Assignment 5

108048110

2022-11-27

```
## Warning: package 'GGally' was built under R version 4.2.1
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

Q1. Predicting the height of the son from the height of the father.

```
##     FatherH         SonmeanH      NumofFather
##  Min.   :62.00   Min.   :65.50   Min.   : 2.0
##  1st Qu.:64.75   1st Qu.:66.80   1st Qu.: 7.5
##  Median :67.50   Median :68.20   Median :17.0
##  Mean   :67.50   Mean   :68.42   Mean   :16.0
##  3rd Qu.:70.25   3rd Qu.:69.70   3rd Qu.:26.0
##  Max.   :73.00   Max.   :72.00   Max.   :27.0
```

- The father's heights were rounded to the nearest inch.
- The average height of the son for fathers of that height is given.
- The number of fathers in each category is given.

***i.* Construct a linear regression model for predicting the height of the son from the height of the father in the best manner given information available.**

- There are several observations given a single value of x, therefore, I will use the generalized least square to compute the weights of each predictor to estimate the parameters for the rest of the questions.
- Since the observed responses $y_i's$ are actually averages of several observations `number of fathers`, we set the wight as the number of observations.

Hard Coding Ver

- Constructing Σ and Σ^{-1} first.

```
##       [,1]      [,2]       [,3]       [,4]       [,5]       [,6]       [,7]
##  [1,]  0.5 0.0000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##  [2,]  0.0 0.1666667 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##  [3,]  0.0 0.0000000 0.08333333 0.00000000 0.00000000 0.00000000 0.00000000
##  [4,]  0.0 0.0000000 0.00000000 0.05263158 0.00000000 0.00000000 0.00000000
##  [5,]  0.0 0.0000000 0.00000000 0.00000000 0.03703704 0.00000000 0.00000000
##  [6,]  0.0 0.0000000 0.00000000 0.00000000 0.00000000 0.03846154 0.00000000
##  [7,]  0.0 0.0000000 0.00000000 0.00000000 0.00000000 0.00000000 0.03846154
##  [8,]  0.0 0.0000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##  [9,]  0.0 0.0000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
## [10,]  0.0 0.0000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
## [11,]  0.0 0.0000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
## [12,]  0.0 0.0000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##            [,8] [,9]      [,10] [,11] [,12]
##  [1,] 0.00000000 0.00 0.00000000 0.000   0.0
##  [2,] 0.00000000 0.00 0.00000000 0.000   0.0
##  [3,] 0.00000000 0.00 0.00000000 0.000   0.0
##  [4,] 0.00000000 0.00 0.00000000 0.000   0.0
##  [5,] 0.00000000 0.00 0.00000000 0.000   0.0
##  [6,] 0.00000000 0.00 0.00000000 0.000   0.0
##  [7,] 0.00000000 0.00 0.00000000 0.000   0.0
##  [8,] 0.03846154 0.00 0.00000000 0.000   0.0
##  [9,] 0.00000000 0.05 0.00000000 0.000   0.0
## [10,] 0.00000000 0.00 0.06666667 0.000   0.0
## [11,] 0.00000000 0.00 0.00000000 0.125   0.0
## [12,] 0.00000000 0.00 0.00000000 0.000   0.2
```

- Since $w_i's$ is proportion to $1/var(\epsilon)$

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
##  [1,]    2    0    0    0    0    0    0    0    0     0     0     0
##  [2,]    0    6    0    0    0    0    0    0    0     0     0     0
##  [3,]    0    0   12    0    0    0    0    0    0     0     0     0
##  [4,]    0    0    0   19    0    0    0    0    0     0     0     0
##  [5,]    0    0    0    0   27    0    0    0    0     0     0     0
##  [6,]    0    0    0    0    0   26    0    0    0     0     0     0
##  [7,]    0    0    0    0    0    0   26    0    0     0     0     0
##  [8,]    0    0    0    0    0    0    0   26    0     0     0     0
##  [9,]    0    0    0    0    0    0    0    0   20     0     0     0
## [10,]    0    0    0    0    0    0    0    0    0    15     0     0
## [11,]    0    0    0    0    0    0    0    0    0     0     8     0
## [12,]    0    0    0    0    0    0    0    0    0     0     0     5
```

- $S = diag(1/\sqrt{w}_1, 1/\sqrt{w}_2, ...)$, then $\Sigma = SS^T$.
- Then we can use OLS to regress $S^{-1}Y$ on $S^{-1}X$.

```
g = lm(sy~sx-1)
summary(g, cor=T)
```

```
##
## Call:
## lm(formula = sy ~ sx - 1)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.39024 -0.77499  0.04766  1.15672  1.67501
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## sx         32.5820     2.2486   14.49 4.87e-08 ***
## sxFatherH   0.5297     0.0332   15.96 1.93e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.147 on 10 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.416e+05 on 2 and 10 DF,  p-value: < 2.2e-16
##
## Correlation of Coefficients:
##           sx
## sxFatherH -1.00
```

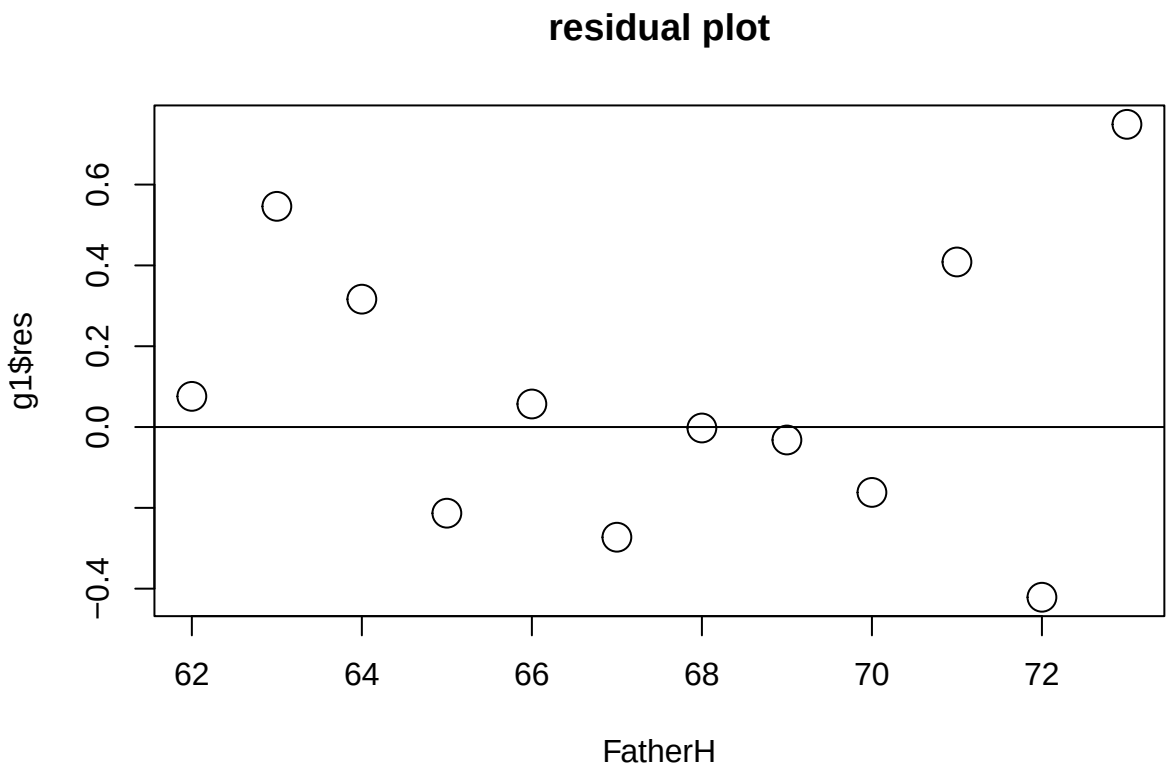
Wighted Least Square ver

- $w_i = n_i$, $S^{-1} = \sqrt(w_i)$

```
Sigi = sqrt(NumofFather)
g1 = lm(SonmeanH~FatherH, weights = NumofFather)
summary(g1, cor=T)
```

```
##
## Call:
## lm(formula = SonmeanH ~ FatherH, weights = NumofFather)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -1.39024 -0.77499  0.04766  1.15672  1.67501
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.5820     2.2486   14.49 4.87e-08 ***
## FatherH       0.5297     0.0332   15.96 1.93e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.147 on 10 degrees of freedom
## Multiple R-squared:  0.9622, Adjusted R-squared:  0.9584
## F-statistic: 254.6 on 1 and 10 DF,  p-value: 1.926e-08
##
## Correlation of Coefficients:
##         (Intercept)
## FatherH -1.00
```

- Though the result seems indifference; however, the latter method included intercept in the model, therefore, its calculated R^2 value will be more vigorous compared to the former hard coding approach.



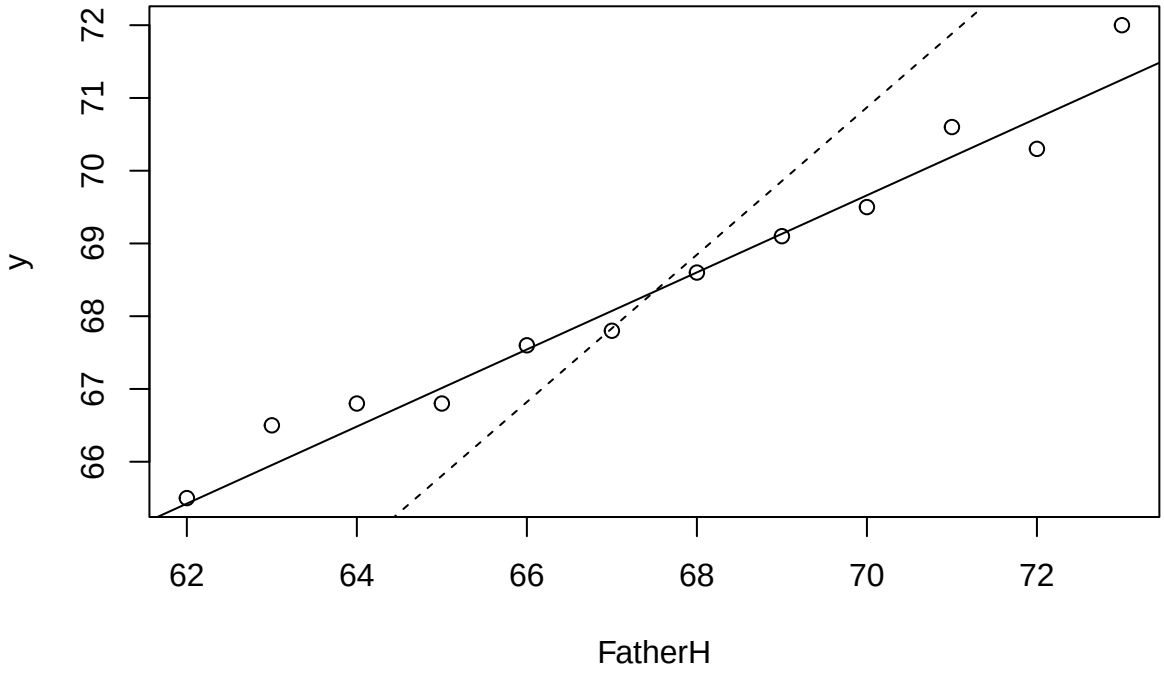
ii. **Can the model be simplified to** $Son\ height = \ father\ height + \epsilon$**?**

- $H_0 : yes\ ;\ H_1 : no$
- Try fitting the regression without weights first.

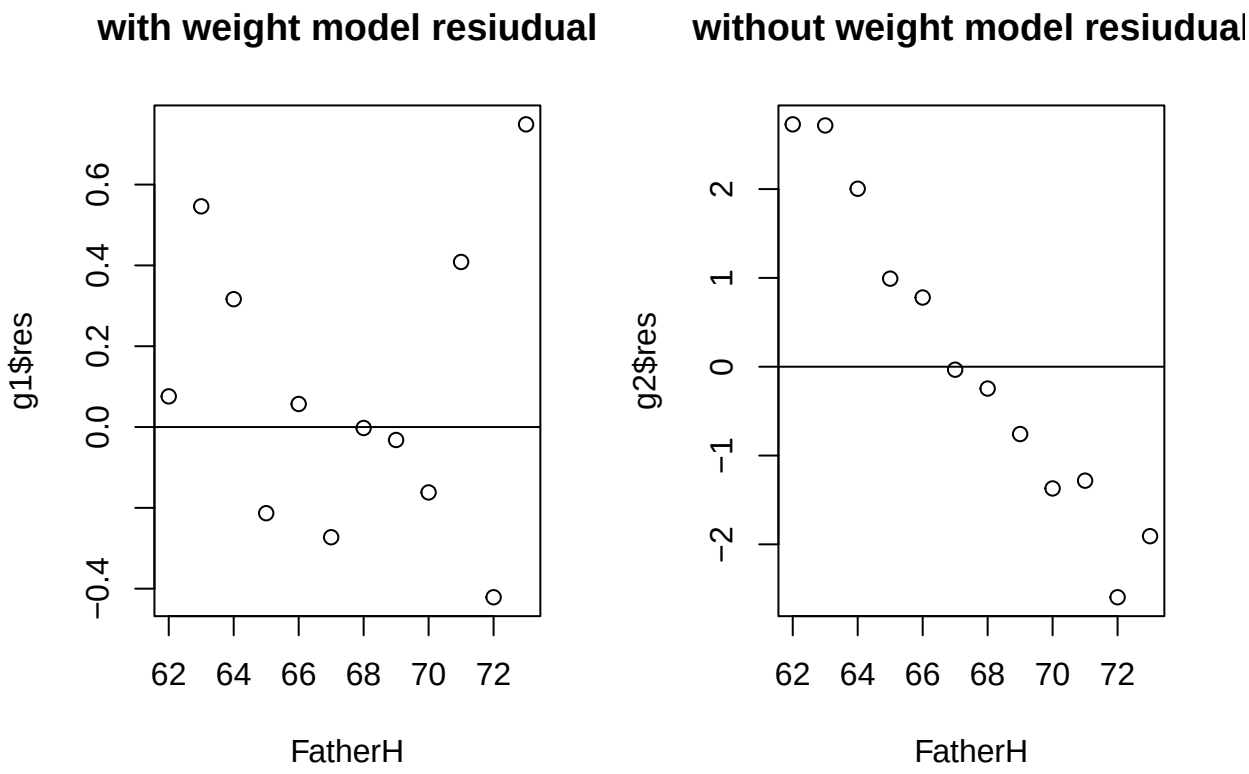
```
g2 = lm(SonmeanH~FatherH-1)
summary(g2)
```

```
##
## Call:
## lm(formula = SonmeanH ~ FatherH - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5955 -1.3050 -0.1395  1.2447  2.7289
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## FatherH 1.012438   0.007615   132.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.783 on 11 degrees of freedom
## Multiple R-squared:  0.9994, Adjusted R-squared:  0.9993
## F-statistic: 1.767e+04 on 1 and 11 DF,  p-value: < 2.2e-16
```

- Comparing two fits.
  - Solid line: fitted line of model with the calculated weights.
  - Dash line: the simplified model with the height of the father to be the only predictor.



- As we can observe from the plot that the solid line seemed to be slightly closer to the response and fit better than the dash line.



- Comparing their residual variance to test if the given model good enough for the prediction.
- the F-statistics with the null hypothesis to be "there isn't much different between two models", and the alternative hypothesis to be "there is a significance difference between two models."

```
## [1] 31.79089
```

```
## [1] 0.0004338593
```

- The p-value larger than the significance level indicates that we reject the null hypothesis, that is, the model can not be simplified.

Q2. Ultrasonic measurements of the depths of defects in the Alaska pipeline in the field and in the lab.

- Take a look at the data

```
##      Field            Lab            Batch
##  Min.   : 5.00   Min.   : 4.30   Min.   :1.000
##  1st Qu.:18.00   1st Qu.:18.35   1st Qu.:2.000
##  Median :35.00   Median :38.00   Median :3.000
##  Mean   :33.58   Mean   :39.10   Mean   :3.234
##  3rd Qu.:46.50   3rd Qu.:55.55   3rd Qu.:5.000
##  Max.   :85.00   Max.   :81.90   Max.   :6.000
```

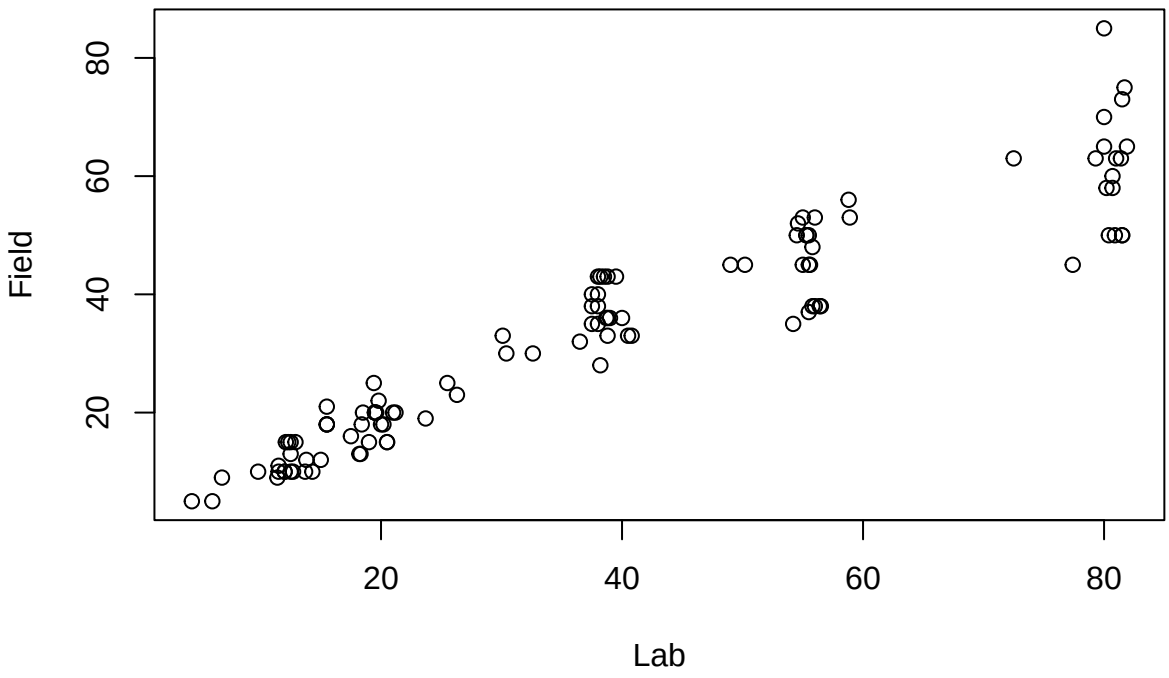
Batch	Freq
1	19
2	20
3	20
4	20
5	21
6	7

- Batch effect is not significant and can be ignored.
- Lab measurements are more accurate than that of measured in the field.

***i.* Fit a regression model Lab~Field. And check for non-constant variance.**

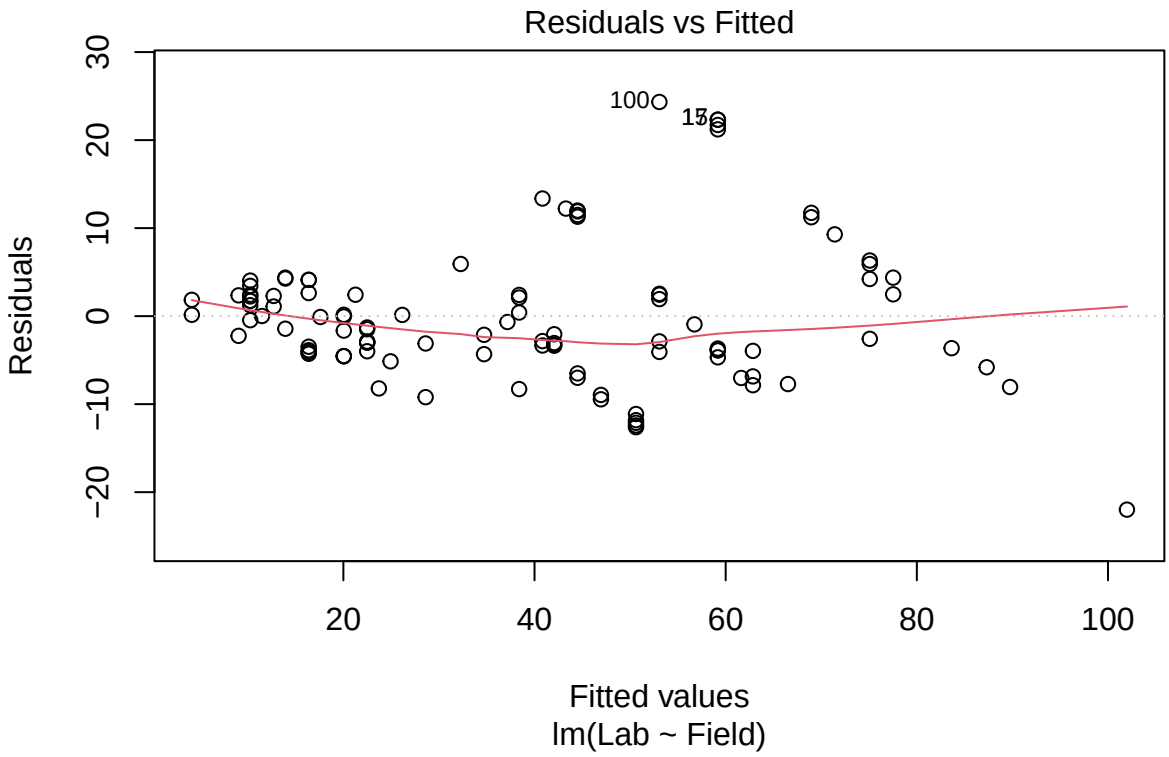
```
g = lm(Lab~Field)
summary(g)
```

```
##
## Call:
## lm(formula = Lab ~ Field)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.985  -4.072  -1.431   2.504  24.334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.96750    1.57479  -1.249    0.214
## Field        1.22297    0.04107  29.778   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.865 on 105 degrees of freedom
## Multiple R-squared:  0.8941, Adjusted R-squared:  0.8931
## F-statistic: 886.7 on 1 and 105 DF,  p-value: < 2.2e-16
```



- It is obvious in the plot that the variance is not constant.
- Also, there are replicate measurements in this dataset.

	Field	Lab	Batch
1	18	20.2	1
5	18	15.5	1
21	18	20.0	2
25	18	15.5	2
43	18	18.4	3
73	18	15.5	4



- We can see that the variability around 0 increases as we move further to the right with bigger fitted values.
- Hence, we can confidently conclude that the data does not have a constant variance.

***ii.* Use weights to account for the non-constant variance.**

Suppose we assume that the variance in the response is linked to the predictor $var(Lab) = \alpha_0 * Field^{\alpha_1}$.

Regress $log(var(Lab))$ on $log(mean(Field))$ to estimate α_0, α_1.

Use this to determine weights in a WLS fit of `Lab` on `Field`. Show the regression summary.

- Computation splitting the range of `Field` into 12 groups of size 9 except for the last group that has only 8 values.
- $log(var(Lab)) = log(\alpha_0) + \alpha_1 log(Field)$

```
g4 = lm(log(varlab)~log(meanfield))
summary(g4)
```

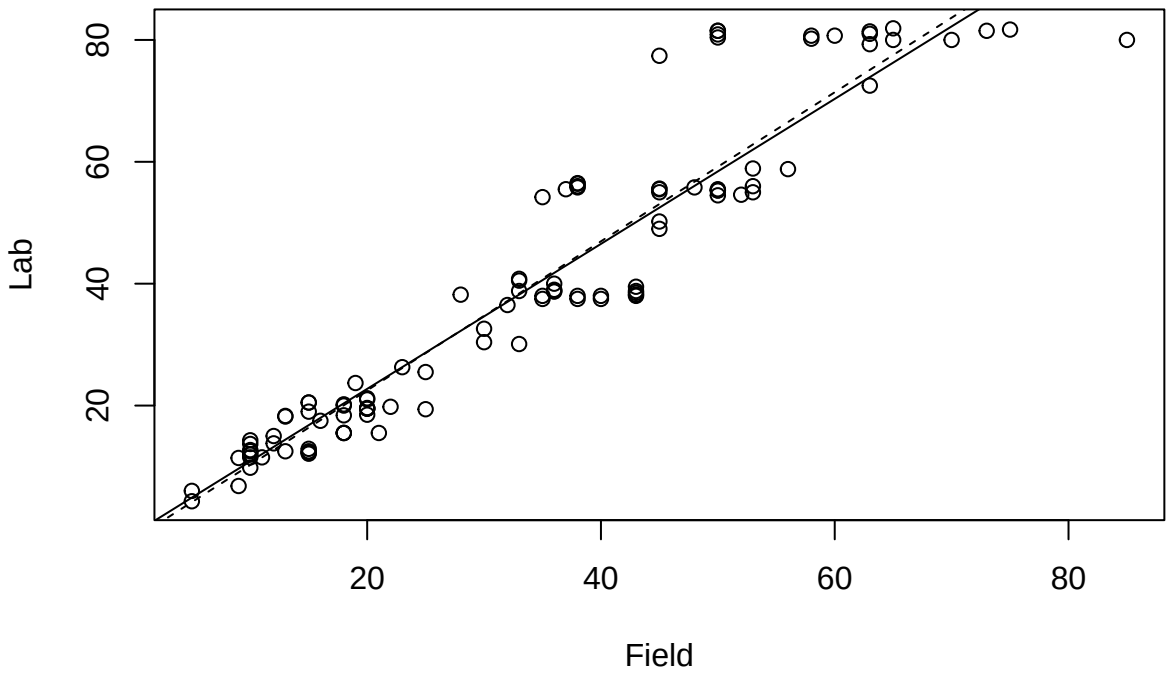
```
##
## Call:
```

```
## lm(formula = log(varlab) ~ log(meanfield))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00477 -0.42268  0.05989  0.37854  0.93815
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.9352     1.0929  -1.771 0.110403
## log(meanfield)   1.6707     0.3296   5.070 0.000672 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.657 on 9 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7118
## F-statistic:  25.7 on 1 and 9 DF,  p-value: 0.0006723
```

- $\alpha_0 = e^{-1.9352} = 0.1444,\ \alpha_1 = 1.6707$
- $var(Lab) = 0.1444 * Field^{1.6707}$
- $\sigma = \sqrt{(0.1444)} = 0.38,\ w_i = \frac{1}{Field^{1.6707}}$

```
w =  1/Field^(1.6707)
g1 = lm(Lab~Field, weights = w)
summary(g1)
```

```
##
## Call:
## lm(formula = Lab ~ Field, weights = w)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66245 -0.25532 -0.09474  0.22675  1.03651
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.05531    0.69766  -1.513    0.133
## Field        1.18963    0.03401  34.984   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3742 on 105 degrees of freedom
## Multiple R-squared:  0.921,  Adjusted R-squared:  0.9202
## F-statistic:  1224 on 1 and 105 DF,  p-value: < 2.2e-16
```



Q3. This dataset provides the data on the outside diameter of crankpins produced by an industrial process.

- All the crankpins should be between 0.7425 and 0.7430 inches.
- The number given in the table are in units of 0.00001 inches deviation from 0.742 inches.

```
summary(data2)
```

```
##       day           diameter
##  Min.   : 1.00   Min.   : 72.00
##  1st Qu.: 6.25   1st Qu.: 85.75
##  Median :11.50   Median : 89.00
##  Mean   :11.50   Mean   : 88.00
##  3rd Qu.:16.75   3rd Qu.: 92.25
##  Max.   :22.00   Max.   :100.00
```

- So the actual outside diameter of crankpins ranges from 0.74272 to 0.743 inches.
- Under control, the average size of the crankpin produced should...
    1. fall near the middle of the specified range
    2. not depend on the time.
- Fit an appropriate model to see of the process is under control and test for lack of fit in the model.

Under control test

- $H_0 : g = process\ is\ under\ control,\ H_1 : g1 = not\ under\ control$
- This is a dataset with replicate values, which show a pattern of unequal variance.
- Calculate the sample variance.
- The data is obviously with inconsistent variation.
- Hence, we calculated weights as 1/sample variance.

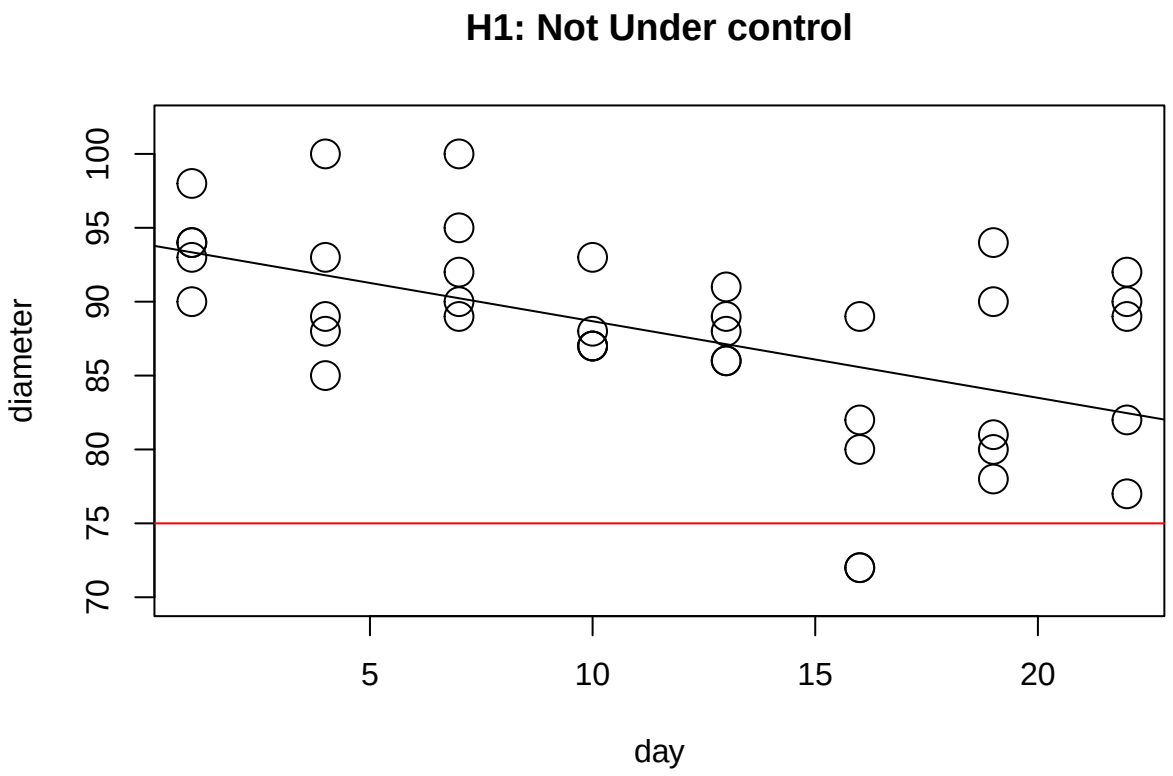
	1.000000	4.000000	7.0000000	10.0000000	13.0000000	16.0000000	19.0000000	22.0000000
dayvar	8.200000	33.500000	19.7000000	6.8000000	4.5000000	52.0000000	48.8000000	39.5000000
daymean	93.800000	91.000000	93.2000000	88.4000000	88.0000000	79.0000000	84.6000000	86.0000000
w	0.010661	0.010989	0.0107296	0.0113122	0.0113636	0.0126582	0.0118203	0.0116279

- Calculated Σ
- Cholesky Decomposition and fit the model.

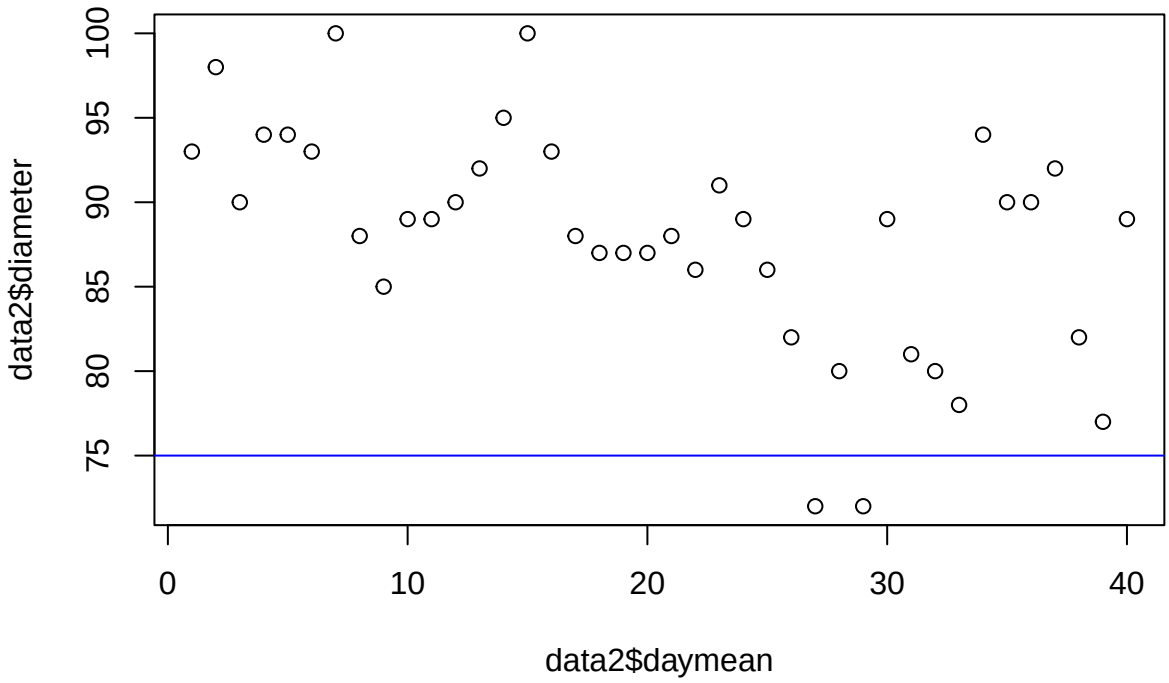
```
S = chol(Sig)
Si = solve(S)

g1 = lm(daymean~unique(day), weights=w)
summary(g1, cor=T)
```

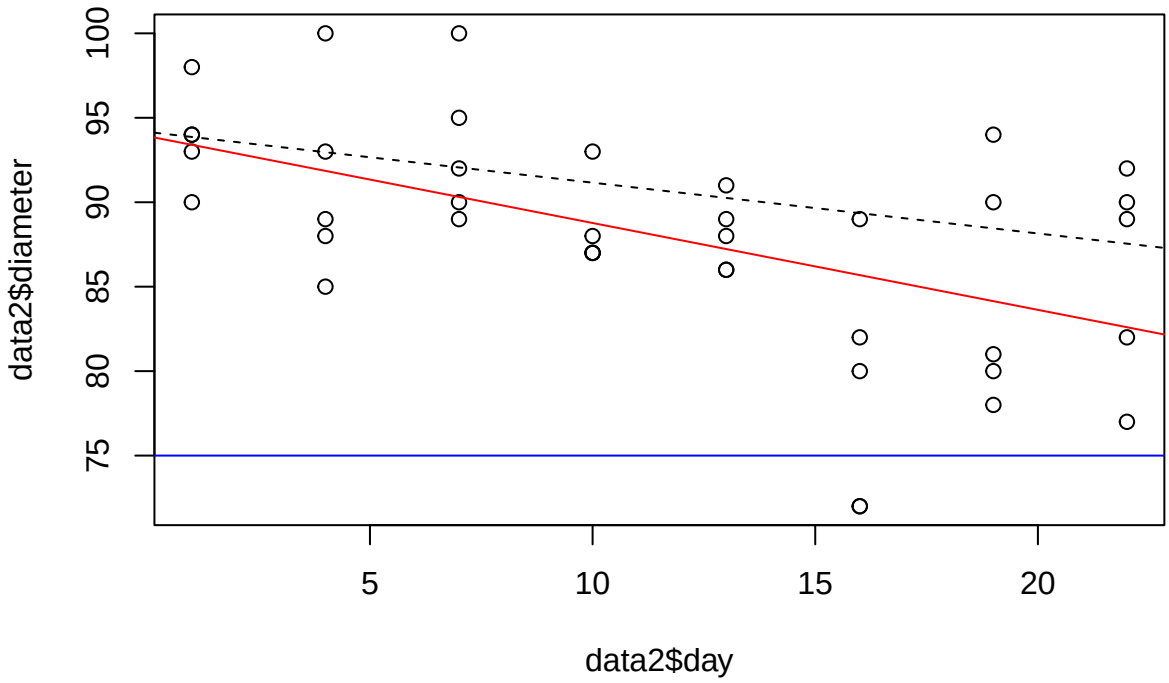
```
##
## Call:
## lm(formula = daymean ~ unique(day), weights = w)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73884 -0.04261  0.05570  0.14712  0.38196
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  93.8581     2.4230  38.736 1.98e-08 ***
## unique(day)  -0.5182     0.1780  -2.912   0.0269 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3672 on 6 degrees of freedom
## Multiple R-squared:  0.5856, Adjusted R-squared:  0.5165
## F-statistic: 8.479 on 1 and 6 DF,  p-value: 0.02692
##
## Correlation of Coefficients:
##             (Intercept)
## unique(day) -0.86
```



- If the process is under control, the average size of the production diameter should fall within the range of 75 (50+100/2), and the result should not be affected by predictor, day.




```
##
## Call:
## lm(formula = diameter ~ data2$daymean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0451  -3.3872  -0.6917   3.7782  10.3459
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   94.16538    1.81233  51.958  < 2e-16 ***
## data2$daymean -0.30075    0.07703  -3.904 0.000375 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.624 on 38 degrees of freedom
## Multiple R-squared:  0.2863, Adjusted R-squared:  0.2675
## F-statistic: 15.24 on 1 and 38 DF,  p-value: 0.0003749
```



- red line: not under control
- dash line: under control

```
## Analysis of Variance Table
##
## Model 1: diameter ~ day
## Model 2: diameter ~ data2$daymean
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1     38 1184.1
## 2     38 1201.9  0   -17.783
```

Test for lack of fit

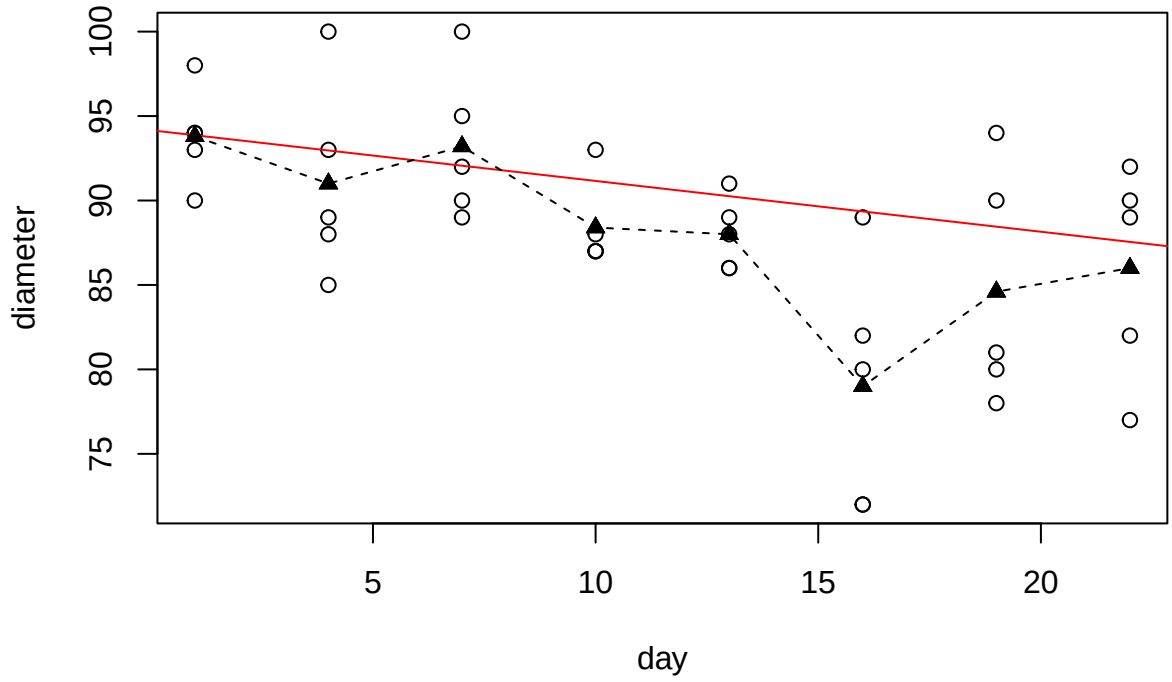
- Saturated model

```
g2 = lm(diameter~factor(day))
summary(g2)
```

```
##
## Call:
## lm(formula = diameter ~ factor(day))
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
##   -9.0   -3.3   -0.6    3.0   10.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     93.800      2.308  40.648  < 2e-16 ***
## factor(day)4    -2.800      3.263  -0.858  0.39728
## factor(day)7    -0.600      3.263  -0.184  0.85529
## factor(day)10   -5.400      3.263  -1.655  0.10776
## factor(day)13   -5.800      3.263  -1.777  0.08503 .
## factor(day)16  -14.800      3.263  -4.535 7.63e-05 ***
## factor(day)19   -9.200      3.263  -2.819  0.00819 **
## factor(day)22   -7.800      3.263  -2.390  0.02290 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.16 on 32 degrees of freedom
## Multiple R-squared:  0.4941, Adjusted R-squared:  0.3834
## F-statistic: 4.464 on 7 and 32 DF,  p-value: 0.001467
```

- The saturated model has 8 parameters
- The residual standard error is 5.16, which is the pure error estimate of true σ.




```
## Analysis of Variance Table
```

```
##
## Model 1: diameter ~ data2$daymean
## Model 2: diameter ~ factor(day)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     38 1201.9
## 2     32  852.0  6     349.9 2.1903 0.06987 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The p-value is above 0.05, we do not reject the null hypothesis that these is no lack of fit.

```
detach(data2)
```