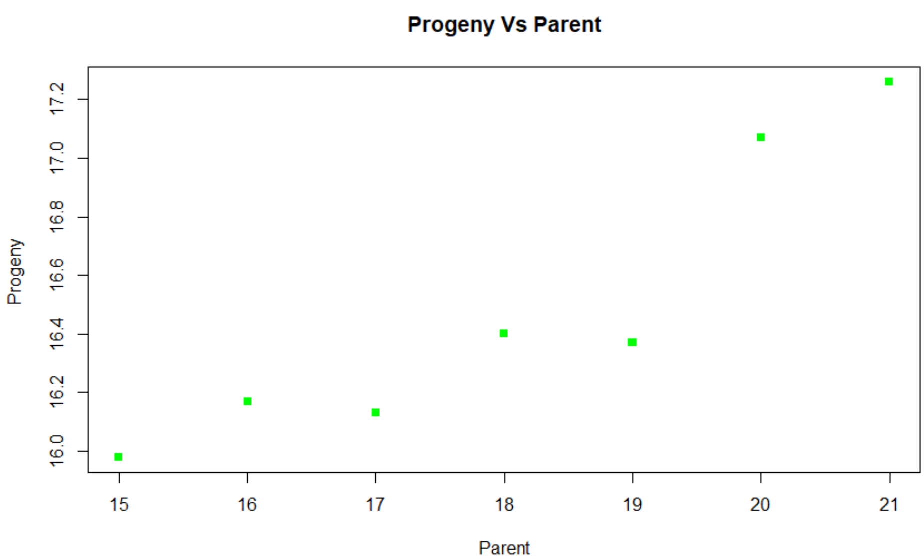
STAT 8030 HW 4

1.1 (7.7.1):

Scatter Plot is shown below:



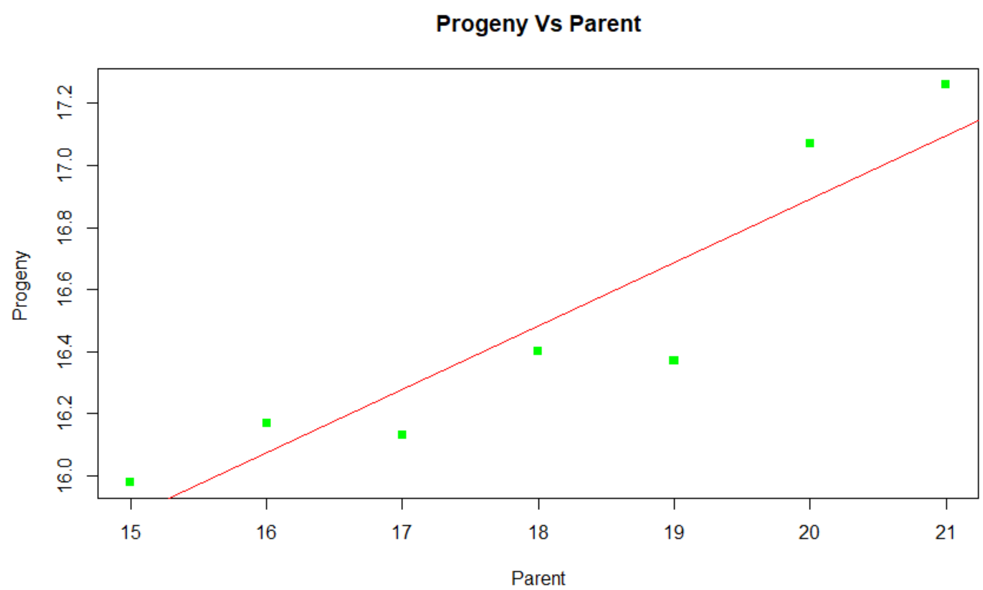
As we see above, there are very few data points. Hence, we cannot have any assumption and conclusion for a pattern.

1.2 (7.7.2):

We can see below the summary for computing weighted regression:

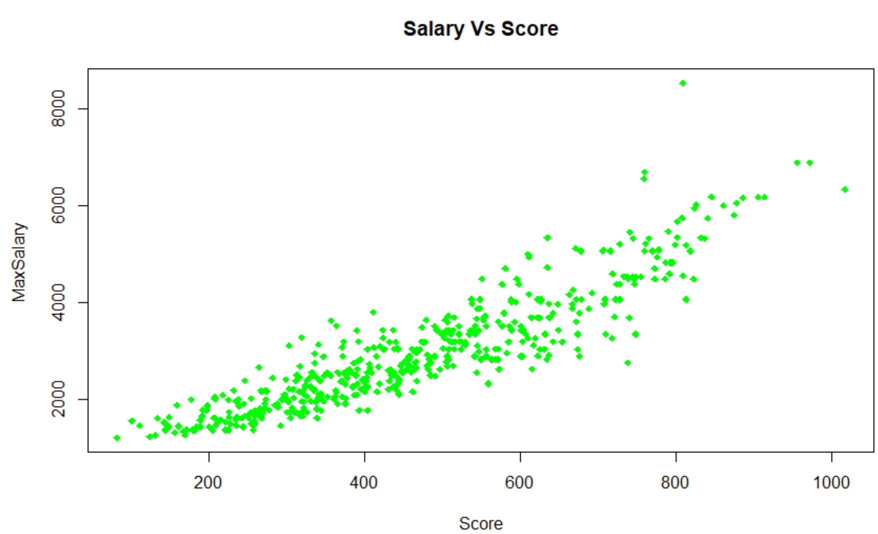
```
> summary(Elmweighted)

Call:
lm(formula = Progeny ~ Parent, data = galtonpeas, weights = 1/SD^2)

Weighted Residuals:
       1        2        3        4        5        6        7 
 0.08187  0.09162 -0.16753 -0.04067 -0.08950  0.06071  0.06328 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.79642    0.68112  18.787 7.87e-06 ***
Parent       0.20480    0.03815   5.368  0.00302 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.11 on 5 degrees of freedom
Multiple R-squared:  0.8521,	Adjusted R-squared:  0.8225 
F-statistic: 28.81 on 1 and 5 DF,  p-value: 0.003021
```

And below is the scatter plot for fitting regression:



2.1

Scatter Plot for Maxsalary vs Score is shown below:



As we see above, the pattern does not look to be linear since the datapoints seems to be making a curve, though we can see an uphill pattern in the datapoints.

2.2

Summary and scatter plot for Linear Model is shown below:

```
> summary(N_lm)

Call:
lm(formula = MaxSalary ~ Score, data = salarygov)

Residuals:
    Min      1Q  Median      3Q     Max 
-1797.9  -284.1   -42.0   248.7  3569.2 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  295.274     62.012   4.762 2.53e-06 ***
Score          5.760      0.123  46.844  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 507.2 on 493 degrees of freedom
Multiple R-squared:  0.8165,	Adjusted R-squared:  0.8162 
F-statistic:  2194 on 1 and 493 DF,  p-value: < 2.2e-16
```

Summary and scatter plot (in black) for Model with degree 2 is shown below:

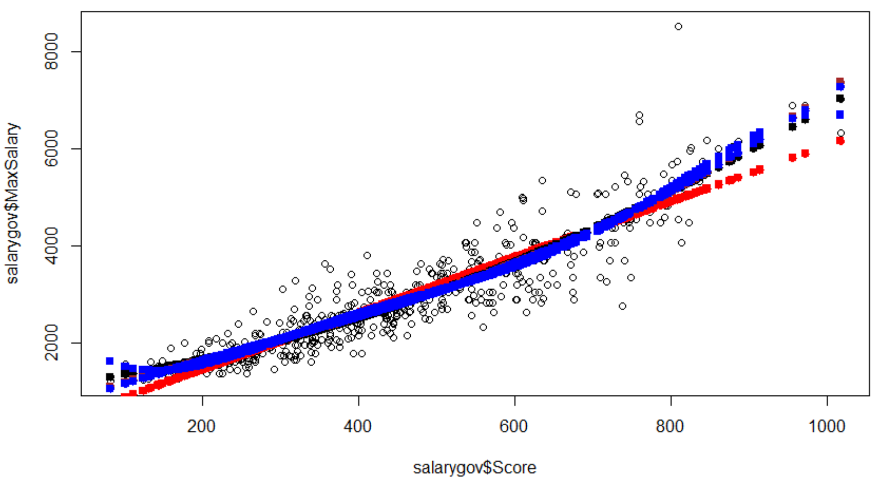
```
> summary(lm_2)

Call:
lm(formula = MaxSalary ~ Score + I(Score^2), data = salarygov)

Residuals:
    Min      1Q  Median      3Q     Max 
-1877.0  -251.8   -67.2   251.2  3344.7 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.102e+03  1.320e+02   8.345 7.23e-16 ***
Score       2.007e+00  5.613e-01   3.575 0.000384 ***
I(Score^2)  3.750e-03  5.484e-04   6.838 2.39e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 485.1 on 492 degrees of freedom
Multiple R-squared:  0.8325,	Adjusted R-squared:  0.8318 
F-statistic:  1222 on 2 and 492 DF,  p-value: < 2.2e-16
```

Summary and scatter plot (in brown) for Model with degree 3 is shown below:

```
Call:
lm(formula = MaxSalary ~ Score + I(Score^2) + I(Score^3), data = salarygov)

Residuals:
    Min      1Q  Median      3Q     Max 
-1842.8  -257.8   -45.7   245.0  3343.5 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  7.077e+02  2.560e+02   2.764  0.00593 **
Score        4.956e+00  1.736e+00   2.855  0.00449 **
I(Score^2)  -2.640e-03  3.602e-03  -0.733  0.46398   
I(Score^3)   4.139e-06  2.306e-06   1.795  0.07334 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 484 on 491 degrees of freedom
Multiple R-squared:  0.8336,	Adjusted R-squared:  0.8325 
F-statistic: 819.7 on 3 and 491 DF,  p-value: < 2.2e-16
```

Summary and scatter plot (in blue) for Model with degree 5 is shown below:

```
> summary(lm_5)

Call:
lm(formula = MaxSalary ~ Score + I(Score^2) + I(Score^3) + I(Score^4) +
    I(Score^5), data = salarygov)

Residuals:
    Min      1Q  Median      3Q     Max
-1831.3  -274.9   -56.1   236.6  3240.5

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.896e+03  8.195e+02   3.534 0.000449 ***
Score       -2.562e+01  1.040e+01  -2.464 0.014080 *
I(Score^2)   1.470e-01  4.812e-02   3.056 0.002366 **
I(Score^3)  -3.245e-04  1.022e-04  -3.174 0.001597 **
I(Score^4)   3.300e-07  1.008e-07   3.273 0.001141 **
I(Score^5)  -1.231e-10  3.732e-11  -3.297 0.001047 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 479.7 on 489 degrees of freedom
Multiple R-squared:  0.8372,    Adjusted R-squared:  0.8355
F-statistic: 502.9 on 5 and 489 DF,  p-value: < 2.2e-16
```



2.3

F-Tests:

- For Fitted model with d=1 and d=2

  Null Hypothesis => Ho: β2 = 0
  Alternate Hypothesis => Ha: β2!=0

```
> anova(N_lm,lm_2)
Analysis of Variance Table

Model 1: MaxSalary ~ Score
Model 2: MaxSalary ~ Score + I(Score^2)
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1    493 126801821
2    492 115797671  1  11004150 46.754 2.393e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

We can see F Statistic: F-Value= 46.754 and P-Value = 2.393e-11

Now, considering alpha = 0.05, we see that p-value is less than alpha. Hence, we reject the null hypothesis. Thus, adding the quadratic term in model can improves its accuracy.

- For Fitted model with d=2 and d=3

  Null Hypothesis => Ho: β3 = 0
  Alternate Hypothesis => Ha: β3!=0

```
> anova(lm_2,lm_3)
Analysis of Variance Table

Model 1: MaxSalary ~ Score + I(Score^2)
Model 2: MaxSalary ~ Score + I(Score^2) + I(Score^3)
  Res.Df       RSS Df Sum of Sq      F  Pr(>F)
1    492 115797671
2    491 115043095  1    754576 3.2205 0.07334 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

We can see F Statistic: F-Value= 3.2205 and P-Value = 0.07334

Now, considering alpha = 0.05, we see that p-value is more than alpha. Hence, we fail to reject the null hypothesis. Thus, adding the quadratic term in model will not improve its accuracy.

- For Fitted model with d=1 and d=5

  Null Hypothesis => Ho: β5 = 0
  Alternate Hypothesis => Ha: β5!=0

```
> anova(lm_2,lm_5)
Analysis of Variance Table

Model 1: MaxSalary ~ Score + I(Score^2)
Model 2: MaxSalary ~ Score + I(Score^2) + I(Score^3) + I(Score^4) + I(Score^5)
  Res.Df       RSS Df Sum of Sq      F   Pr(>F)
1    492 115797671
2    489 112536963  3   3260708 4.7229 0.002931 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

We can see F Statistic: F-Value= 4.7229 and P-Value = 0.002931

Now, considering alpha = 0.05, we see that p-value is less than alpha. Hence, we reject the null hypothesis. Thus, adding the quadratic term in model can improves its accuracy.

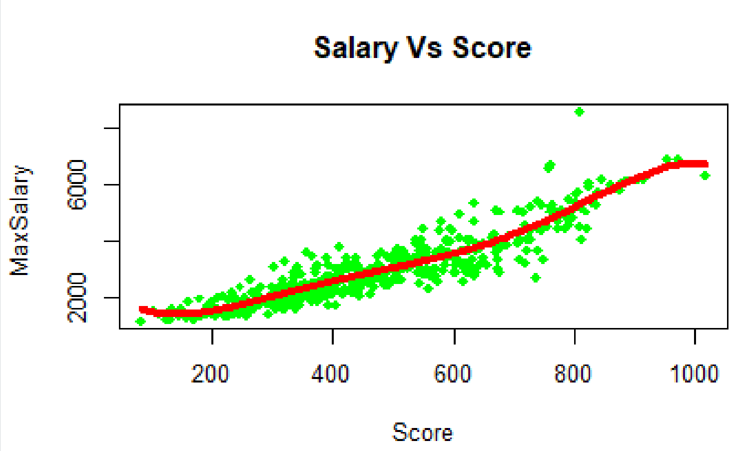
R-squared values for the models:

```
> summary(N_lm)$r.squared
[1] 0.8165477
> summary(lm_2)$r.squared
[1] 0.8324681
> summary(lm_3)$r.squared
[1] 0.8335598
> summary(lm_5)$r.squared
[1] 0.8371856
>
```

As we see above, R-squared value for model d=5 is changed the max. Thus, confirming that the model with d=5 is the most accurate model and the coefficients for the same (d=5) is shown below:

```
> lm_5

Call:
lm(formula = MaxSalary ~ Score + I(Score^2) + I(Score^3) + I(Score^4) +
    I(Score^5), data = salarygov)

Coefficients:
(Intercept)        Score   I(Score^2)   I(Score^3)   I(Score^4)   I(Score^5)
  2.896e+03   -2.562e+01    1.470e-01   -3.245e-04    3.300e-07   -1.231e-10
```

Scatter Plot with fitted mean function is shown below:



2.4

We can use the NE column to find the weights for the total number of employees currently employed in specific job class.

```
> AW = salarygov$NE
```

2.5

Coefficients for weighted model is shown below:

```
> #2.5
>
> MW <- lm(MaxSalary~Score, data = salarygov, weights = AW)
> summary(MW)

Call:
lm(formula = MaxSalary ~ Score, data = salarygov, weights = AW)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-6549.1  -558.7   -83.7   497.8 10445.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 368.3847    43.9524   8.381  5.5e-16 ***
Score         5.5961     0.1125  49.722  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1282 on 493 degrees of freedom
Multiple R-squared:  0.8337,    Adjusted R-squared:  0.8334
F-statistic:  2472 on 1 and 493 DF,  p-value: < 2.2e-16
```

Coefficients for unweighted model is shown below:

```
> summary(N_lm)

Call:
lm(formula = MaxSalary ~ Score, data = salarygov)

Residuals:
    Min      1Q  Median      3Q     Max
-1797.9  -284.1   -42.0   248.7  3569.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  295.274     62.012   4.762 2.53e-06 ***
Score          5.760      0.123  46.844  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 507.2 on 493 degrees of freedom
Multiple R-squared:  0.8165,    Adjusted R-squared:  0.8162
F-statistic:  2194 on 1 and 493 DF,  p-value: < 2.2e-16
```

Seeing above, we can say that the increase in maximum salary is 5.5961$ for weighted model and is 5.76$ for the unweighted model.