professor is given by $104,759.0 + 1172.5x^* = 104,759.0 + 1172.5(6) = \$111,794$.

1. (Exercise 4.2) Consider the regression model $Y_i = \beta x_i + e_i$ with $\mathbb{V}(e_i|x_i) = x_i^2\sigma^2$. We can translate this to a weighted regression problem by letting $w_i = 1/x_i$. Using these weights, we have the regression model $w_iY = w_i\beta x_i + w_ie_i$. Then, given $\{x\}_{i=1}^n$ we have

$$\mathbb{V}(w_iY_i|x_i) = \mathbb{V}(w_ie_i|x_i) = w_i^2\mathbb{V}(e_i|x_i) = \frac{1}{x_i^2}x_i^2\sigma^2 = \sigma^2$$

That is we have constant variance. So we can use the original RSS formulation with this adjusted model to estimate $\beta$.

$$\begin{aligned} WRSS(b) &= \sum_{i=1}^n (w_iy_i - w_ibx_i)^2 \\ &= \sum_{i=1}^n w_i^2 (y_i - bx_i)^2 \end{aligned}$$

Now our estimate of $\beta$ will be given by $\arg\min_b WRSS(b)$. We will find this minimum by differentiating $WRSS$ with respect to $b$ and setting the result equal to zero.

$$\frac{d}{db}WRSS(b) = -2\sum_{i=1}^n x_iw_i^2(y_i - bx_i) \overset{set}{=} 0$$

$$b\sum_{i=1}^n (x_iw_i)^2 = \sum_{i=1}^n x_iy_iw_i^2$$

$$\widehat{\beta} = \frac{\sum_{i=1}^n x_iy_iw_i^2}{\sum_{i=1}^n (x_iw_i)^2} = \frac{1}{n}\sum_{i=1}^n \frac{y_i}{x_i}$$

2. (Exercise 4.3)

   (a) By collecting our data by subdivision, we add another level of complexity to our model. In some cases, subdivisions may have a small sample size, while others have quite a few data points. Our data was aggregated at the division level, so to account for the differing number of samples, we should give more weight in our model to those points with several samples and less weight to those with small sample size. This of course corresponds to $w_i = n_i$.

   (b) It appears that the response variable $Y$ is highly right skewed. Our linear model cannot take into account this higher order growth. The fitted values vs standardized residuals plots show this trend. While the variance appears to be constant, our trend assumptions is incorrect.

   (c) I would use a log, or square root transformation of $Y$. In this way, we could better address the trend issue. The explanatory variables are in the form of percentages, so our estimate coefficients could be interpreted as % changes.

3. First note Joe's model is just OLS. Next note that for Sue, $\overline{x}_w = \sum_{i=1}^n w_i x_i \big/ \sum_{i=1}^n w_i x_i = 2\sum_{i=1}^n x_i/2n = \overline{x}$. A similar result shows for Sue's model that $\overline{y}_w = \overline{y}$. Now, for Sue, the WLS coefficients are given by

$$\widehat{\beta}_{1WLS} = \frac{\sum_{i=1}^n w_i(x_i - \overline{x}_w)(y_i - \overline{y}_w)}{\sum_{i=1}^n w_i(x_i - \overline{x}_w)^2} = \frac{2\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{2\sum_{i=1}^n (x_i - \overline{x})^2} = \widehat{\beta}_{1OLS}$$

$$\widehat{\beta}_{0WLS} = \overline{y}_w - \widehat{\beta}_{1WLS}\overline{x}_W = \overline{y} - \widehat{\beta}_{1OLS}\overline{x} = \widehat{\beta}_{0OLS}$$

This shows that the coefficients do not differ. Hence, since variance is a well defined function we have that

$$\mathbb{V}(\widehat{\beta}_{1WLS}|X) = \mathbb{V}(\widehat{\beta}_{1OLS}|X) = \frac{\sigma^2}{SXX}$$

$$\mathbb{V}(\widehat{\beta}_{0WLS}|X) = \mathbb{V}(\widehat{\beta}_{0OLS}|X) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{SXX}\right)$$

We now turn to the standard error the residuals associated with the residuals of the model. Here we see that the standard error is given by the following form.

$$\sqrt{\frac{WRSS}{n-2}} = \sqrt{\frac{\sum_{i=1}^n w_i(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{2}\sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{2}\sqrt{\frac{RSS}{n-2}}$$

Thus we see that the standard error increases by introducing the weights. We calculated the associated $F$ statistic for the model. Recall that in $OLS$ analogue, our model is given by

$$\sqrt{w_i}Y_i = \beta_0\sqrt{w_i} + \beta_1\sqrt{w_i}x_i + \sqrt{w_i}e_i \Longleftrightarrow Y_{i,new} = \beta_0\sqrt{w_i} + \beta_1\sqrt{w_i}x_i + e_{i,new}$$

. Using this, along with our normal forms of $F$ statistics in regression we have

$$F_{WLS} = \frac{(n-2)SSreg}{RSS} = \frac{(n-2)\sum_{i=1}^n(\sqrt{w_i}\widehat{\beta}_{0WLS} + \sqrt{w_i}\widehat{\beta}_{1WLS} - \overline{y}_{new})^2}{\sum_{i=1}^n w_i(y_i - \widehat{\beta}_{0WLS} + \widehat{\beta}_{1WLS}x_i)^2}$$

$$= \frac{(n-2)\sum_{i=1}^n(\sqrt{w_i}\widehat{\beta}_{0WLS} + \sqrt{w_i}\widehat{\beta}_{1WLS} - \sqrt{w_i}\overline{y})^2}{\sum_{i=1}^n w_i(y_i - \widehat{\beta}_{0WLS} + \widehat{\beta}_{1WLS}x_i)^2}$$

$$= \frac{2(n-2)\sum_{i=1}^n(\widehat{\beta}_{0OLS} + \widehat{\beta}_{1OLS} - \overline{y})^2}{2\sum_{i=1}^n(y_i - \widehat{\beta}_{0OLS} + \widehat{\beta}_{1OLS}x_i)^2} = F_{OLS}$$

Therefore, we see that the for identical weights, that only the standard error changes (scaled by a function of $\sqrt{w}$).

4. (a) Recall for the regression $Y \sim X_2$, our estimated slope was given by

$$\widehat{\beta}_2 = \frac{\sum_{i=1}^n (x_{i2} - \overline{x_2})(y_i - \overline{y})}{\sum_{i=1}^n (x_{i2} - \overline{x_2})^2}$$

(b) The regression $X_1 \sim X_2$, our estimated slope is given by

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_{i2} - \overline{x_2})(x_{i1} - \overline{x_1})}{\sum_{i=1}^n (x_{i1} - \overline{x_1})^2} = 0$$

This is due to the fact that the sample correlation between $X_1$ and $X_2$ is zero.

(c) Here we look to show that the slope coefficient of $\hat{e}^{y|x_2} \sim \hat{e}^{y|x_1}$ is the same as the coefficient given by $Y \sim X_1$.

First note that the data in the problem is given by the following general forms

$$\hat{e}_i^{y|X_2} = y_i - \widehat{\beta}_0 - \widehat{\beta}_2 x_{i2} = y_i - \overline{y} + \widehat{\beta}_2(\overline{x_2} - x_{i2})$$

$$\hat{e}_i^{X_1|X_2} = x_{i1} - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i2} = x_{i1} - \overline{x}_1 + \widehat{\beta}_1(\overline{x_2} - x_{i2}) = x_{i1} - \overline{x}_1$$

Notice that

$$\overline{\hat{e}_i^{Y|X_2}} = \frac{1}{n}\sum_{i=1}^n (y_i - \overline{y}) + \frac{\widehat{\beta}_2}{n}\sum_{i=1}^n (\overline{x}_2 - x_{i2}) = 0$$

$$\overline{\hat{e}_i^{X_1|X_2}} = \frac{1}{n}\sum_{i=1}^n (x_i 1 - \overline{x}_1) = 0$$

Hence for our regression, we have the following estimated slope

$$\widehat{\beta}_* = \frac{\sum_{i=1}^n (\hat{e}_i^{Y|X_2} - \overline{\hat{e}^{Y|X_2}})(\hat{e}_i^{X_1|X_2} - \overline{\hat{e}^{X_1|X_2}})}{\sum_{i=1}^n (\hat{e}_i^{X_1|X_2} - \overline{e^{X_1|X_2}})^2}$$

$$= \frac{\sum_{i=1}^n (\hat{e}_i^{Y|X_2} \hat{e}_i^{X_1|X_2})}{\sum_{i=1}^n (\hat{e}_i^{X_1|X_2})^2}$$

$$= \frac{\sum_{i=1}^n (y_i - \overline{y} + \widehat{\beta}_2(\overline{x_2} - x_{i2}))(x_{i1} - \overline{x}_1)}{\sum_{i=1}^n (x_{i1} - \overline{x}_1)^2}$$

$$= \frac{\sum_{i=1}^n (y_i - \overline{y})(x_{i1} - \overline{x}_1)}{\sum_{i=1}^n (x_{i1} - \overline{x}_1)^2} + \frac{\widehat{\beta}_2 \sum_{i=1}^n (\overline{x_2} - x_{i2})(x_{i1} - \overline{x}_1)}{\sum_{i=1}^n (x_{i1} - \overline{x}_1)^2}$$

$$= \frac{\sum_{i=1}^n (y_i - \overline{y})(x_{i1} - \overline{x}_1)}{\sum_{i=1}^n (x_{i1} - \overline{x}_1)^2}$$

The last equality is again due to the fact that $x_1, x_2$ have zero sample correlation. You'll notice that $\widehat{\beta}_*$ is the exact form given by the slope coefficient given by the regression $Y \sim X_1$.