Homework 4

Subhankar Ghosh

ghosh17

Question 1

To know if there is some difference between the operators we will first fit a model using default coding.

```r
library(faraway)

pulp_model = lm(bright~factor(operator), data = pulp)
summary(pulp_model)
```

```
##
## Call:
## lm(formula = bright ~ factor(operator), data = pulp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.440 -0.195 -0.070  0.175  0.560
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        60.2400     0.1458 413.243   <2e-16 ***
## factor(operator)b  -0.1800     0.2062  -0.873   0.3955
## factor(operator)c   0.3800     0.2062   1.843   0.0839 .
## factor(operator)d   0.4400     0.2062   2.134   0.0486 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.326 on 16 degrees of freedom
## Multiple R-squared:  0.4408, Adjusted R-squared:  0.3359
## F-statistic: 4.204 on 3 and 16 DF,  p-value: 0.02261
```

p-value for the F-statistic (0.02261) is quite small (<0.05) so there is some difference between the operators.

To understand the differences we will run the Tuskey's HSD test to come up with the confidence intervals.

```r
TukeyHSD(aov(bright ~ factor(operator), pulp))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = bright ~ factor(operator), data = pulp)
##
## $`factor(operator)`
##      diff         lwr       upr     p adj
## b-a -0.18 -0.76981435 0.4098143 0.8185430
## c-a  0.38 -0.20981435 0.9698143 0.2903038
## d-a  0.44 -0.14981435 1.0298143 0.1844794
## c-b  0.56 -0.02981435 1.1498143 0.0657945
## d-b  0.62  0.03018565 1.2098143 0.0376691
## d-c  0.06 -0.52981435 0.6498143 0.9910783
```

We can see that only d-b difference is significant since for all other pair comparison the confidence interval contains 0 and their corresponding p-values are higher.

Question 2

To know if there is some difference between the operators we will first fit a model using default coding.

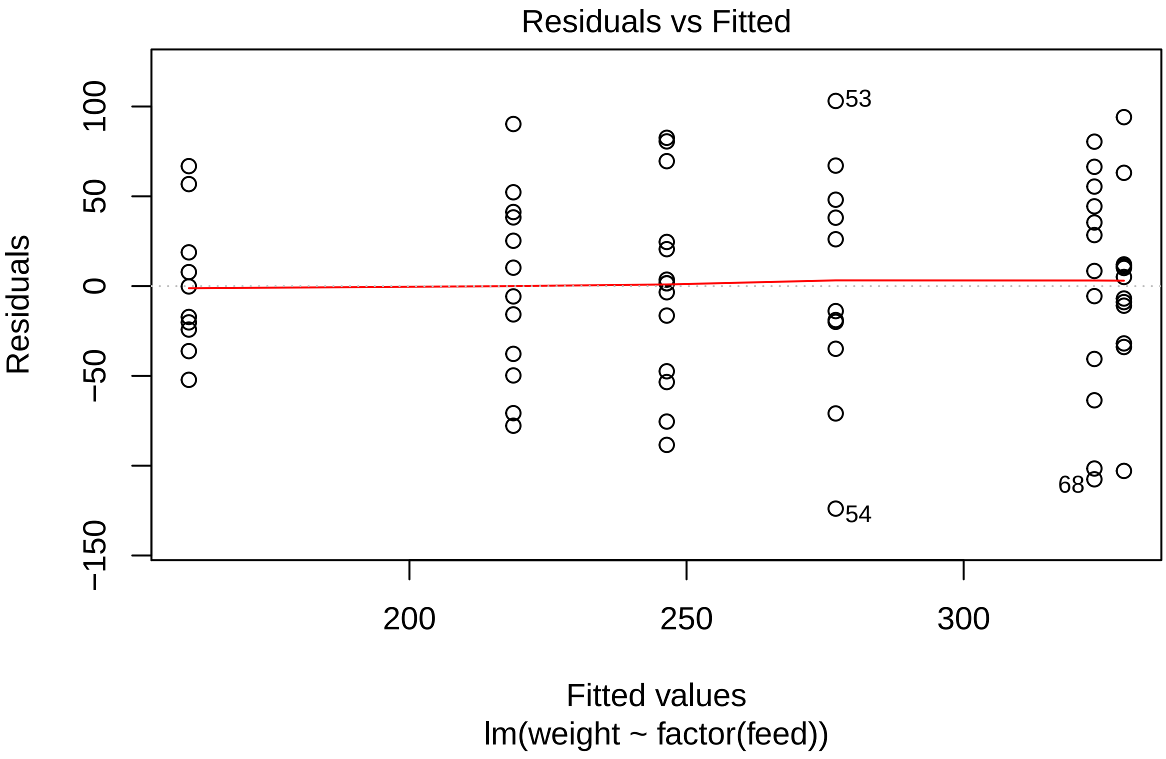
```
chick_model = lm(weight ~ factor(feed), chickwts)
summary(chick_model)
```

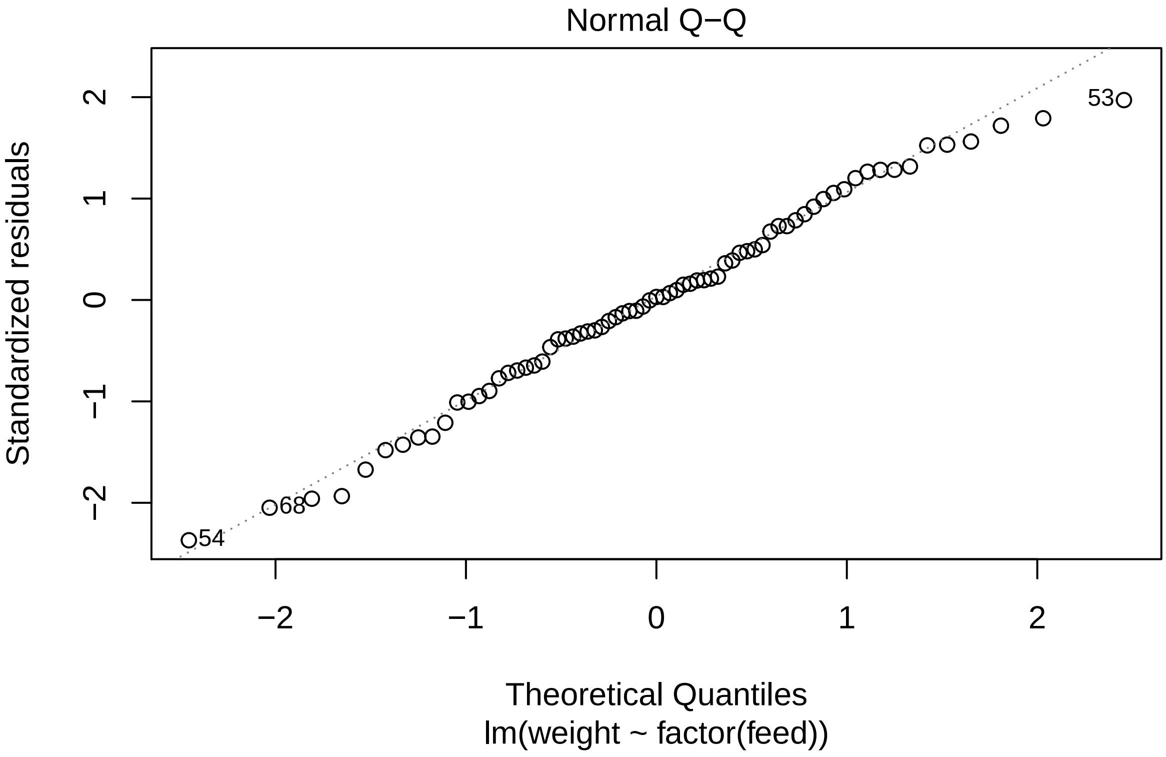
```
##
## Call:
## lm(formula = weight ~ factor(feed), data = chickwts)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.909  -34.413    1.571   38.170  103.091
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             323.583     15.834  20.436  < 2e-16 ***
## factor(feed)horsebean -163.383     23.485  -6.957 2.07e-09 ***
## factor(feed)linseed   -104.833     22.393  -4.682 1.49e-05 ***
## factor(feed)meatmeal   -46.674     22.896  -2.039 0.045567 *
## factor(feed)soybean    -77.155     21.578  -3.576 0.000665 ***
## factor(feed)sunflower    5.333     22.393   0.238 0.812495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.85 on 65 degrees of freedom
## Multiple R-squared:  0.5417, Adjusted R-squared:  0.5064
## F-statistic: 15.36 on 5 and 65 DF,  p-value: 5.936e-10
```

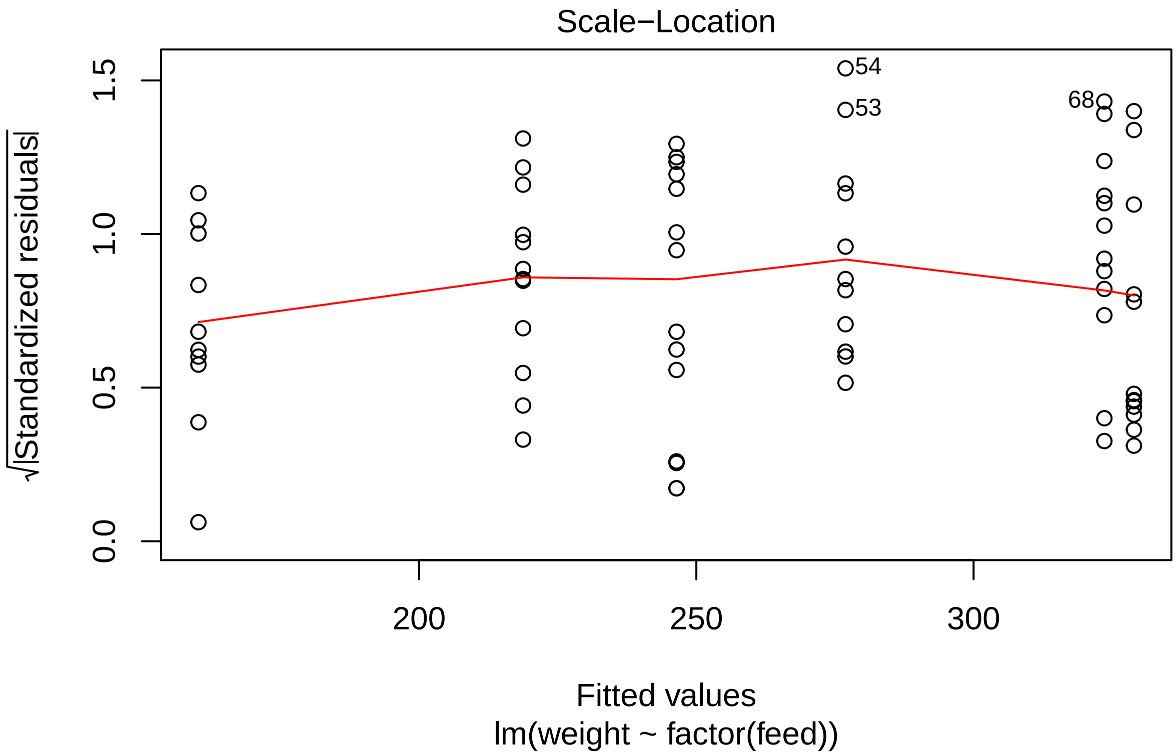
The p-value for F-statistics is very less(almost close to 0) this shows there are lot of differences in the treatments.

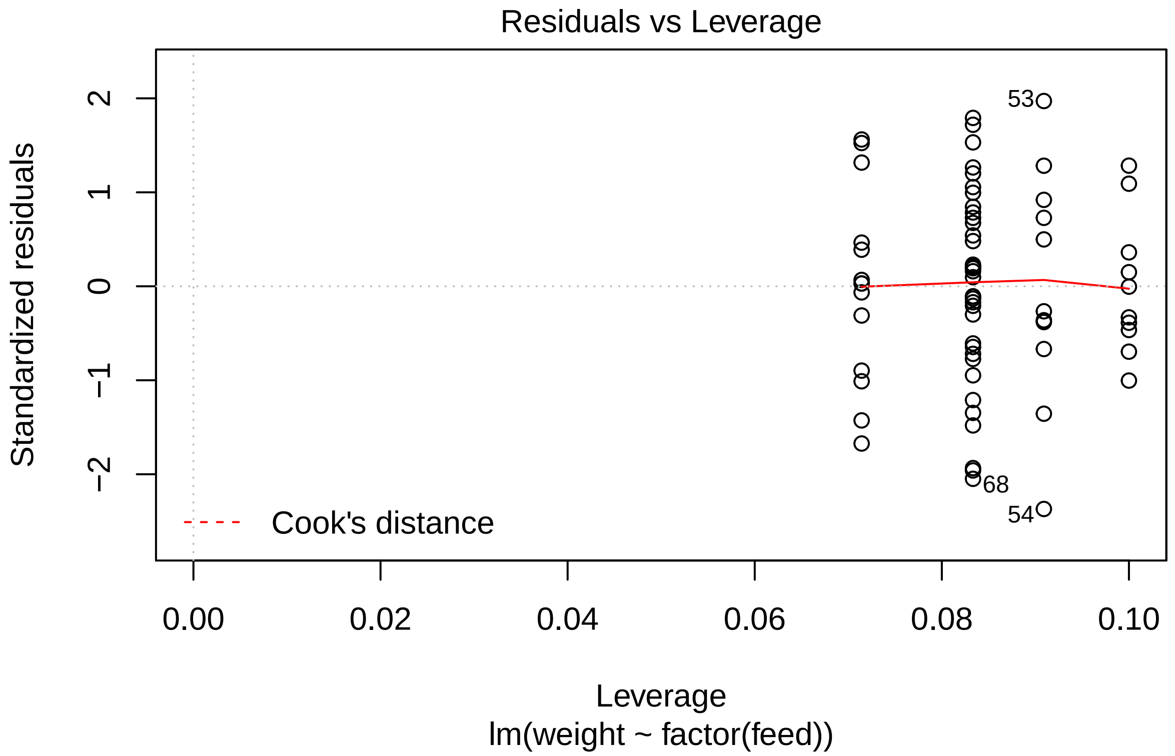
Before we look at the Tuskey's HSD test let us look at the diagnostics plots.

```
plot(chick_model)
```


Residuals vs Fitted
Residuals
100
50
0
−50
−150
200
250
300
53
54
68
Fitted values
lm(weight ~ factor(feed))


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−2
−1
0
1
2
53
68
54
Theoretical Quantiles
lm(weight ~ factor(feed))


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
200
250
300
Fitted values
lm(weight ~ factor(feed))
54
53
68



The residuals vs fitted value plots show that variance is constant. The QQ plot also confirm normality of the errors, although there are some points in the QQ-plot in the top-right corner but overall the it satisfies both constant variance as well as normality assumptions.

From Stardardized Residuals VS Leverage plot we notice point 53, 54 and 68 to be high leverage points.

To understand the differences we will run the Tuskey's HSD test to come up with the confidence intervals.

```
TukeyHSD(aov(chick_model))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = chick_model)
##
## $`factor(feed)`
##                            diff         lwr       upr     p adj
## horsebean-casein    -163.383333 -232.346876 -94.41979 0.0000000
## linseed-casein      -104.833333 -170.587491 -39.07918 0.0002100
## meatmeal-casein      -46.674242 -113.906207  20.55772 0.3324584
## soybean-casein       -77.154762 -140.517054 -13.79247 0.0083653
## sunflower-casein       5.333333  -60.420825  71.08749 0.9998902
## linseed-horsebean     58.550000  -10.413543 127.51354 0.1413329
## meatmeal-horsebean   116.709091   46.335105 187.08308 0.0001062
## soybean-horsebean     86.228571   19.541684 152.91546 0.0042167
## sunflower-horsebean  168.716667   99.753124 237.68021 0.0000000
## meatmeal-linseed      58.159091   -9.072873 125.39106 0.1276965
## soybean-linseed       27.678571  -35.683721  91.04086 0.7932853
## sunflower-linseed    110.166667   44.412509 175.92082 0.0000884
## soybean-meatmeal     -30.480519  -95.375109  34.41407 0.7391356
## sunflower-meatmeal    52.007576  -15.224388 119.23954 0.2206962
## sunflower-soybean     82.488095   19.125803 145.85039 0.0038845
```

Significant difference can be seen between horsebean-casein, linseed-casein, soybean-casein, meatmeal-horsebean, soybean-horsebean, sunflower-linseed, sunflower-soybean.

Question 3

Lets examine the Latin Square

```
matrix(alfalfa$inoculum, 5, 5)
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,] "A"  "D"  "C"  "E"  "B"
## [2,] "B"  "E"  "D"  "A"  "C"
## [3,] "D"  "B"  "A"  "C"  "E"
## [4,] "C"  "A"  "E"  "B"  "D"
## [5,] "E"  "C"  "B"  "D"  "A"
```

Fit the model

```
yield_model = lm(yield ~ shade + irrigation + inoculum, data = alfalfa)
summary(yield_model)
```

```
##
## Call:
## lm(formula = yield ~ shade + irrigation + inoculum, data = alfalfa)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.024 -0.604  0.036  1.016  1.936
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.464      1.263  26.500 5.11e-12 ***
## shade2         2.460      1.108   2.221 0.046343 *
## shade3         3.020      1.108   2.727 0.018376 *
## shade4         4.980      1.108   4.496 0.000731 ***
## shade5         5.080      1.108   4.587 0.000625 ***
## irrigation2   -0.060      1.108  -0.054 0.957688
## irrigation3   -0.740      1.108  -0.668 0.516684
## irrigation4   -1.020      1.108  -0.921 0.375214
## irrigation5   -2.240      1.108  -2.023 0.065993 .
## inoculumB     -0.720      1.108  -0.650 0.527883
## inoculumC     -0.080      1.108  -0.072 0.943607
## inoculumD     -0.860      1.108  -0.776 0.452490
## inoculumE     -6.600      1.108  -5.959 6.62e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.751 on 12 degrees of freedom
## Multiple R-squared:  0.876,  Adjusted R-squared:  0.7519
## F-statistic: 7.062 on 12 and 12 DF,  p-value: 0.0009624
```

The treatment effects are significant since the p-value of the F-statistic 0.0009624 is much much less than 0.05 and thus we need to do a TukeyHSD test to understand the nature of the differences.

Inoculum E seems to be the most significant of the treatments. This also calls for TukeyHSD test.

```
TukeyHSD(aov(yield ~ shade + irrigation + inoculum, data = alfalfa))$inoculum
```

```
##      diff        lwr       upr        p adj
## B-A -0.72  -4.250202  2.810202 0.9633432745
## C-A -0.08  -3.610202  3.450202 0.9999928279
## D-A -0.86  -4.390202  2.670202 0.9326392350
## E-A -6.60 -10.130202 -3.069798 0.0005166455
## C-B  0.64  -2.890202  4.170202 0.9759058775
## D-B -0.14  -3.670202  3.390202 0.9999331812
## E-B -5.88  -9.410202 -2.349798 0.0014163428
## D-C -0.78  -4.310202  2.750202 0.9515868499
## E-C -6.52 -10.050202 -2.989798 0.0005764154
## E-D -5.74  -9.270202 -2.209798 0.0017334480
```

The pairs E-A, E-B, E-C and E-D are significant since they do not have 0 in their confidence interval and their p-adj value is less than significance level 0.05.

Question 4

Looking at the data

```
eggprod
```

```
##    treat block eggs
## 1      O     1  330
## 2      O     2  288
## 3      O     3  295
## 4      O     4  313
## 5      E     1  372
## 6      E     2  340
## 7      E     3  343
## 8      E     4  341
## 9      F     1  359
## 10     F     2  337
## 11     F     3  373
## 12     F     4  302
```

Here the treatment factor is the treat variable and the blocking factor is block column

```
egg_model = lm(eggs ~ factor(block) + treat, data = eggprod)
summary(egg_model)
```

```
##
## Call:
## lm(formula = eggs ~ factor(block) + treat, data = eggprod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.667  -8.125   2.083   5.521  26.000
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       369.92      13.91  26.596 1.87e-07 ***
## factor(block)2    -32.00      16.06  -1.992   0.0934 .
## factor(block)3    -16.67      16.06  -1.038   0.3394
## factor(block)4    -35.00      16.06  -2.179   0.0721 .
## treatF             -6.25      13.91  -0.449   0.6690
## treatO            -42.50      13.91  -3.056   0.0224 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.67 on 6 degrees of freedom
## Multiple R-squared:  0.7381, Adjusted R-squared:  0.5199
## F-statistic: 3.382 on 5 and 6 DF,  p-value: 0.08504
```

In this case we do not get any evidence of differences among treatment factors since the p-value for F-statitic is not small enough. To confirm this observation we will perform a TukeyHSD test

```
TukeyHSD(aov(eggs ~ factor(block) + treat, data = eggprod))$treat
```

```
##       diff       lwr        upr      p adj
## F-E  -6.25 -48.92641 36.4264142 0.89650787
## O-E -42.50 -85.17641  0.1764142 0.05078737
## O-F -36.25 -78.92641  6.4264142 0.08917699
```

As expected we could not find any significant difference among the treatment factors.