Examples for 11/19/15, Part 2

One-Way ANOVA and Multiple Linear Regression

```
Wheat <- c(5.2,4.5,6.0,6.1,6.7,5.7)
Barley <- c(6.5,8.0,6.1,7.5,5.9,5.6)
Maize <- c(5.8,4.7,6.4,4.9,6.0,5.2)
Oats <- c(8.3,6.1,7.8,7.0,5.6,7.2)

Grain <- c(rep("Wheat",6), rep("Barley",6), rep("Maize",6), rep("Oats",6))
Thiamin <- c(Wheat, Barley, Maize, Oats)

Cereal <- data.frame(Grain, Thiamin)
is.factor(Cereal$Grain)
```

```
## [1] TRUE
```

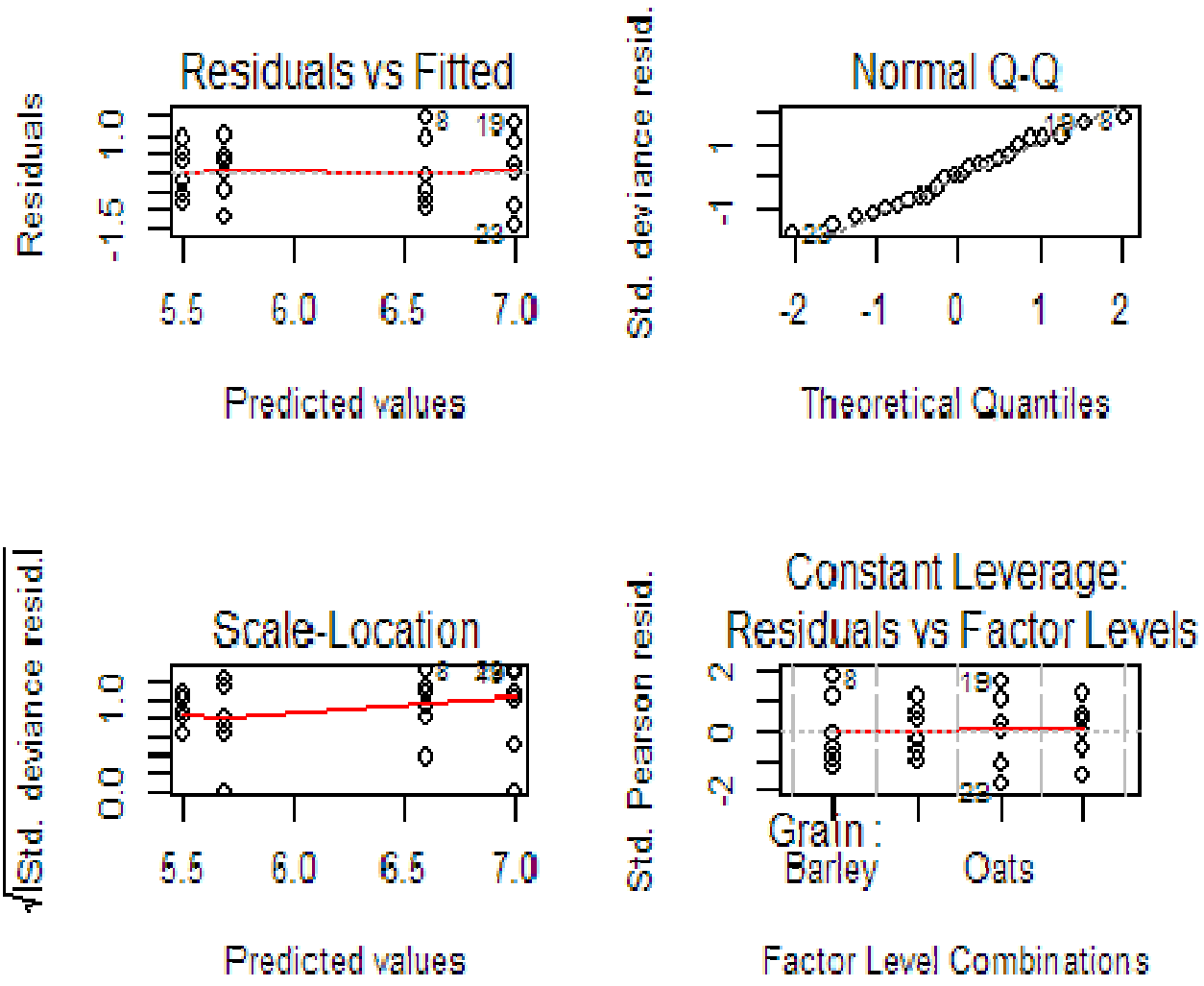
```
fit <- glm(Thiamin ~ Grain, data = Cereal)
summary(fit)
```

```
##
## Call:
## glm(formula = Thiamin ~ Grain, data = Cereal)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.400  -0.625   0.000   0.575   1.400
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.6000     0.3517  18.768 3.62e-14 ***
## GrainMaize   -1.1000     0.4973  -2.212   0.0388 *
## GrainOats     0.4000     0.4973   0.804   0.4307
## GrainWheat   -0.9000     0.4973  -1.810   0.0854 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.742)
##
##     Null deviance: 24.08  on 23  degrees of freedom
## Residual deviance: 14.84  on 20  degrees of freedom
## AIC: 66.572
##
## Number of Fisher Scoring iterations: 2
```

```r
summary(aov(fit))
```

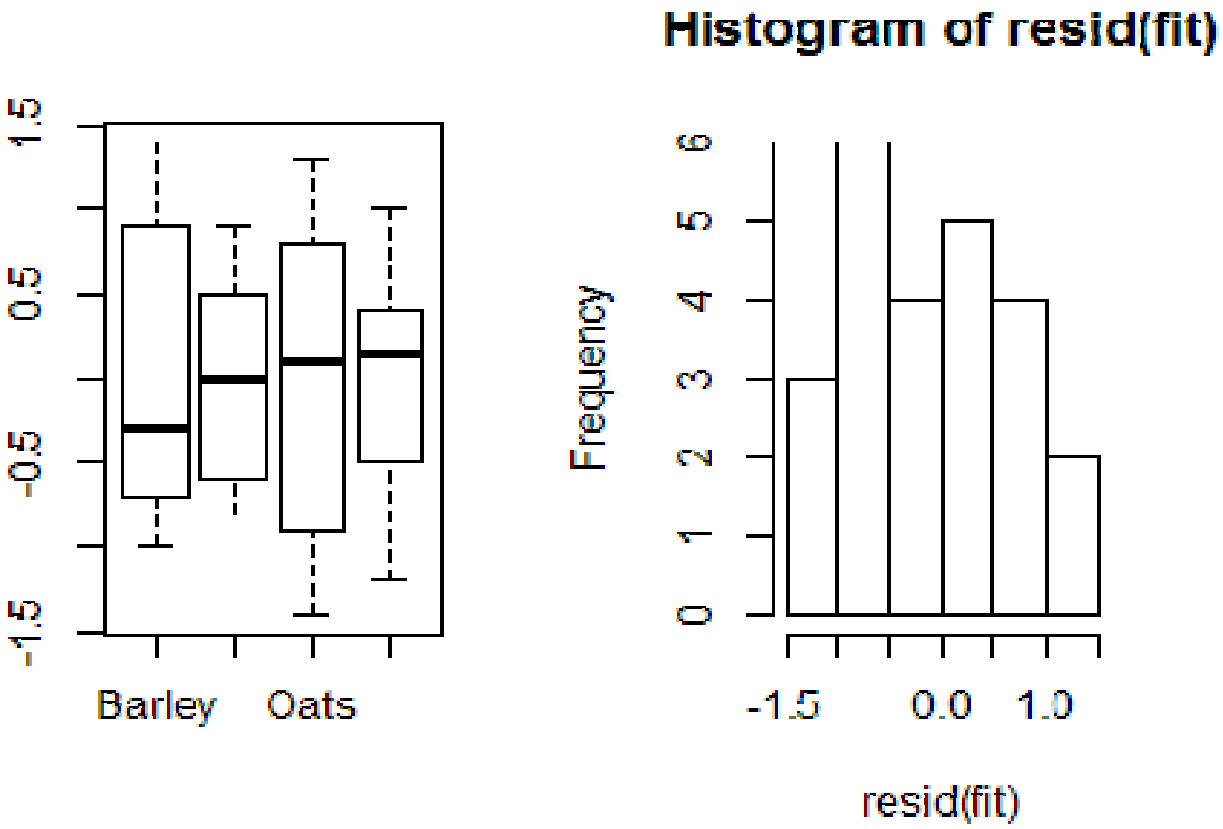
```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Grain        3   9.24   3.080   4.151 0.0194 *
## Residuals   20  14.84   0.742
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
par(mfrow=c(2,2))
plot(fit)
```



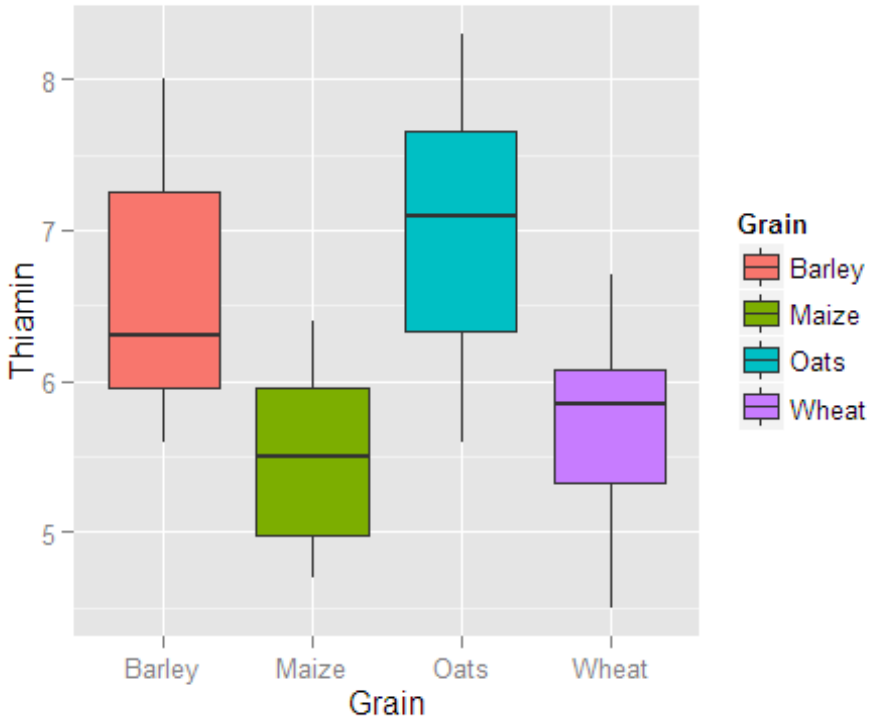

```r
dev.off()
```

```r
par(mfrow=c(1,2))
boxplot(resid(fit) ~ Grain, data = Cereal)
hist(resid(fit))
```




```r
dev.off()
```

```
library(ggplot2)
p <- ggplot(Cereal, aes(Grain, Thiamin))
p + geom_boxplot(aes(fill = Grain))
```

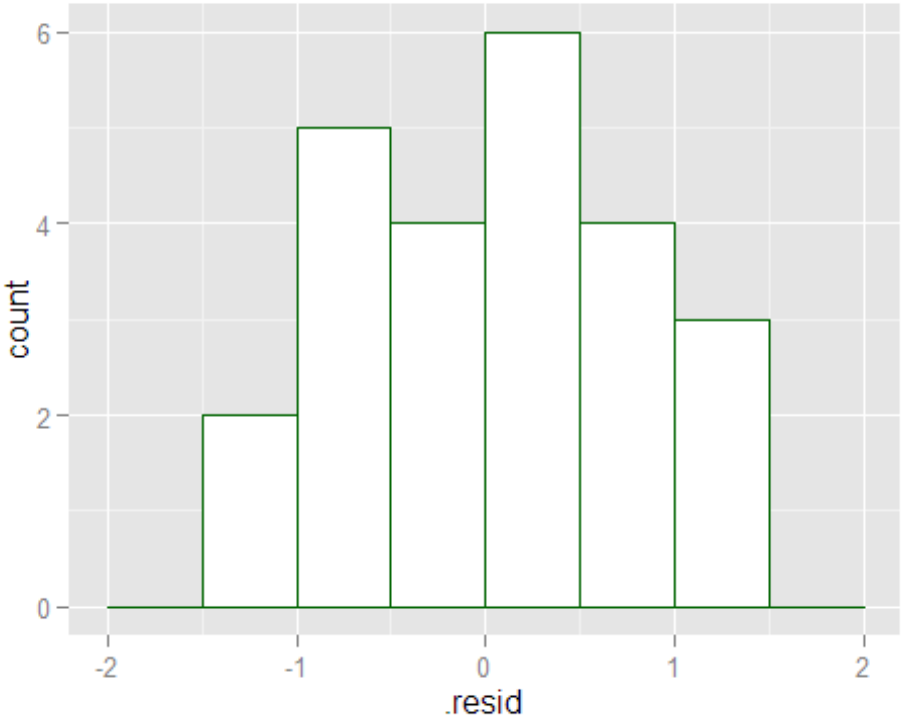



```
library(broom)

head(augment(fit))
```

```
##   Thiamin Grain .fitted   .se.fit        .resid      .hat    .sigma
## 1     5.2 Wheat     5.7 0.3516627 -5.000000e-01 0.1666667 0.8747932
## 2     4.5 Wheat     5.7 0.3516627 -1.200000e+00 0.1666667 0.8307257
## 3     6.0 Wheat     5.7 0.3516627  3.000000e-01 0.1666667 0.8805501
## 4     6.1 Wheat     5.7 0.3516627  4.000000e-01 0.1666667 0.8780361
## 5     6.7 Wheat     5.7 0.3516627  1.000000e+00 0.1666667 0.8472867
## 6     5.7 Wheat     5.7 0.3516627 -1.776357e-15 0.1666667 0.8837718
##         .cooksd .std.resid
## 1 2.021563e-02 -0.6358559
## 2 1.164420e-01 -1.5260541
## 3 7.277628e-03  0.3815135
## 4 1.293801e-02  0.5086847
## 5 8.086253e-02  1.2717117
## 6 2.551572e-31  0.0000000
```

```r
m <- ggplot(augment(fit), aes(x=.resid))
m + geom_histogram(colour = "darkgreen", fill = "white", binwidth = 0.5)
```




```r
shapiro.test(resid(fit))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit)
## W = 0.9704, p-value = 0.6775
```

```r
wheat <- c(rep(1,6),rep(0,18))
barley <- c(rep(0,6),rep(1,6),rep(0,12))
maize <- c(rep(0,12),rep(1,6),rep(0,6))
oats <- c(rep(0,18),rep(1,6))

thiamin <- Thiamin

cereal <- data.frame(wheat, barley, maize, oats, thiamin)
cereal
```

```
##    wheat barley maize oats thiamin
## 1      1      0     0    0     5.2
## 2      1      0     0    0     4.5
## 3      1      0     0    0     6.0
## 4      1      0     0    0     6.1
## 5      1      0     0    0     6.7
## 6      1      0     0    0     5.7
## 7      0      1     0    0     6.5
## 8      0      1     0    0     8.0
## 9      0      1     0    0     6.1
## 10     0      1     0    0     7.5
## 11     0      1     0    0     5.9
## 12     0      1     0    0     5.6
## 13     0      0     1    0     5.8
## 14     0      0     1    0     4.7
## 15     0      0     1    0     6.4
## 16     0      0     1    0     4.9
## 17     0      0     1    0     6.0
## 18     0      0     1    0     5.2
## 19     0      0     0    1     8.3
## 20     0      0     0    1     6.1
## 21     0      0     0    1     7.8
## 22     0      0     0    1     7.0
## 23     0      0     0    1     5.6
## 24     0      0     0    1     7.2
```

$$Y_i = \beta_0 + \beta_1 \texttt{maize}_i + \beta_2 \texttt{oats}_i + \beta_3 \texttt{wheat}_i + \varepsilon_i, \qquad i = 1, 2, \ldots, 24.$$

μ_{Barley}	β_0
μ_{Maize}	$\beta_0 + \beta_1$
μ_{Oats}	$\beta_0 + \beta_2$
μ_{Wheat}	$\beta_0 + \beta_3$

$$H_0: \mu_{\text{Barley}} = \mu_{\text{Maize}} = \mu_{\text{Oats}} = \mu_{\text{Wheat}}$$

$$\Leftrightarrow \qquad H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

```r
fit2 <- lm(thiamin ~ maize + oats + wheat, data = cereal)
summary(fit2)
```

```
##
## Call:
## lm(formula = thiamin ~ maize + oats + wheat, data = cereal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.400 -0.625  0.000  0.575  1.400
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.6000     0.3517  18.768 3.62e-14 ***
## maize        -1.1000     0.4973  -2.212   0.0388 *
## oats          0.4000     0.4973   0.804   0.4307
## wheat        -0.9000     0.4973  -1.810   0.0854 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8614 on 20 degrees of freedom
## Multiple R-squared:  0.3837, Adjusted R-squared:  0.2913
## F-statistic: 4.151 on 3 and 20 DF,  p-value: 0.01936
```

```r
anova(lm(thiamin ~ 1, data = cereal), fit2)
```

```
## Analysis of Variance Table
##
## Model 1: thiamin ~ 1
## Model 2: thiamin ~ maize + oats + wheat
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1     23 24.08
## 2     20 14.84  3      9.24 4.1509 0.01936 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
fit3 <- lm(thiamin ~ maize + oats + wheat + barley, data = cereal)
summary(fit3)
```

```
##
## Call:
## lm(formula = thiamin ~ maize + oats + wheat + barley, data = cereal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.400 -0.625  0.000  0.575  1.400
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.6000     0.3517  18.768 3.62e-14 ***
## maize        -1.1000     0.4973  -2.212   0.0388 *
## oats          0.4000     0.4973   0.804   0.4307
## wheat        -0.9000     0.4973  -1.810   0.0854 .
## barley            NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8614 on 20 degrees of freedom
## Multiple R-squared:  0.3837, Adjusted R-squared:  0.2913
## F-statistic: 4.151 on 3 and 20 DF,  p-value: 0.01936
```

```r
fit4 <- lm(thiamin ~ maize + oats + wheat + barley + 0, data = cereal)
summary(fit4)
```

```
##
## Call:
## lm(formula = thiamin ~ maize + oats + wheat + barley + 0, data = cereal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.400 -0.625  0.000  0.575  1.400
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## maize    5.5000     0.3517   15.64 1.11e-12 ***
## oats     7.0000     0.3517   19.91 1.18e-14 ***
## wheat    5.7000     0.3517   16.21 5.72e-13 ***
## barley   6.6000     0.3517   18.77 3.62e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8614 on 20 degrees of freedom
```

```
## Multiple R-squared:  0.9843, Adjusted R-squared:  0.9812
## F-statistic: 313.9 on 4 and 20 DF,  p-value: < 2.2e-16
```

```
grain <- c(rep(1,6),rep(2,6),rep(3,6),rep(4,6))
cereal2 <- data.frame(grain,thiamin);cereal2
```

```
##    grain thiamin
## 1      1     5.2
## 2      1     4.5
## 3      1     6.0
## 4      1     6.1
## 5      1     6.7
## 6      1     5.7
## 7      2     6.5
## 8      2     8.0
## 9      2     6.1
## 10     2     7.5
## 11     2     5.9
## 12     2     5.6
## 13     3     5.8
## 14     3     4.7
## 15     3     6.4
## 16     3     4.9
## 17     3     6.0
## 18     3     5.2
## 19     4     8.3
## 20     4     6.1
## 21     4     7.8
## 22     4     7.0
## 23     4     5.6
## 24     4     7.2
```

```
lm(thiamin ~ grain, data = cereal2)
```

```
##
## Call:
## lm(formula = thiamin ~ grain, data = cereal2)
##
## Coefficients:
## (Intercept)        grain
##        5.50         0.28
```

```
lm(thiamin ~ factor(grain), data = cereal2)
```

```
##
## Call:
## lm(formula = thiamin ~ factor(grain), data = cereal2)
##
## Coefficients:
##    (Intercept)  factor(grain)2  factor(grain)3  factor(grain)4
##            5.7             0.9            -0.2             1.3
```

```r
is.factor(grain)
```

```
## [1] FALSE
```