Using the Ericksen dataset, we want to predict the percentage by which the population of a region is undercounted (undercount).

1. Which of these three models is most useful for predicting undercount and why?

   modm: undercount ~ minority
   modp: undercount ~ poverty
   modc: undercount ~ crime

2. Sequential sums of squares.

   modmp = lm(undercount ~ minority + poverty, Ericksen)
   modpm = lm(undercount ~ poverty + minority, Ericksen)
   anova(modmp)

   Analysis of Variance Table
   Response: undercount
   Df  Sum Sq  Mean Sq    F value      Pr(>F)
   minority 1 195.80   195.80   65.331 2.61e-11
   poverty  1  12.10    12.10    4.048  0.048
   Residuals 63 188.80    3.00

   anova(modpm)

   Analysis of Variance Table
   Response: undercount
   Df  Sum Sq  Mean Sq    F value      Pr(>F)
   poverty  1   63.69    63.69   21.335 2.00e-05
   minority 1 144.21   144.21   48.106 2.60e-09
   Residuals 63 188.80    3.00

(a) What’s different between the two models?

(b) What’s the same?

3. Adding crime: undercount ~ minority + poverty + crime
Problem due Tuesday, April 16

The dataframe `Baseball21` has data on all baseball team seasons in the 21st century. The number of runs scored by a baseball team in a year \((R)\) is obviously related to the things it does on offense. The following variables are measures of the offense of a baseball team:

- \(H\) hits
- \(X2B\) doubles
- \(X3B\) triples
- \(HR\) home runs
- \(B\) walks
- \(SO\) strike outs
- \(SB\) stolen bases
- \(CS\) caught stealing

1. Which variable is most useful in explaining the number the of runs a team scores in a year?

2. Which second variable would you add to the model and why?

3. The model that explains runs from all of these variables is too big. Fit that model and suggest at three variables to remove with good reasons for that.