Stat 425 Homework 5

Sharvi Tomar

Contents

Problem 1

Use the chickwts data to fit a one-way ANOVA with weights as the response and feed as the predictor.

```
library(faraway)
attach(chickwts)
attributes(feed)
```

```
## $levels
## [1] "casein"    "horsebean" "linseed"   "meatmeal"  "soybean"   "sunflower"
##
## $class
## [1] "factor"
```

```
contrasts(feed)=contr.treatment(6)
g=lm(weight~feed)
summary(g)
```

```
##
## Call:
## lm(formula = weight ~ feed)
##
## 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 ***
## feed2       -163.383     23.485  -6.957 2.07e-09 ***
## feed3       -104.833     22.393  -4.682 1.49e-05 ***
## feed4        -46.674     22.896  -2.039 0.045567 *
## feed5        -77.155     21.578  -3.576 0.000665 ***
## feed6          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
```

(a) Determine whether there are any differences in the weights of chickens according to their feed.

```
null = lm(weight ~ 1)
anova(null, g)
```

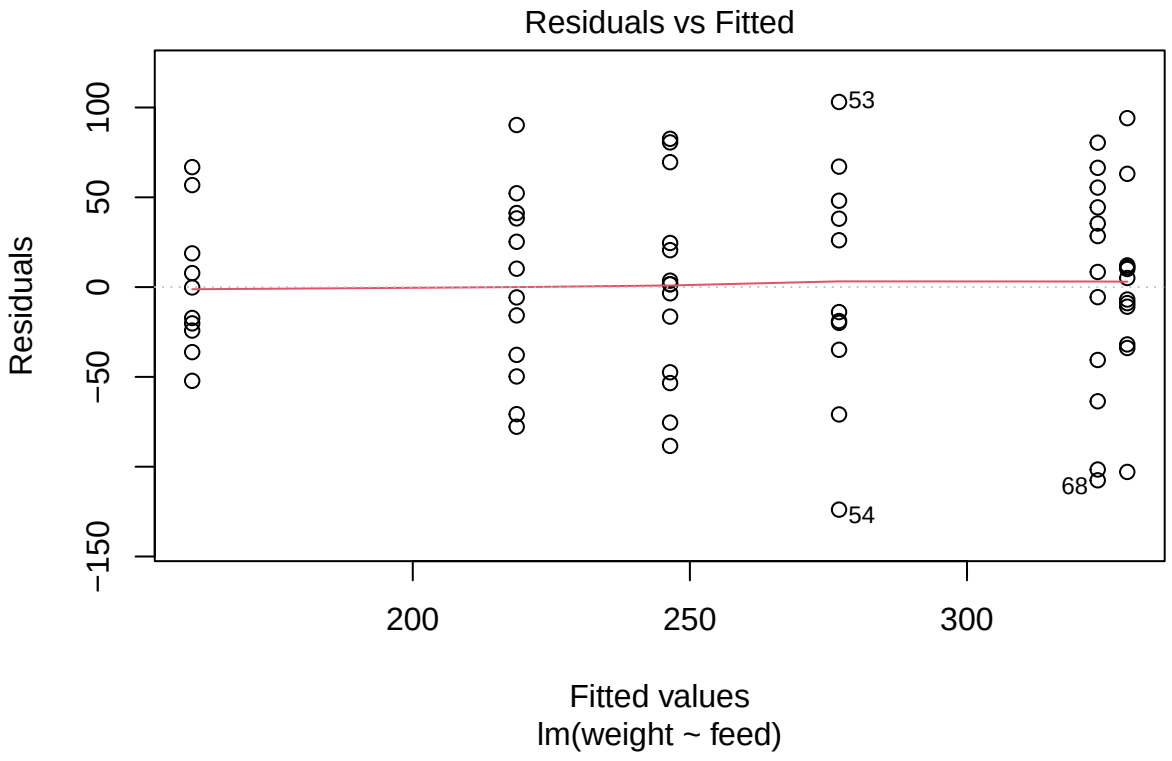
```
## Analysis of Variance Table
##
## Model 1: weight ~ 1
## Model 2: weight ~ feed
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     70 426685
## 2     65 195556  5    231129 15.365 5.936e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value of 5.936e-10 <0.05, hence we reject the null hypothesis and conclude that there is difference in weights of chicken according to their feed.

(b) Perform all necessary model diagnostics.

1. Performing Levene's test for equality of variance

```
plot(g, which=1)
```



```
summary(lm(abs(g$res) ~ feed))
```

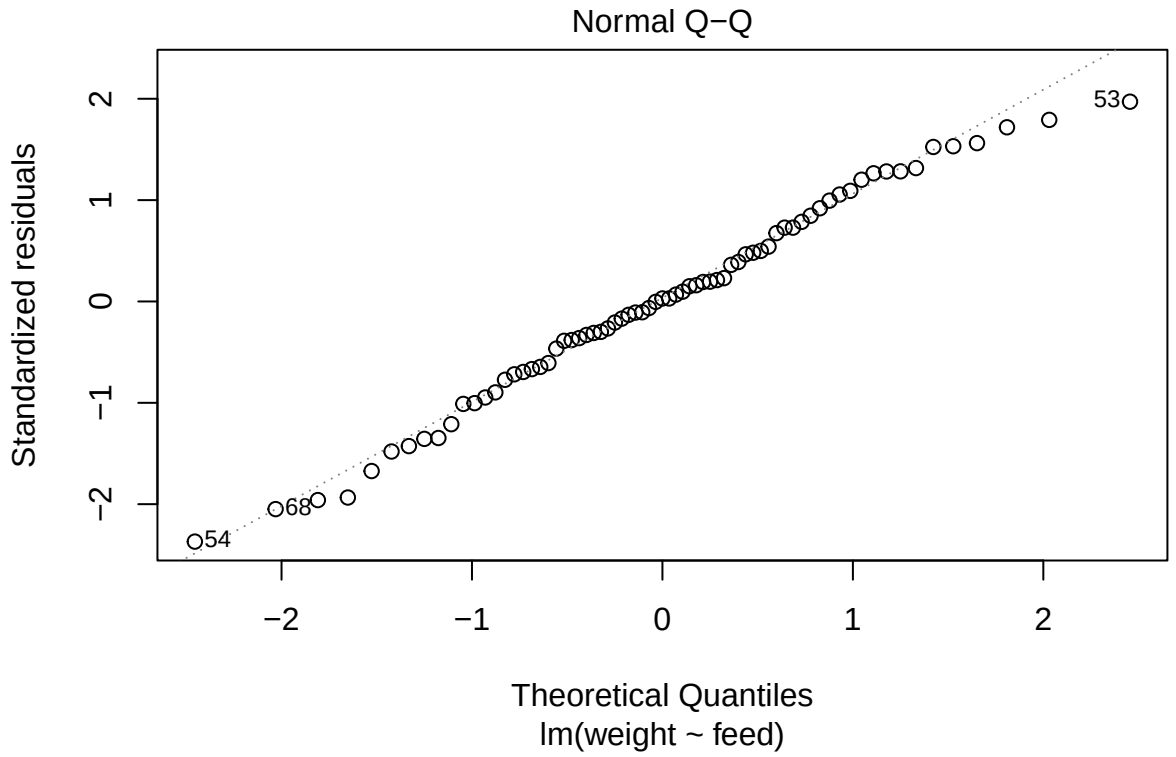
```
##
## Call:
## lm(formula = abs(g$res) ~ feed)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47.57 -23.95  -5.84  24.46  72.55
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   53.153      9.173   5.794 2.18e-07 ***
## feed2        -23.113     13.606  -1.699   0.0942 .
## feed3        -10.236     12.973  -0.789   0.4330
## feed4         -1.797     13.265  -0.136   0.8926
## feed5        -12.500     12.501  -1.000   0.3211
## feed6        -20.569     12.973  -1.586   0.1177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.78 on 65 degrees of freedom
## Multiple R-squared:  0.07059,    Adjusted R-squared:  -0.0009059
## F-statistic: 0.9873 on 5 and 65 DF,  p-value: 0.4324
```

From the residuals vs. fitted values plot, we see that the points appear to be randomly spread out about the line, with no discernible non-linear trends or indications of non-constant variance. Hence we conclude that the constant variance assumption holds true.

The same result is obtained from Levene's test since the p-value>0.01, there is no evidence of unequal variances.

2. Performing Shapiro-Wilk test for normality

```
plot(g, which=2)
```




```
shapiro.test(residuals(g))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(g)
## W = 0.98616, p-value = 0.6272
```

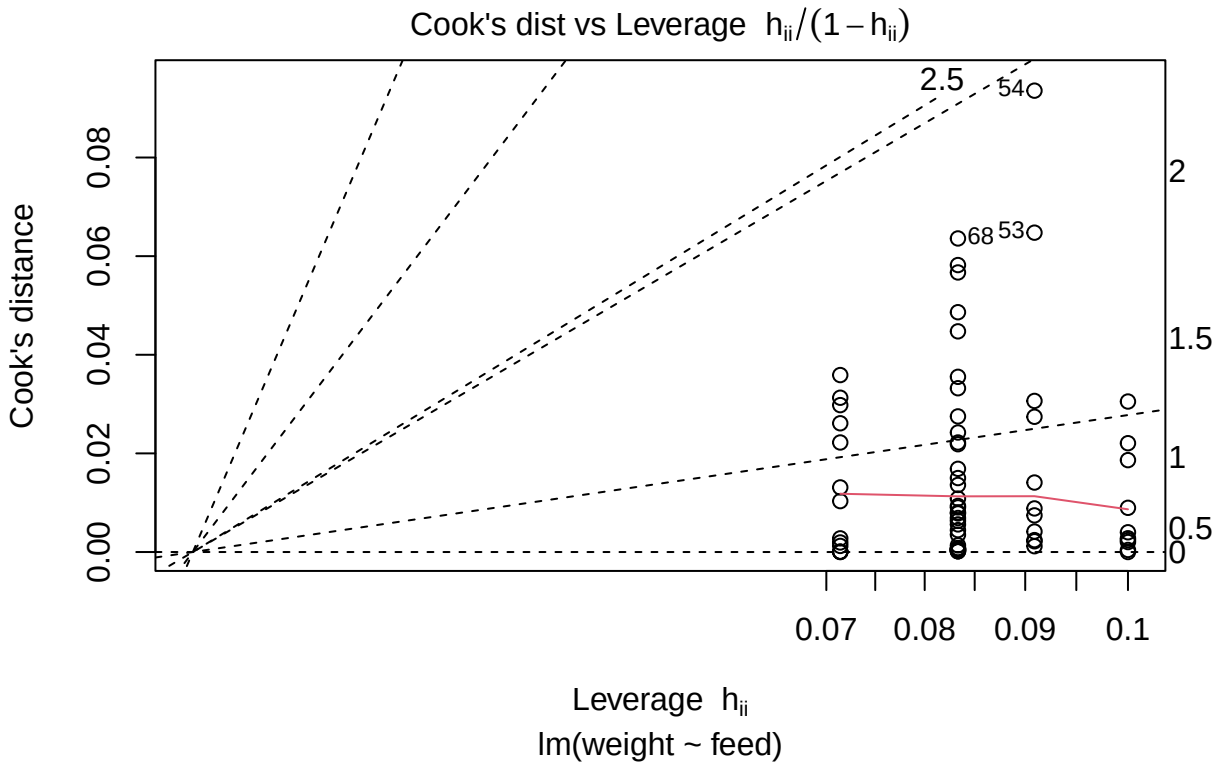
From the Q-Q plot, we can see that there is a straight and hence the normality assumption holds true.

We get the same confirmation from the Shapiro-Wilk normality test. The p-value>0.05, hence we fail to reject the null hypothesis and conclude that the normality of residuals assumption holds true.

3. Detecting unusual observations

a. Leverage

```
plot(g, which=6)
```



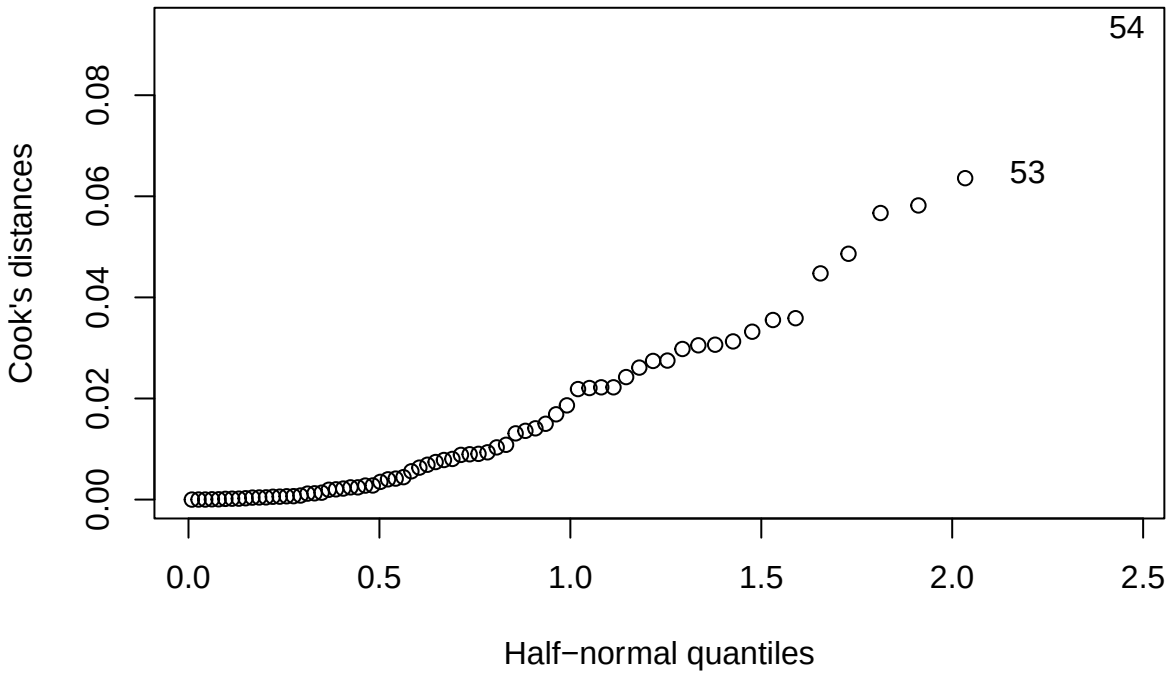
Observations which have high Cook's distance value as well as leverage require special handling. Such observations are: 54, 53 and 64 are worth investigating further.

b. Influential Observations

```
cook = cooks.distance(g)
sort(cook, decreasing = TRUE)[1:5]
```

```
##         54         53         68         42         69
## 0.09355994 0.06476261 0.06358824 0.05819132 0.05669330
```

```
halfnorm(cook, labs=row.names(chickwts), ylab="Cook's distances")
```



Observations with high Cook's distance are: Observations-54, 53 and 68. Observation 54 has the highest cook's distance value (although not close 1), but a very high cook's distance value in comparison to other observations it may be considered as an influential observation and should be handled differently. We may remove this observation for model fitting.

c. Outliers

```
jack=rstudent(g)
nn=dim(chickwts)[1]
qt(0.5/(2*nn),g$df.residual-1)
```

```
## [1] -2.784499
```

```
sort(abs(jack),decreasing=TRUE)[1:10]
```

```
##       54       68       53       42       69       37       11       35
## 2.459615 2.101780 2.017235 2.004747 1.977177 1.823298 1.745395 1.697061
##       27       64
## 1.580108 1.547652
```

Observation 54 has a studentized quantile value close to ~2.7., hence it may be considered outlier. We may remove this observation for model fitting.

Problem 2

Use the infmort data to fit a one-way ANOVA with income as the response and region as the predictor.

```
#infmort
attach(infmort)
attributes(region)
```

```
## $levels
## [1] "Africa"   "Europe"   "Asia"     "Americas"
##
## $class
## [1] "factor"
```

```
contrasts(region)=contr.treatment(4)
g2=lm(income~region)
summary(g2)
```

```
##
## Call:
## lm(formula = income ~ region)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2634.2  -515.9  -192.2     7.8  4583.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    273.2      180.5   1.514   0.1332
## region2       2767.0      306.8   9.020 1.29e-14 ***
## region3        365.6      263.6   1.387   0.1685
## region4        666.6      284.1   2.346   0.0209 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1052 on 101 degrees of freedom
## Multiple R-squared:  0.4641, Adjusted R-squared:  0.4482
## F-statistic: 29.16 on 3 and 101 DF,  p-value: 1.157e-13
```

(a) Determine whether income varies with region. Perform all necessary model diagnostics.

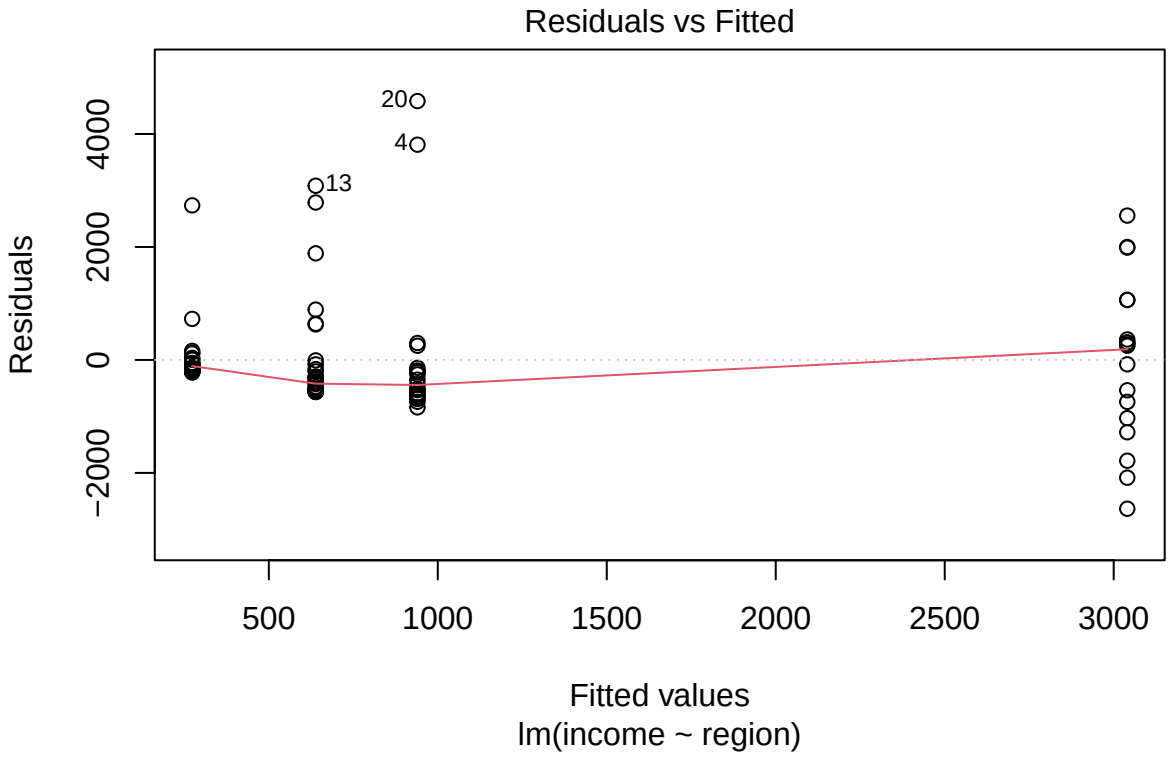
```
null2 = lm(income ~ 1)
anova(null2, g2)
```

```
## Analysis of Variance Table
##
## Model 1: income ~ 1
## Model 2: income ~ region
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1    104 208736257
## 2    101 111857493  3  96878763 29.158 1.157e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value<0.05, hence we reject the null hypothesis and conclude that there is difference in income according to the region.

1. Performing Levene's test for equality of variance

```
plot(g2, which=1)
```




```
summary(lm(abs(g2$res) ~ region))
```

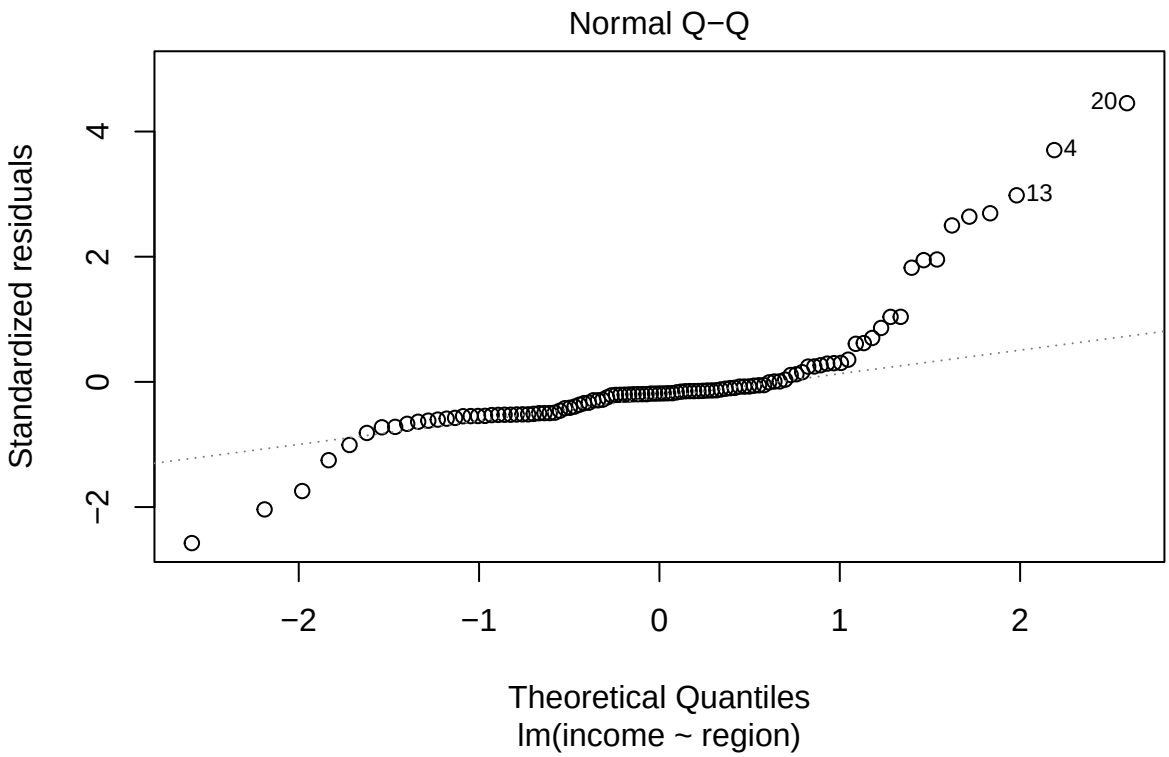
```
##
## Call:
## lm(formula = abs(g2$res) ~ region)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1052.9  -283.5  -118.5   -38.1  3805.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    230.3      132.2   1.742 0.084576 .
## region2        899.8      224.7   4.004 0.000119 ***
## region3        431.0      193.1   2.232 0.027827 *
## region4        547.6      208.1   2.631 0.009853 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 770.9 on 101 degrees of freedom
## Multiple R-squared:  0.1485, Adjusted R-squared:  0.1232
## F-statistic: 5.869 on 3 and 101 DF,  p-value: 0.0009796
```

From the residuals vs. fitted values plot, we see that the points don't appear to be randomly spread out about the line. We should investigate further using Levene's test.

From Levene's test since the p-value<0.01,we reject the null hypothesis and conclude that there is evidence of unequal variances.

2. Performing Shapiro-Wilk test for normality

```
plot(g2, which=2)
```




```
shapiro.test(residuals(g2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(g2)
## W = 0.75507, p-value = 6.061e-12
```

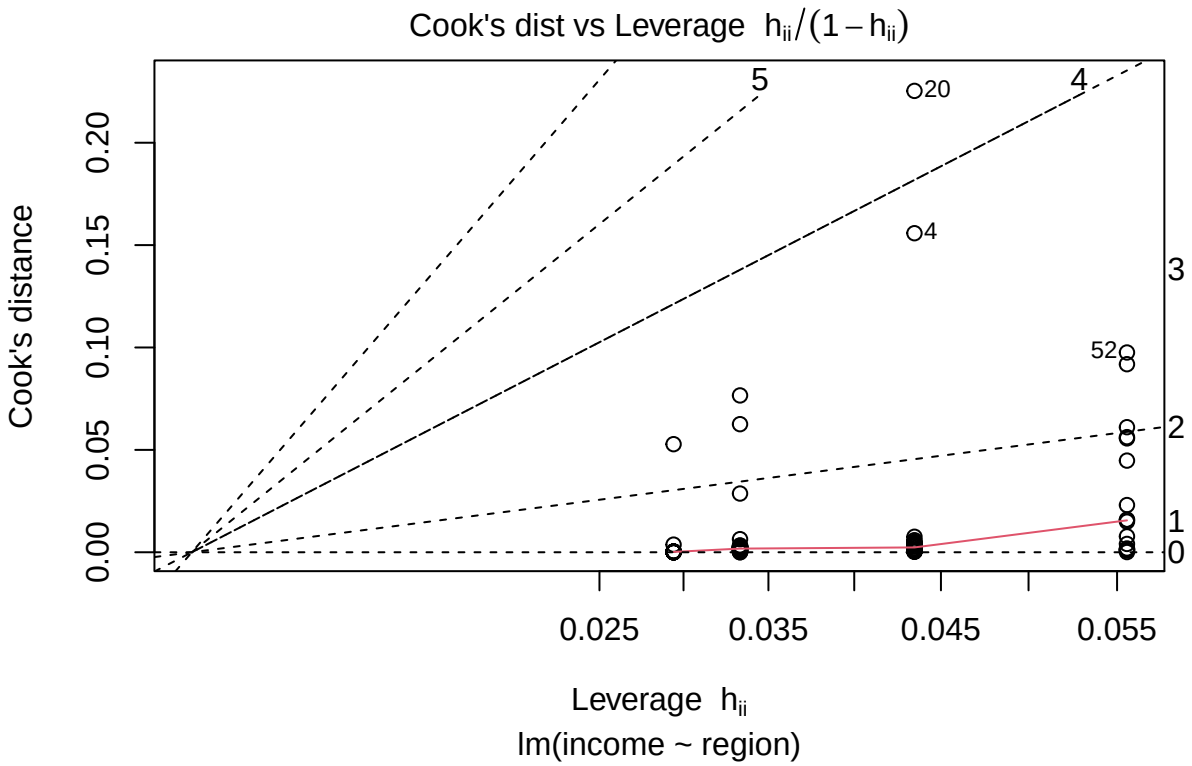
From the Q-Q plot, we can see that there is an increasing trend however not a perfect straight. We should investigate further with Shapiro-Wilk test.

We get the same confirmation from the Shapiro-Wilk normality test. The p-value<0.05, hence we reject the null hypothesis and conclude that the normality of residuals assumption does not hold true.

3. Detecting unusual observations

a. Leverage

```
plot(g2, which=6)
```



Observations which have high Cook's distance value as well as leverage require special handling. Such observations are: 20 and 52 are worth investigating further.

b. Influential Observations

```
cook2 = cooks.distance(g2)
sort(cook2, decreasing = TRUE)[1:5]
```

```
##         20          4         52         17         13
## 0.22532180 0.15580680 0.09756090 0.09183688 0.07659278
```

Observations 20 has a very high cook's distance value as compared to other observations, we may consider removing it for model fitting.

c. Outliers

```
jack2=rstudent(g2)
nn2=dim(infmort)[1]
qt(0.5/(2*nn2),g2$df.residual-1)
```

```
## [1] -2.887335
```

```
sort(abs(jack2),decreasing=TRUE)[1:10]
```

```
##       20        4       13        1       26       52       17       15
## 4.942432 3.963280 3.105683 2.782119 2.722136 2.651459 2.567212 2.070814
##        8        5
## 1.983539 1.972203
```

Observations 20, 4, 13 have a studentized quantile value close $> 2.8.$, hence they may be considered outliers. We may remove this observation for model fitting.

(b) In case income varies with region, determine which pairs of regions are different.

We use Pairwise t-test with Bonferroni Correction to determine the pairs of region which have difference in incomes.

```
pairwise.t.test(income,region,p.adjust.method = "bonferroni")
```

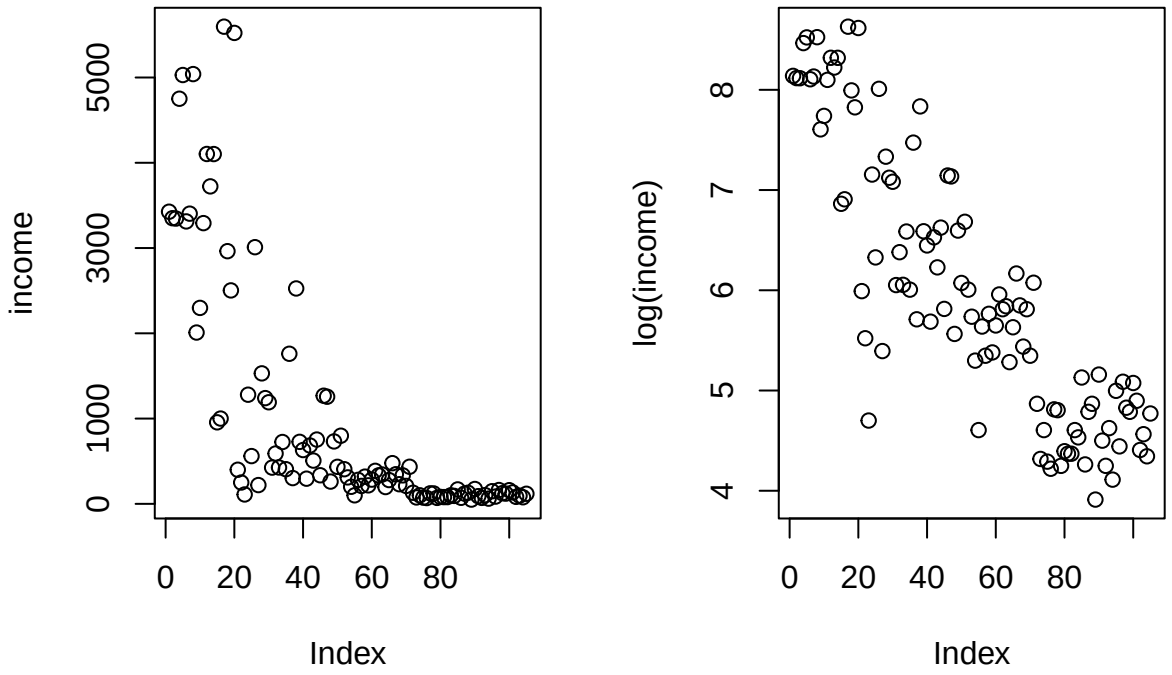
```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  income and region
##
##          Africa  Europe  Asia
## Europe   7.7e-14 -       -
## Asia     1.00    7.2e-11 -
## Americas 0.13    3.9e-08 1.00
##
## P value adjustment method: bonferroni
```

For pair-wise computed p-values<0.05 we conclude those regions to have significant differences in incomes. Such regions where there is significant difference in income are:

1. Europe and Africa
2. Asia and Europe
3. Americas and Europe

(c) In case you need a transformation for the response, re-fit the model and make a comparison with the previous results.

```
par(mfrow=c(1,2))
plot(income)
plot(log(income))
```



Log transformation of response variable results in a linearly distributed scatter plot. We transform the variable with a log transformation.

```
transformed_income=log(income)
contrasts(region)=contr.treatment(4)
g2_2=lm(transformed_income~region)
summary(g2_2)
```

```
##
## Call:
## lm(formula = transformed_income ~ region)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85632 -0.68738 -0.08462  0.41350  2.92786
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.0818     0.1630  31.182  < 2e-16 ***
## region2       2.7808     0.2770  10.039  < 2e-16 ***
## region3       0.5800     0.2380   2.437   0.0166 *
## region4       1.2586     0.2566   4.906 3.57e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9503 on 101 degrees of freedom
## Multiple R-squared:  0.5159, Adjusted R-squared:  0.5015
## F-statistic: 35.88 on 3 and 101 DF,  p-value: 7.171e-16
```

Since the p-value<0.05, hence we reject the null hypothesis and conclude that there is difference in log(income) according to the region.

```
pairwise.t.test(transformed_income,region,p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  transformed_income and region
##
##          Africa  Europe  Asia
## Europe   4.4e-16 -       -
## Asia     0.099   4.1e-11 -
## Americas 2.1e-05 1.0e-05 0.069
##
## P value adjustment method: bonferroni
```

For pair-wise computed p-values<0.05 we conclude those regions to have significant differences in log(incomes). Such regions where there is significant difference in income are:

1. Europe and Africa
2. Asia and Europe
3. Americas and Africa
4. Americas and Europe

** Comparison with previous results:**

After the transformation of response variable, we still get the same result that the income varies with region as well as log(income) varies with region.

After the transformation in addition to the regions obtained previously, we get 'Americas and Africa' regions have signficant diffences in log of incomes.

Problem 3

Use the PlantGrowth data to determine whether there are any differences between the groups.

```
#PlantGrowth
attach(PlantGrowth)
```

```
## The following object is masked from chickwts:
##
##     weight
```

```
attributes(group)
```

```
## $levels
## [1] "ctrl" "trt1" "trt2"
##
## $class
## [1] "factor"
```

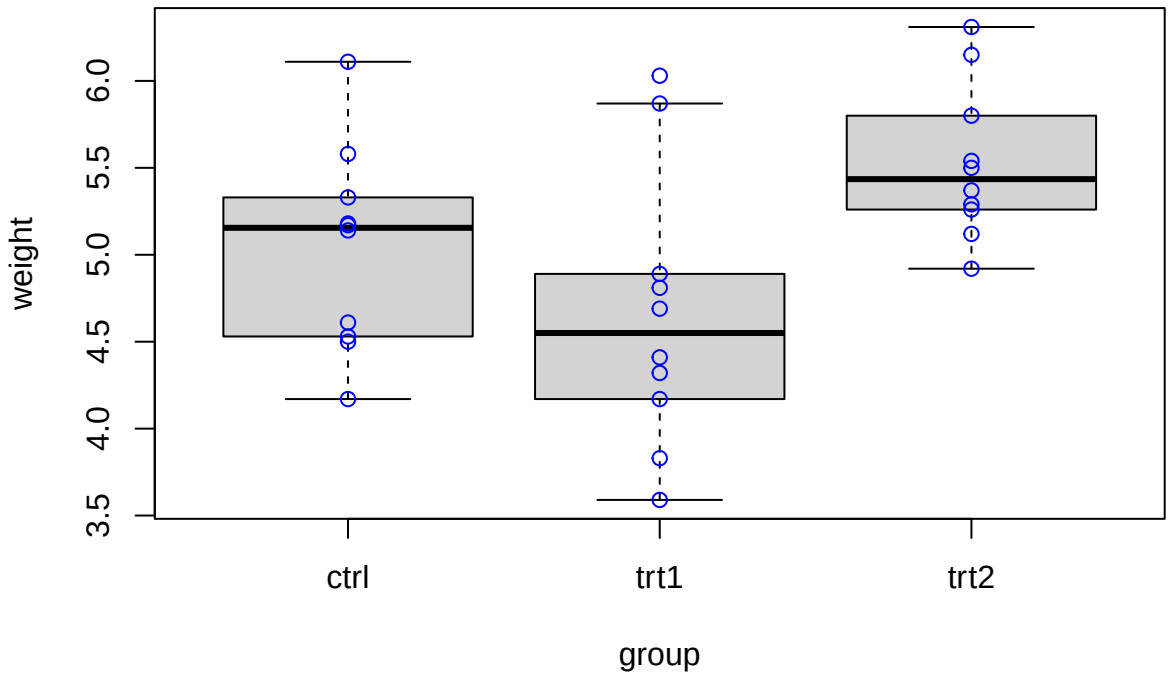
```
contrasts(group)=contr.treatment(3)
g3=lm(weight~group)
summary(g3)
```

```
##
## Call:
## lm(formula = weight ~ group)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0710 -0.4180 -0.0060  0.2627  1.3690
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.0320     0.1971  25.527   <2e-16 ***
## group2       -0.3710     0.2788  -1.331   0.1944
## group3        0.4940     0.2788   1.772   0.0877 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6234 on 27 degrees of freedom
## Multiple R-squared:  0.2641, Adjusted R-squared:  0.2096
## F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591
```

Since the p-value of 0.01591<0.05, we reject the null hypothesis and conclude that there is a difference in weights according to the groups.

(a) Make an appropriate plot to help you explain the nature of these differences. Comment on your results.

```
boxplot(weight~group,outline=FALSE)
stripchart(weight~group,vertical=TRUE,
           add=TRUE,col="blue",pch=1)
```

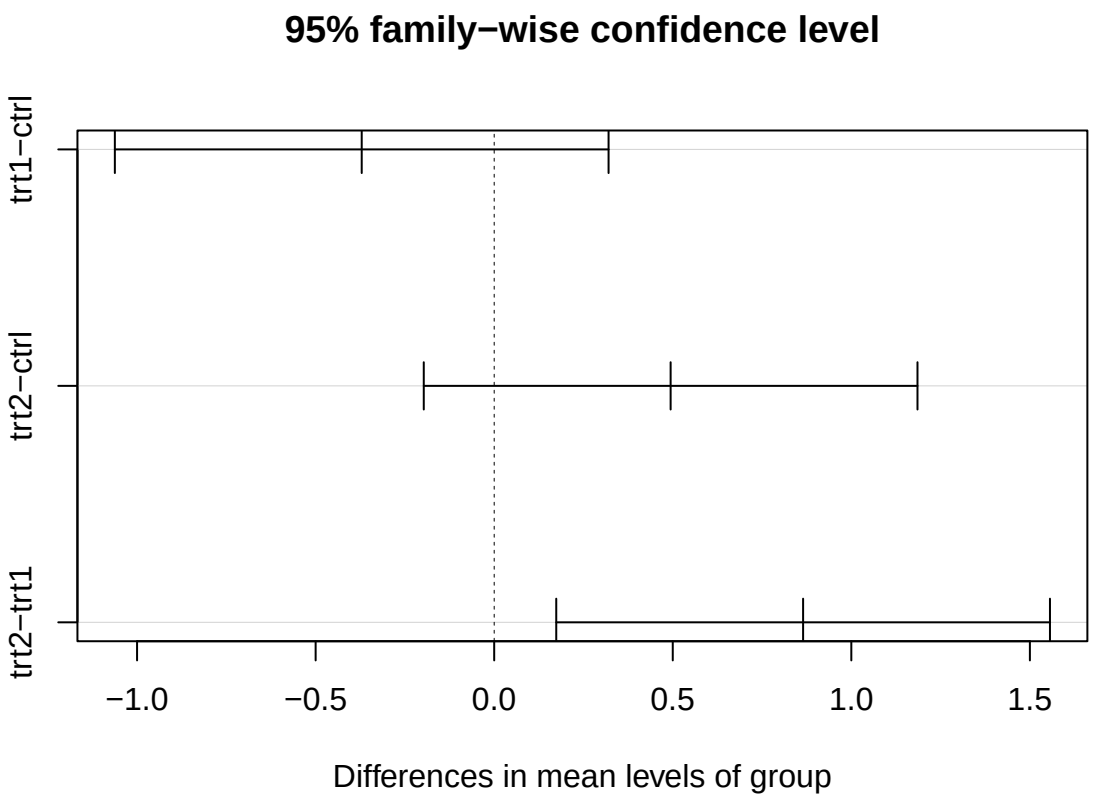


```
g3aov=aov(weight~group,data=PlantGrowth)
```

```
tci=TukeyHSD(g3aov, conf.level=.95)
tci
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight ~ group, data = PlantGrowth)
##
## $group
##              diff        lwr       upr     p adj
## trt1-ctrl -0.371 -1.0622161 0.3202161 0.3908711
## trt2-ctrl  0.494 -0.1972161 1.1852161 0.1979960
## trt2-trt1  0.865  0.1737839 1.5562161 0.0120064
```

```
plot(tci)
```



There are 3 groups - which gives 3 possible pairwise comparisons between them: trt1-ctrl, trt2-ctrl, and trt2-trt1.

The plot has all these comparisons displayed at different heights with the label on the left y axis side.

The x-axis represent the mean differences that were found between those pairs. The extended lines show the 95% confidence intervals.

In case if the confidence interval crosses the 0 point - the difference would not be statistically significant.

Based on this we could reason that there is significant difference in 'trt1 and trt2' while all the other comparisons 'trt2 and ctrl' & 'trt1 and ctrl' are not.

(b) Test whether there is a significant difference between the average yield for the two treatments and the control.

```
library(DescTools)
ScheffeTest(g3aov,contrasts=matrix(c(1,-0.5,-0.5),ncol=1))
```

```
##
##   Posthoc multiple comparisons of means: Scheffe Test
##     95% family-wise confidence level
##
## $group
##                   diff     lwr.ci    upr.ci   pval
## ctrl-trt1,trt2 -0.0615 -0.6868163 0.5638163 0.9681
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the p-val>0.05 hence we fail to reject the null hypothesis and conclude that the is no significant difference between the difference between the average yields for the two treatments and control.