ST 518 Homework 4

Eric Warren

September 22, 2023

Contents

1 Problem 1

First we are going to read in the `battery` data and will show what it looks like.

```
library(tidyverse)
battery <- read_table("battery.txt")
battery$TypeBat <- factor(battery$TypeBat)
battery
```

```
## # A tibble: 16 x 5
##    TypeBat  Life UCost  LPUC Order
##    <fct>   <dbl> <dbl> <dbl> <dbl>
##  1 1         602 0.985   611     1
##  2 2         863 0.935   923     2
##  3 1         529 0.985   537     3
##  4 4         235 0.495   476     4
##  5 1         534 0.985   542     5
##  6 1         585 0.985   593     6
##  7 2         743 0.935   794     7
##  8 3         232 0.52    445     8
##  9 4         282 0.495   569     9
## 10 2         773 0.935   827    10
## 11 2         840 0.935   898    11
## 12 3         255 0.52    490    12
## 13 4         238 0.495   480    13
## 14 3         200 0.52    384    14
## 15 4         228 0.495   460    15
## 16 3         215 0.52    413    16
```

1.1 Part A

Write out the factorial effects model for the LPUC variable (lifetime per unit cost) that allows for the mean to vary across battery types. Include a random variable for an error term that has the same variance for all four battery types.

We know when the variances should be the same for all treatments groups that our model for One Factor Experiments should be $Y_{ij} = \mu + \tau_i + E_{ij}$ for $i = 1, 2, ..., t$ and $j = 1, 2, ..., n$ where t is the number of treatments and n is the sample size in each treatment. Also E_{ij} are i.i.d $N(0, \sigma^2)$ errors. In this case, $t = 4$ and $n = 4$ given that there are 4 battery types and 4 batteries in each type. So knowing all of this our model should be written as $Y_{ij} = \mu + \tau_i + E_{ij}$ for $i = 1, 2, 3, 4$ and $j = 1, 2, 3, 4$ and E_{ij} are i.i.d $N(0, \sigma^2)$ errors.

1.2 Part B

Report the mean and variance for the 4 measurements from each of the four battery types.

We can do this by using the `group_by()` function in R to group all the data by battery type and then use the `summarize()` function to find the `mean` and `variance` of `LPUC` value by each battery type. I am assuming sample variance and sample mean in the question so our functions of `mean()` get the sample mean and `var()` get the sample variance.

```
mean_variance_output <- battery %>%
  group_by(TypeBat) %>%
  summarize(mean = mean(LPUC), variance = var(LPUC))
mean_variance_output
```

```
## # A tibble: 4 x 3
##   TypeBat  mean variance
##   <fct>   <dbl>    <dbl>
## 1 1        571.    1360.
## 2 2        860.    3619
## 3 3        433     2065.
## 4 4        496.    2427.
```

Here we can see the mean and variance of each battery type.

1.3 Part C

Either by hand or using software, obtain the ANOVA table for your model. It should have three rows for Battery type, error and the corrected total sum of squares.

Here we are going to use our `anova()` function to get an ANOVA table for battery types.

```
battery_model <- lm(LPUC ~ TypeBat, battery)
anova(battery_model)
```

```
## # A tibble: 2 x 5
##      Df `Sum Sq` `Mean Sq` `F value`     `Pr(>F)`
##   <int>    <dbl>     <dbl>     <dbl>        <dbl>
## 1     3  427915.   142638.      60.2  0.000000166
## 2    12   28412.     2368.      NA   NA
```

This ANOVA table is basically complete. We just have to get the corrected total degrees of freedom and sum of squares. The degrees of freedom is just $n * t - 1 = 4 * 4 - 1 = 15$ and this makes sense because the error degrees of freedom is 12 and the battery type degrees of freedom is 3 so added together is 15, which is what we expected. To get the Sum of Squares Total we add the Sum of Squares for `Battery Type` and the Sum of Squares for `Error`. The Sum of Squares for `Battery Type` is 427915.25 and the Sum of Squares for `Error` is 28412.5 so added together gets us 456327.75, which is our Sum of Squares for `Corrected Total`. Thus, we have all the values we need for our final ANOVA table.

Source	DF	SS	MS	F-value	p-value
Battery Type	3	427915.25	142638.4166667	60.243238	0.0000002
Error	12	28412.5	2367.7083333		
Corrected Total	15	456327.75			

1.4 Part D

Two versions of the fitted model from PROC GLM, with parameter estimates $\hat{\mu}, \hat{\tau_1}, \hat{\tau_2}, \hat{\tau_3}, \hat{\tau_4}$ are given. Use these models, say `fit1` and `fit2` to report estimates of each of the linear combinations of parameters below.

Parameter	`fit1`	`fit2`
$\theta_1 = \mu + \tau_1$	$496.25 + 74.50 = 570.75$	$570.75 + 0 = 570.75$
$\theta_2 = \tau_1$	74.50	0
$\theta_3 = \tau_1 - \tau_3$	$-63.25 - 74.50 =$ -137.75	-137.75
$\theta_4 = \tau_3$	-63.25	-137.75
$\theta_5 = \mu + \tau_4$	$496.25 + 0 = 496.25$	$570.75 + (-74.5) = 496.25$
$\theta_6 = \mu + \frac{1}{4}\sum_1^4 \tau_i$	$496.25 + \frac{1}{4} * (74.5 + 364.25 + (-63.25) + 0) = 590.125$	$570.75 + \frac{1}{4} * (0 + 289.75 + (-137.75) + (-74.5)) = 590.125$
$\theta_7 = \tau_2 - \tau_4$	$364.25 - 0 = 364.25$	$289.75 - (-74.5) = 364.25$

1.5 Part E

Which of the seven parameters considered in the table from **part (d)** are uniquely estimable? For each estimable parameter, give a linear combination of the data that has the parameter as its expectation.

We have the following linear combinations that exist:

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Knowing this, we can look at all of them one by one to figure this out.

1.5.1 Part i

We know that θ_1 is estimable. This is because the linear combination of $(1, 1, 0, 0, 0)$ can be used to get this parameter expectation.

1.5.2 Part ii

We know that θ_2 is **NOT** estimable. This is because the linear combination of $(0, 1, 0, 0, 0)$ has be used to get this parameter expectation and there is no way to manipulate the data linear combinations to get to the one we need for this.

1.5.3 Part iii

We know that θ_3 is an estimable. This is because the linear combination of $(0, 1, 0, -1, 0)$ can be used to get this parameter expectation. We can take the linear combination of $(1, 1, 0, 0, 0)$ minus the linear combination of $(1, 0, 0, 1, 0)$ to equal $(0, 1, 0, -1, 0)$.

1.5.4 Part iv

We know that θ_4 is **NOT** estimable. This is because the linear combination of $(0, 0, 0, 1, 0)$ has be used to get this parameter expectation and there is no way to manipulate the data linear combinations to get to the one we need for this.

1.5.5 Part v

We know that θ_5 is estimable. This is because the linear combination of $(1, 0, 0, 0, 1)$ can be used to get this parameter expectation.

1.5.6 Part vi

We know that θ_6 is estimable. This is because the linear combination of $(1, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ can be used to get this parameter expectation. We can get this by adding linear combination of $1, 1, 0, 0, 0)$ to $(1, 0, 1, 0, 0)$ to $(1, 0, 0, 1, 0)$ to $(1, 0, 0, 0, 1)$ to to get $(4, 1, 1, 1, 1)$. Then we can multiply the scalar of $\frac{1}{4}$ to get the linear combination of $\frac{1}{4}(4, 1, 1, 1, 1)$ to get $(1, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Thus, we have shown this estimable.

1.5.7 Part vii

We know that θ_7 is an estimable. This is because the linear combination of $(0, 0, 1, 0, -1)$ can be used to get this parameter expectation. We can take the linear combination of $(1, 0, 1, 0, 0)$ minus the linear combination of $(1, 0, 0, 0, 1)$ to equal $(0, 0, 1, 0, -1)$.

1.6 Part F

Is θ_1 a contrast? How about θ_3?

We know that a contrast is when the $\sum_1^t c_j = 0$.

From **part e**, we know that θ_1 has the linear combination of $(1, 1, 0, 0, 0)$. Thus by taking the $\sum_1^t c_j = \sum_1^4 c_j \neq 0$. We know this because all the values are non-negative (with $c_1 = 1$). So θ_1 is not a contrast.

From **part e**, we know that θ_3 has the linear combination of $(0, 1, 0, -1, 0)$. Thus by taking the $\sum_1^t c_j = \sum_1^4 c_j = 1 + 0 - 1 + 0 = 0$. So θ_3 is a contrast.

1.7 Part G

Compute the standard error of $\theta_3 = \bar{y_1} - \bar{y_3}$.

We know that the standard error of $\hat{\theta}$ is $\hat{SE\hat{\theta}} = \sqrt{MS(E) * \sum_{i=1}^{i=t} \frac{c_i^2}{n_i}}$. To find the standard error for θ_3, we know from *part c* that the MS(E) is 2367.7083333 and $n_i = 5$ for all $i = 1, 2, ..., t$. Now to find $\sum_{i=1}^{i=t} \frac{c_i^2}{n_i} = \frac{1^2}{4} + 0 + \frac{-1^2}{4} + 0 = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. So, $\hat{SE\hat{\theta_3}} = \sqrt{MS(E) * \sum_{i=1}^{i=t} \frac{c_i^2}{n_i}} = (2367.7083333 * \frac{1}{2})^{\frac{1}{2}} =$ 34.4071819. So $\hat{SE(\hat{\theta_3})} = 34.4071819$.

1.8 Part H

Compute the contrast sum of squares for $\hat{\theta_3}$ and also for the least squares estimator of $\hat{\theta_3} = \bar{y_2} - \bar{y_4}$. Which contrast explains more of the battery type effect, $\hat{\theta_3}$ or $\hat{\theta_7}$?

We know that the $SS(\hat{\theta}) = \frac{\hat{\theta}^2}{\frac{c_1^2}{n_1} + ... + \frac{c_t^2}{n_t}}$

Some things we should note that we solved from *part b*:

- $\bar{y_1} = 570.75$
- $\bar{y_2} = 860.5$

- $\bar{y_3} = 433$
- $\bar{y_4} = 496.25$

So we know that $\hat{\theta_3} = \bar{y_1} - \bar{y_3} = 137.75$ and $\hat{\theta_7} = \bar{y_2} - \bar{y_4} = 364.25$.

Now we can solve the sum of squares for both $\hat{\theta_3}$ and $\hat{\theta_7}$.

$SS(\hat{\theta_3}) = \frac{\hat{\theta_3}^2}{\frac{c_1^2}{n_1}+...+\frac{c_t^2}{n_t}} = 137.75\ ^2\ /\ (\frac{1^2}{4} + 0 + \frac{-1^2}{4} + 0) = 18975.0625\ /\ \frac{1}{2} = 37950.125$. So $SS(\hat{\theta_3}) = 37950.125$.

$SS(\hat{\theta_7}) = \frac{\hat{\theta_7}^2}{\frac{c_1^2}{n_1}+...+\frac{c_t^2}{n_t}} = 364.25\ ^2\ /\ (0 + \frac{1^2}{4} + 0 + \frac{-1^2}{4}) = 132678.0625\ /\ \frac{1}{2} = 265356.125$. So $SS(\hat{\theta_7}) = 265356.125$.

The higher sum of squares for $\hat{\theta_7}$ means that the contrast with $\hat{\theta_7}$ explains more of the battery type effect than $\hat{\theta_3}$.

1.9 Part I

Are the contrasts θ_3 and θ_7 orthogonal? What is the covariance of the least squares estimators?

We know that contrasts are orthogonal if $\sum_{i=1}^{t} \frac{c_i * d_i}{n_i} = 0$ For θ_3 the linear combination is (1, 0, -1, 0) and the for θ_7 the linear combination is (0, 1, 0, -1). Therefore $\sum_{i=1}^{t} \frac{c_i * d_i}{n_i} = \frac{1*0}{n_1} + \frac{0*1}{n_2} + \frac{-1*0}{n_3} + \frac{0*-1}{n_4} = 0*0*0*0 = 0$ since $n_i \neq 0$ for $i = 1, 2, ..., t$. Therefore, θ_3 and θ_7 are orthogonal. Moreover, since they are orthogonal, their covariance is equal to 0 since orthogonal values are statistically uncorrelated and moreover independent if normally distributed. Therefore, θ_3 and θ_7 are orthogonal and have a covariance of 0.

2 Problem 2

First we are going to read in the `tensile_strength` data and will show what it looks like.

```
library(tidyverse)
tensile_strength <- read_table("tensile_strength.txt")
tensile_strength$CottonWtPer <- factor(tensile_strength$CottonWtPer)
tensile_strength
```

```
## # A tibble: 25 x 2
##    CottonWtPer TensileStrength
##    <fct>                 <dbl>
##  1 15                        7
##  2 20                       12
##  3 25                       14
##  4 30                       19
##  5 35                        7
##  6 15                        7
##  7 20                       17
##  8 25                       19
##  9 30                       25
## 10 35                       10
## # i 15 more rows
```

Let Y_{ij} denote the observed tensile strength of unit j from cotton weight percentage level i for i, j = 1, . . . , 5. Consider a linear model for Y_{ij} with factorial effects for cotton weight percentage.

2.1 Part A

Report the sample mean, $\bar{y}_i$ and sample variance, s_i^2 for each level of cotton wt. percentage, i.

We can do this by using the `group_by()` function in R to group all the data by cotton weight and then use the `summarize()` function to find the `mean` and `variance` of `TensileStrength` value by each cotton weight type. I am assuming sample variance and sample mean in the question so our functions of `mean()` get the sample mean and `var()` get the sample variance.

```
mean_variance_output2 <- tensile_strength %>%
  group_by(CottonWtPer) %>%
  summarize(sample_mean = mean(TensileStrength),
            sample_variance = var(TensileStrength))
mean_variance_output2
```

```
## # A tibble: 5 x 3
##   CottonWtPer sample_mean sample_variance
##   <fct>             <dbl>           <dbl>
## 1 15                  9.8            11.2
## 2 20                 15.4             9.8
## 3 25                 17.6             4.3
## 4 30                 21.6             6.8
## 5 35                 10.8             8.2
```

Here we found our sample mean and sample variance of the `tensile strength` for each cotton weight type.

2.2 Part B

Write out a statistical model for Y_{ij} which assumes the response varies normally about a weight specific mean, with constant variance.

We know when the variances should be the same for all treatments groups that our model for One Factor Experiments should be $Y_{ij} = \mu + \tau_i + E_{ij}$ for $i = 1, 2, ..., t$ and $j = 1, 2, ..., n$ where t is the number of treatments and n is the sample size in each treatment. Also E_{ij} are i.i.d $N(0, \sigma^2)$ errors. In this case, $t = 5$ and $n = 5$ given that there are 4 battery types and 4 batteries in each type. So knowing all of this our model should be written as $Y_{ij} = \mu + \tau_i + E_{ij}$ for $i = 1, 2, 3, 4, 5$ and $j = 1, 2, 3, 4, 5$ and E_{ij} are i.i.d $N(0, \sigma^2)$ errors.

2.3 Part C

Report the fitted model of the form $\hat{Y_{ij}} = \hat{\mu} + \hat{\tau_i}$ and supply numerical estimates for all location parameters, or functions of location parameters, that would enable a user to predict a future observation for a give cotton weight percentage.

We should look at the values it takes by using the `summary()` function on our fitted model (which uses the `lm()` function). The code is below.

```
tensile_strength_model <- lm(TensileStrength ~ CottonWtPer, tensile_strength)
summary(tensile_strength_model)
```

```
##
## Call:
## lm(formula = TensileStrength ~ CottonWtPer, data = tensile_strength)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -3.8   -2.6    0.4    1.4    5.2
##
## Coefficients:
##               Estimate Std. Error t value    Pr(>|t|)
## (Intercept)      9.800      1.270   7.719 0.000000202 ***
## CottonWtPer20    5.600      1.796   3.119    0.005409 **
## CottonWtPer25    7.800      1.796   4.344    0.000315 ***
## CottonWtPer30   11.800      1.796   6.572 0.000002108 ***
## CottonWtPer35    1.000      1.796   0.557    0.583753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.839 on 20 degrees of freedom
## Multiple R-squared:  0.7469, Adjusted R-squared:  0.6963
## F-statistic: 14.76 on 4 and 20 DF,  p-value: 0.000009128
```

As we can see our model can be written out as $Y_i = 9.8 + 5.6 * x_{2i} + 7.8 * x_{3i} + 11.8 * x_{4i} + 1 * x_{5i}$ where x_{ti} represents an indicator variable if the specific value matched up with the corresponding treatment t (it is equal to 1 if it does match and 0 if it does not match). Note, $t = 1$ for cotton weight of 15, $t = 2$ for cotton weight of 20, $t = 3$ for cotton weight of 25, $t = 4$ for cotton weight of 30, and $t = 5$ for cotton weight of 35.

2.4 Part D

Using level of significance $\alpha = .05$, test the null hypothesis that mean tensile strength is constant across the levels of cotton weight percentage considered in this experiment. Provide all the elements of the test (H_0, critical value, observed value of the test statistic, conclusion).

Here we are going to say that our null hypothesis is that $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. Our alternative hypothesis is that H_A: At least one of the μ_i terms from $i = 1, 2, 3, 4, 5$ does not equal the rest. We can do a F-test to look at this below.

```
anova(tensile_strength_model)
```

```
## # A tibble: 2 x 5
##      Df `Sum Sq` `Mean Sq` `F value`    `Pr(>F)`
##   <int>    <dbl>     <dbl>     <dbl>       <dbl>
## 1     4     476.    119.        14.8  0.00000913
## 2    20     161.      8.06      NA   NA
```

As we can see, we get a F-observed value of 14.7568238 with a numerator degrees of freedom (the `CottonWtPer` df) of 4 and a denominator degrees of freedom (the `Residuals` df) of 20. Using our `qf()` function, we can put in our alpha level of 0.05, our numerator degrees of freedom, and our denominator degrees of freedom to get a F-critical value of 2.8660814. Since our F-observed value is greater than our F-critical value we should reject our null hypothesis that $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. Therefore, we have statistically significant evidence to say that at least one of the μ_i values from $i = 1, 2, 3, 4, 5$ is different from the others. Moreover this means that the mean tensile strength is not constant for all cotton weight percentages considered.

2.5 Part E

Provide an ANOVA table to accompany the analysis.

From *part d*, we saw a partial ANOVA, which is also shown below:

```
anova(tensile_strength_model)
```

```
## # A tibble: 2 x 5
##      Df ‘Sum Sq‘ ‘Mean Sq‘ ‘F value‘    ‘Pr(>F)‘
##   <int>    <dbl>     <dbl>     <dbl>       <dbl>
## 1     4     476.    119.         14.8  0.00000913
## 2    20     161.      8.06       NA   NA
```

We will fill in the rest of the table by filling in the `Total` or also known as the `Corrected Total` section of the table. We need the degrees of freedom and Sum of Squares values. For degrees of freedom, we can get that by adding the `CottonWtPer` (which is 4) and the `Residuals` (which is 20). Therefore the `Corrected Total` degrees of freedom is 24. To get the sum of square value, we can also add the SS(`CottonWtPer`) with the SS(`Residual`). Note: SS(Total) = SS(Model) + SS(Error). Therefore we can see from our partial ANOVA table the SS(`CottonWtPer`) = 475.76 and the SS(`Residual`) = 161.2. So adding them together we can get the SS(`Corrected Total`) = 636.96. Now we can fill in the rest of our table below with these values.

Source	DF	SS	MS	F-value	p-value
Cotton Weight	4	475.76	118.94	14.7568238	0.0000091
Error	20	161.2	8.06		
Corrected Total	24	636.96			

2.6 Part F

What proportion of variability in tensile strength is explained by Cotton Weight Percentage?

We know to find the proportion of variability we need to find the R^2 value. To do this, we can use a formula we learned that $R^2 = \frac{SS(R)}{SS(Tot)}$. Note, from *part e*, we found that the SS(Tot) = 636.96 and the SS(R) = 475.76. Therefore, $R^2 = \frac{SS(R)}{SS(Tot)}$ = 475.76 / 636.96 = 0.7469229. Therefore, we can say that 74.6922884% of the variability in tensile strength is explained by Cotton Weight Percentage.

2.7 Part G

Consider a linear contrast among the five treatment means,

$$\hat{\theta_L} = -2\bar{y_1} - \bar{y_2} + \bar{y_4} + 2\bar{y_5}$$

If μ_i denotes the population mean of units treated with cotton weight percentage i and the vector μ is defined by μ = (μ_1, μ_2, μ_3, μ_4, μ_5)’ then

$$\hat{\theta_L} = (-2, -1, 0, 1, 2)\mu$$

2.7.1 Part i

Compute $\hat{\theta_L}$.

We know that $\hat{\theta_L} = -2\bar{y_1} - \bar{y_2} + \bar{y_4} + 2\bar{y_5}$. From *part a* we know that,

- $\bar{y_1} = 9.8$
- $\bar{y_2} = 15.4$

- $\bar{y_4} = 21.6$
- $\bar{y_5} = 10.8$

Therefore, $\hat{\theta_L} = -2\bar{y_1} - \bar{y_2} + \bar{y_4} + 2\bar{y_5}$ = -2 x 9.8 - 15.4 + 21.6 + 2 x 10.8 = 8.2. So, $\hat{\theta_L}$ = 8.2.

2.7.2 Part ii

Report the standard error for the estimated contrast in part (i), $\hat{SE(}\ \hat{\theta_L})$.

We know that $\hat{SE(}\ \hat{\theta_L}) = \sqrt{MS(E) * \sum_{i=1}^{i=t} \frac{c_i^2}{n_i}}$. To find the standard error for $\hat{\theta_L}$, we know from *part e* that the MS(E) is 8.06 and $n_i = 5$ for all $i = 1, 2, ..., t$. Now to find $\sum_{i=1}^{i=t} \frac{c_i^2}{n_i} = \frac{-2^2}{5} + \frac{-1^2}{5} + 0 + \frac{1^2}{5} + \frac{2^2}{5} = \frac{4}{5} + \frac{1}{5} + \frac{1}{5} + \frac{4}{5} = 2$. So, $\hat{SE(}\ \hat{\theta_L}) = \sqrt{MS(E) * \sum_{i=1}^{i=t} \frac{c_i^2}{n_i}}$ = (8.06 * 2) $^{\frac{1}{2}}$ = 4.014972. So $\hat{SE(}\ \hat{\theta_L})$ = 4.014972.

2.7.3 Part iii

Report the sum of squares associated with the contrast i part (i). Of the variability among the five observed treatment means, what proportion is explained by this contrast? Note that sums of squares are used to quantify observed variability.

We know that the $SS(\hat{\theta_L}) = \frac{\hat{\theta_L}^2}{\frac{c_1^2}{n_1} + ... + \frac{c_t^2}{n_t}}$. We also know from before that:

- $c_1 = -2$ From information about *part g*
- $c_2 = -1$ From information about *part g*
- $c_3 = 0$ From information about *part g*
- $c_4 = 1$ From information about *part g*
- $c_5 = 2$ From information about *part g*
- $n_i = 5$ for $i = 1, 2, ..., t$ From *part g part ii*
- $\frac{c_1^2}{n_1} + ... + \frac{c_t^2}{n_t} = 2$
- $\hat{\theta_L} = 8.2$ From *part g part i*

So to plug it all in we know that $SS(\hat{\theta_L}) = \frac{\hat{\theta_L}^2}{\frac{c_1^2}{n_1} + ... + \frac{c_t^2}{n_t}}$ = 8.2 ^{2} / 2 = 33.62. So $SS(\hat{\theta_L}) = 33.62$.

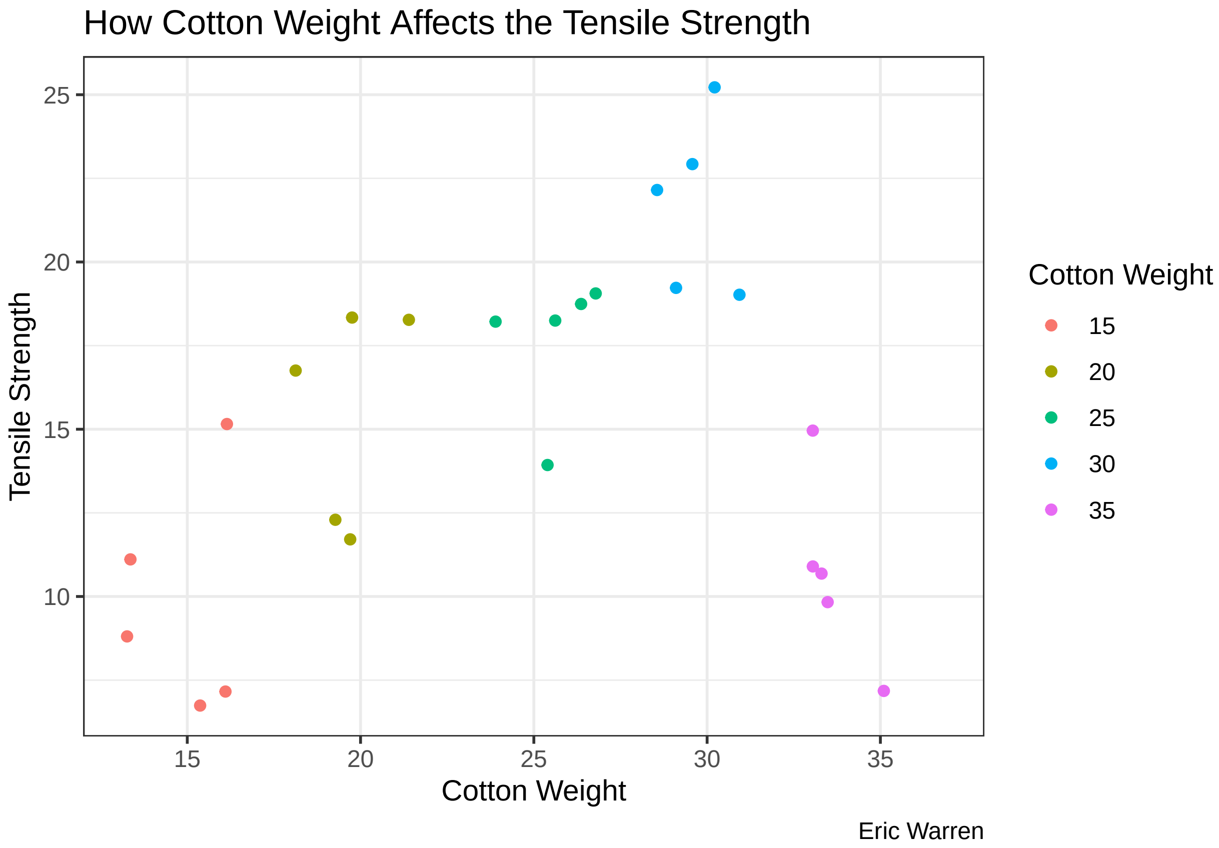
We know to find the proportion of variability we need to find the R^2 value. To do this, we can use a formula we learned that Variability among 5 treatment means = $\frac{SS(\hat{\theta_L})}{SS(R)}$. Note, from *part e*, we found that the SS(R) = 475.76 and the SS($\hat{\theta_L}$) = 33.62. Thus this variability explained = $\frac{SS(R)}{SS(R)}$ = 33.62 / 475.76 = 0.0706659. Therefore, we can say that 7.0665882% of the variability among the five observed treatment means is explained by this contrast.

2.7.4 Part iv

Is too much cotton weight a bad thing? Before conducting any kind of test, do you see, from the table or from a plot of the means, evidence of non-linearity? Explain briefly. Then, using $\alpha = 0.05$, conduct a lack-of-fit for model in which mean tensile strength is a linear function of cotton weight percentage.

First we are going to plot what the data looks like to see if we can see any kind of trend.

```
ggplot(data = tensile_strength, aes(x = CottonWtPer, y = TensileStrength)) +
  geom_point(aes(color = CottonWtPer), position = "jitter") +
  labs(x = "Cotton Weight",
       y = "Tensile Strength",
       title = "How Cotton Weight Affects the Tensile Strength",
       color = "Cotton Weight",
       caption = "Eric Warren") +
  theme_bw()
```



As we can see, there seems to be an increase in strength until we get to the cotton weight of 30 (where it peaks) and then it seems to fall rapidly in strength. From this alone, there seems to be a point where too much cotton weight is a bad thing, as strength seems to be a quadratic trend. To back up this claim we will need to do a lack of fit test. Our null hypothesis H_0 : The Factorial Effects Model can be demonstrated by a Simple Linear Regression Model and our alternative hypothesis H_A : The Factorial Effects Model *cannot* be demonstrated by a Simple Linear Regression Model. First, we should look at the simple linear regression model.

```
# Convert back to numeric before modeling
tensile_strength2 <- tensile_strength
tensile_strength2$CottonWtPer <- as.numeric(tensile_strength2$CottonWtPer)

# Make SLR model
tensile_strength_slr_model <- lm(TensileStrength ~ CottonWtPer, tensile_strength2)

# Print the summary of the model
summary(tensile_strength_slr_model)
```

```
##
## Call:
## lm(formula = TensileStrength ~ CottonWtPer, data = tensile_strength2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9.68  -4.40   1.60   3.78   9.14
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.5800     2.4023   5.237 0.000026 ***
## CottonWtPer   0.8200     0.7243   1.132    0.269
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.122 on 23 degrees of freedom
## Multiple R-squared:  0.05278,    Adjusted R-squared:  0.0116
## F-statistic: 1.282 on 1 and 23 DF,  p-value: 0.2693
```

```
# Print the anova table of model
anova(tensile_strength_slr_model)
```

```
## # A tibble: 2 x 5
##      Df `Sum Sq` `Mean Sq` `F value` `Pr(>F)`
##   <int>    <dbl>     <dbl>     <dbl>    <dbl>
## 1     1     33.6      33.6      1.28    0.269
## 2    23    603.       26.2     NA      NA
```

As we can see here our SS(R_{poly}) $= 33.62$ and our SS(E_{poly}) $= 603.34$. Now this will help us get our F-statistic which can be calculated by $F = \frac{SS(LackofFit)/(t-1-p)}{MS(PureError)}$. Now it is time to get all the parts and plug it in to get our F-statistic.

- We know $MS(PureError) = MS(E_{full})$. From *part e*, we showed that $MS(E_{full}) = 8.06$.
- We know that $t-1-p = t-1-1 = 5-1-1 = 3$. We know from the general description of *Problem 2* that there are 5 treatment groups so $t = 5$. We also know $p = 1$ because our lack of fit test is using a simple linear regression model as a comparison which means the number of predictors in that model is $p = 1$. So $t-1-p = 3$.
- We know that $SS(LackofFit) = SS(Trt) - SS(R_{poly})$. From this part, we showed that SS(R_{poly}) $=$ 33.62. To find SS(Trt), we can look back at *part e* in our ANOVA table to see that this is equal to 475.76. So $SS(Trt) = 475.76$. Therefore, we can now show that $SS(LackofFit) = SS(Trt) - SS(R_{poly}) =$ 442.14. So $SS(LackofFit) = 442.14$.

Now we can get our F-statistic which is $F = \frac{SS(LackofFit)/(t-1-p)}{MS(PureError)}$ = (442.14 / 3) / 8.06 = 18.2853598. So with our F-statistic of 18.2853598, we need to compare this with our F-critical value. This can be found by using F(α, numerator degrees of freedom, denominator degrees of freedom). We were told that $\alpha = 0.05$, the numerator degrees of freedom = 3 (which is just $t-1-p$), and our denominator degrees of freedom = 20 (the same as our Error degrees of freedom from the full one-factor model). So using our `qf()` function in R to get the F-critical value we should get `F(0.05, 3, 20)` = 3.0983912.

Since our F-observed value is greater than our F-critical value we should reject our null hypothesis that the Factorial Effects Model can be demonstrated by a Simple Linear Regression Model. Therefore, we have statistically significant evidence to say that the Factorial Effects Model *cannot* be demonstrated by a Simple Linear Regression Model. Moreover this means that we cannot consider a simple linear regression model to predict the tensile strength from the cotton weight (as this relationship is not linear).