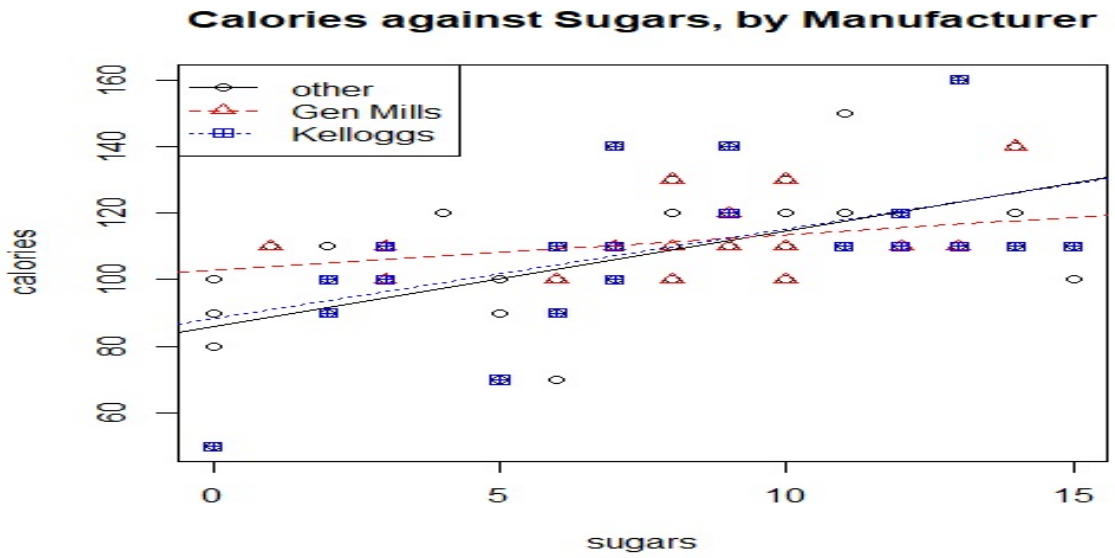
Stat 6021: Guided Question Set 6 Solutions

1. There appears to be a linear relationship between amount of sugar and calories for each of the manufacturers of the cereals. The slope for General Mills appears to be different from Kellogg's and other manufacturers, indicating interaction between amount of sugar and manufacturer. For a 1 unit increase in amount of sugar, the increase in calories for General Mills appears to be less than for Kellogg's and the other manufacturers.



2. The estimated regression equation is

$\hat{calories} = 85.9275 + 2.8666 sugars + 17.0672 GeneralMills + 2.2794 Kelloggs - 1.8145 sugars \times GeneralMills - 0.1583 sugars \times Kelloggs$

```
Call:
lm(formula = calories ~ sugars * mfr)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  85.9275     4.6691  18.404  < 2e-16 ***
```

```
sugars        2.8666     0.6367   4.502 2.57e-05 ***
mfrG         17.0672     9.3197   1.831   0.0712 .
mfrK          2.2794     8.1973   0.278   0.7818
sugars:mfrG  -1.8145     1.1153  -1.627   0.1082
sugars:mfrK  -0.1583     0.9990  -0.158   0.8745
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.25 on 71 degrees of freedom
Multiple R-squared:  0.3502,    Adjusted R-squared:  0.3044
F-statistic: 7.652 on 5 and 71 DF,  p-value: 8.471e-06
```

Note: I set other manufacturers as the reference class.

3. $H_0 : \beta_4 = \beta_5 = 0, H_a$: at least one of the coefficients in H_0 is nonzero.

   The Partial F statistic is 1.4318 with a p-value of 0.2457, which is greater than 0.05. We fail to reject the null hypothesis. This informs us we should go with the simpler model without interactions.

```
Analysis of Variance Table

Model 1: calories ~ sugars + mfr
Model 2: calories ~ sugars * mfr
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     73 19505
2     71 18749  2    756.21 1.4318 0.2457
```

   The estimated regression equation without interactions is

   $\hat{calories} = 88.556 + 2.412sugars + 3.622GeneralMills + 1.893Kelloggs.$

   Breaking this equation down for each manufacturer, we have:

   - For General Mills:

$$\begin{aligned} \hat{calories} &= 88.556 + 2.412sugars + 3.622GeneralMills + 1.893Kelloggs \\ &= 88.556 + 2.412sugars + 3.622 \times 1 + 1.893 \times 0 \\ &= 92.178 + 2.412sugars \end{aligned}$$

   - For Kelloggs:

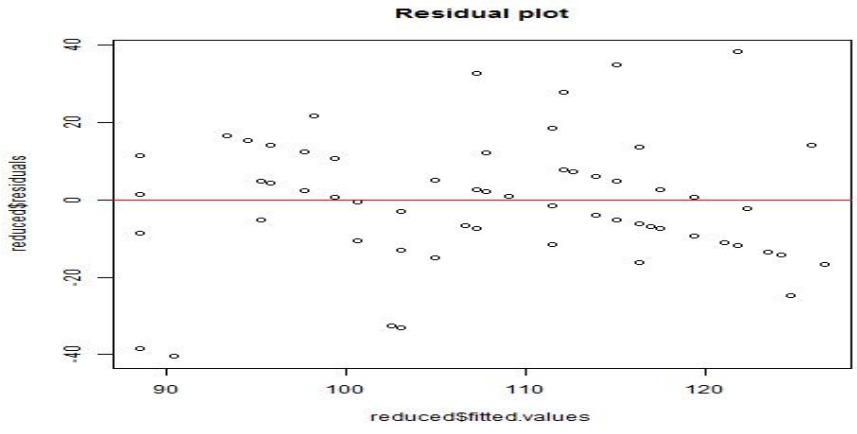
$$\begin{aligned} \hat{calories} &= 88.556 + 2.412sugars + 3.622GeneralMills + 1.893Kelloggs \\ &= 88.556 + 2.412sugars + 3.622 \times 0 + 1.893 \times 1 \\ &= 90.449 + 2.412sugars \end{aligned}$$

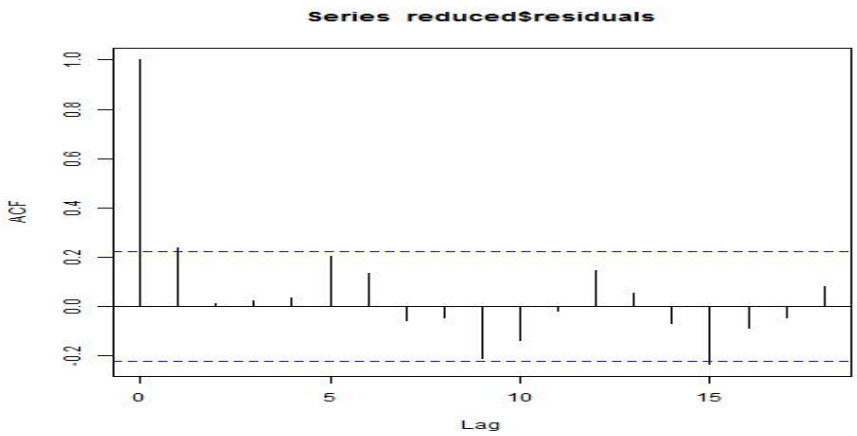
- For other manufacturers:

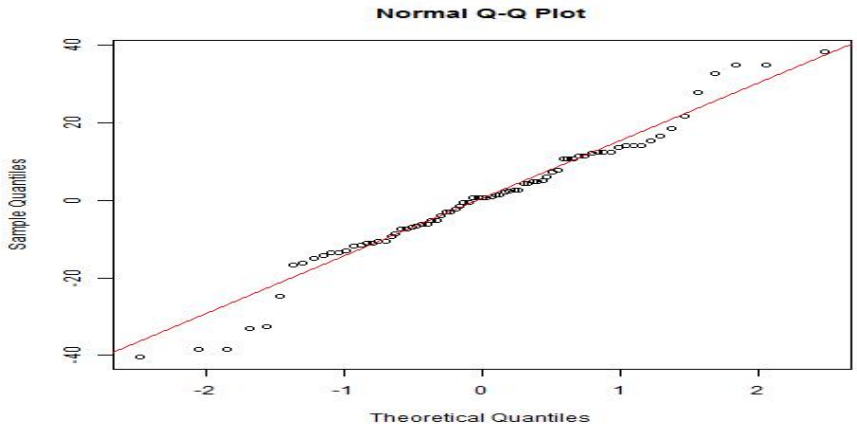
$$\begin{aligned} \hat{calories} &= 88.556 + 2.412sugars + 3.622GeneralMills + 1.893Kelloggs \\ &= 88.556 + 2.412sugars + 3.622 \times 0 + 1.893 \times 0 \\ &= 88.556 + 2.412sugars \end{aligned}$$

```
Call:
lm(formula = calories ~ sugars + mfr)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   88.556      3.831  23.117  < 2e-16 ***
sugars         2.412      0.435   5.545 4.45e-07 ***
mfrG           3.622      4.625   0.783    0.436
mfrK           1.893      4.535   0.417    0.678
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.35 on 73 degrees of freedom
Multiple R-squared:  0.3239,    Adjusted R-squared:  0.2962
F-statistic: 11.66 on 3 and 73 DF,  p-value: 2.507e-06
```

4. Based on the residual plot, the residuals are evenly scattered around 0 with no apparent pattern. The vertical spread of the residuals also appears to be fairly constant. The ACF plot of the residuals indicates the residuals are generally not significantly correlated, although at lag 1 it appears to have a small correlation. The QQ plot indicates the distribution of the residuals may be more heavily tailed than the normal distribution. Since we have a categorical variable, we also need to check if the variance of the response variable is equal among all classes of manufacturers. From Levene's test, the null hypothesis is not rejected, so the assumption of equal variances across all classes is reasonable.








```
> levene.test(calories,mfr)
data:  calories
Test Statistic = 2.6938, p-value = 0.07425
```

5. We note that the p-values for all pairwise comparisons are insignificant. This informs us that the differences in mean calories for all possible pairs of manufacturers are all statistically insignificant when controlling for amount of sugars, i.e. the difference in mean calories between General Mills and other manufacturers is statistically insignificant when controlling for amount of sugars; the difference in mean calories between Kellogg's and other manufacturers is statistically insignificant when controlling for amount of sugars; the difference in mean calories between General Mills and Kellogg's is statistically insignificant when controlling for amount of sugars.

```
> library(multcomp)
> pairwise<-glht(reduced, linfct = mcp(mfr= "Tukey"))
> summary(pairwise)

Fit: lm(formula = calories ~ sugars + mfr)
```

```
Linear Hypotheses:
               Estimate Std. Error t value Pr(>|t|)
G - other == 0    3.622      4.625   0.783    0.714
K - other == 0    1.893      4.535   0.417    0.908
K - G == 0       -1.729      4.878  -0.354    0.933
```