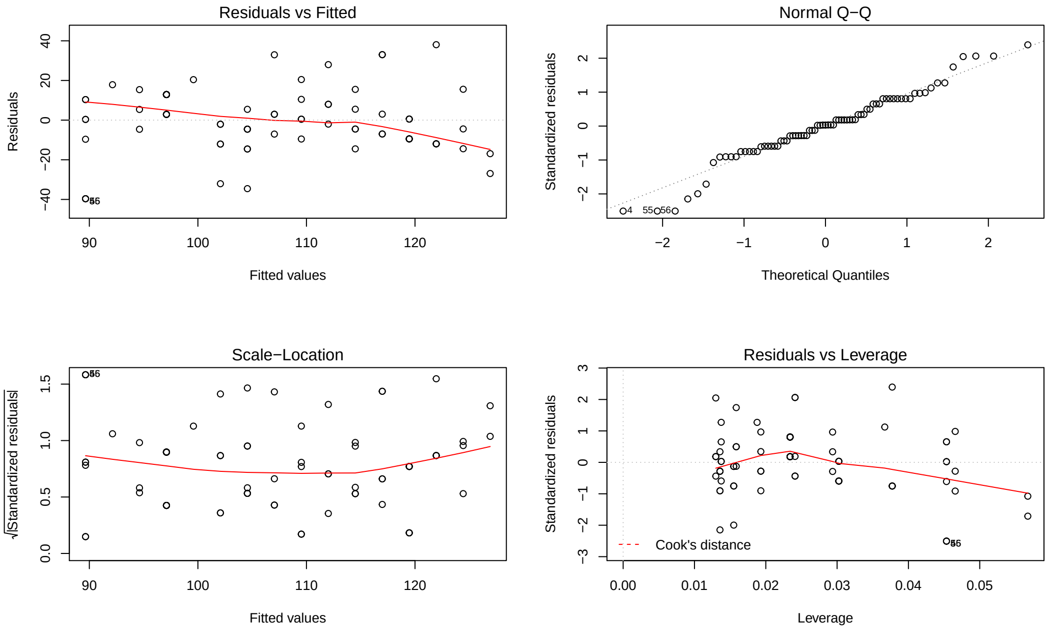
STAT 217: SLR Assumptions Practice (in class 3/30)

A study was done about nutritional and marketing information on US cereals. Data on the calories and sugar per serving (one cup) of 77 randomly selected US cereals was collected. Use the residual plots below to assess the simple linear regression model assumptions.

```
cereal.fit <- lm(calories~sugar, data = cereal)
summary(cereal.fit)

##
## Call:
## lm(formula = calories ~ sugar, data = cereal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -39.65  -9.47   0.47  10.47  38.05
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    89.65       3.45   26.00  < 2e-16 ***
## sugar           2.48       0.42    5.92  9.2e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.2 on 75 degrees of freedom
## Multiple R-squared:  0.318,Adjusted R-squared:  0.309
## F-statistic:   35 on 1 and 75 DF,  p-value: 9.17e-08
```

```
par(mfrow=c(2,2))
plot(cereal.fit)
```

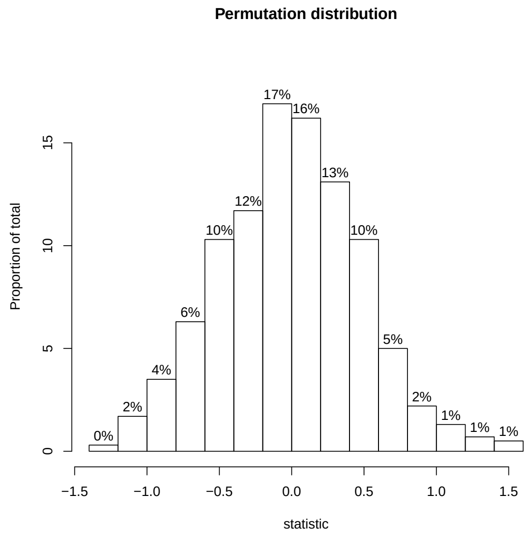


Let's consider using a permutation test to test the hypotheses $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$. The following gives you some general steps for conducting a permutation test. Answer the bulleted questions about how to conduct a permutation test in this context.

```
b1obs <- cereal.fit$coef[2]

B <- 1000
slope <- matrix(NA, nrow=B)
for(b in (1:B)){
  slope[b]<-lm(calories~shuffle(sugar), data=cereal)$coef[2]
}
```

1. Simulate 1000 new datasets under the assumption that the null hypothesis is true.
   - Explain specifically how this is done. Use the R-code above to explain your answer.
2. With each of the 1000 datasets, calculate the statistic of interest.
   - What is the statistic of interest in this example?
3. You should now have 1000 statistics, one from each of the 1000 samples. Make a histogram of those statistics.



   - What was the observed statistic from the original sample?
   - Draw a vertical line at the observed statistic on the plot above.
4. The p-value is found by calculating the proportion of permutation statistics (generated under the assumption that the null hypothesis is true) that were as or more extreme than the observed statistic.
   - Estimate the p-value. Explain your answer.
   - Is this a two-tailed test or a one-tailed test?
5. Write a conclusion.