Name: _____

Consider a multiple regression model on acorn seeds:

 $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2$ $Y = \ln \text{Range}$ $X_1 = \ln \text{Acorn Size}$ $X_2 = \text{Location: 1=Atlantic, 0=California}$ $X_3 = \text{Size of the Tree}$

By fitting two different linear models we get the following ANOVA tables:

```
1. > anova(lm(ln.range ~ ln.size* location + Tree_Height))
  Analysis of Variance Table
  Response: ln.range
                   Df Sum Sq Mean Sq F value
                                                Pr(>F)
  ln.size
                    1 0.715
                               0.715 0.7324
                                               0.39810
  location
                    1 73.288 73.288 75.0595 4.015e-10 ***
  Tree_Height
                    1 2.004
                               2.004 2.0524
                                               0.16109
  ln.size:location 1 4.872
                               4.872 4.9902
                                               0.03218 *
                   34 33.197
  Residuals
                               0.976
2. > anova(lm(ln.range ~ ln.size* location))
  Analysis of Variance Table
  Response: ln.range
                   Df Sum Sq Mean Sq F value
                                                Pr(>F)
  ln.size
                    1 0.715
                               0.715 0.6939
                                               0.41049
                    1 73.288 73.288 71.1147 6.001e-10 ***
  location
  ln.size:location 1 4.004
                               4.004 3.8856
                                               0.05665 .
  Residuals
                   35 36.070
                               1.031
```

According to $R_{a,p}^2$, C_p , and SBC_p , which model (1. or 2.) seems better? Which one would you use?

Solution:

$$\begin{aligned} R_{a,p}^2 &= 1 - \frac{MSE_p}{SSTO/n - 1} \\ C_p &= \frac{SSE_p}{MSE(F)} - (n - 2p) \\ SBC_p &= n\ln SSE_p - n\ln n + (\ln n)p \end{aligned}$$

1.

$$R_{a,p}^{2} = 1 - \frac{0.976}{114.077/38} = 0.675 \quad \checkmark$$
$$C_{p} = \frac{33.197}{0.976} - (39 - 2 \cdot 5) = 5 \quad \checkmark$$
$$SBC_{p} = 39 \ln 33.197 - 39 \ln 39 + (\ln 39)5 = 12.03$$

2.

$$\begin{aligned} R_{a,p}^2 &= 1 - \frac{1.031}{114.077/38} = 0.657 \\ C_p &= \frac{36.070}{0.976} - (39 - 2 \cdot 4) = 5.96 \\ SBC_p &= 39 \ln 36.07 - 39 \ln 39 + (\ln 39)4 = 11.61 \quad \checkmark \end{aligned}$$

 $R_{a,p}^2$ and C_p choose model 1, SBC_p chooses model 2. I'd probably use model 2. Doesn't seem to me that tree height is adding much to the model even though the criteria become optimized. Note the p-value on tree height in model 1. The p-value addresses the hypothesis that β_3 is zero given the size and location of the acorn are in the model (without the interaction!). So, tree height isn't significant given those first two main effects. It's possible that the interaction will make tree height significant, but that doesn't make much sense given what we know about the variables.