WU #13 - non-nested choices

Math 158 - Jo Hardin

Thursday 3/10/2022

Name: _____

Names of people you worked with: _____

Consider the regression model handouts concerning the birth weight data. The task below is to compare two un-nested models. That is, you have to decide if you'd prefer to have smoking `habit` in the model or `term` (the length of the pregnancy, as a categorical variable).

According to $R^2_{a,p}$, C_p, and BIC_p, which model (Model 1 or Model 2?) seems better? Which one would you use?

Model 1 is:

$$E[Y] = \beta_0 + \beta_1 gained + \beta_2 mage + \beta_4 term_{full} + \beta_5 term_{late}$$

Model 2 is:

$$E[Y] = \beta_0 + \beta_1 gained + \beta_2 mage + \beta_3 habit$$

You might need the following output (in addition to the output from the larger handout):

```
anova(lm(weight ~ gained + habit + mage + term, data = births14))
```

```
## Analysis of Variance Table
##
## Response: weight
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## gained      1   33.86  33.860 25.7373 4.718e-07 ***
## habit       1   25.30  25.299 19.2295 1.291e-05 ***
## mage        1    6.42   6.417  4.8774   0.02745 *
## term        2  223.27 111.637 84.8555 < 2.2e-16 ***
## Residuals 935 1230.10   1.316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Solution: Notice that there is no hypothesis test here!! They are *not* nested models. But we can still consider which one might be better.

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

Model 1

$$\begin{aligned} R^2_{adj,p} &= 1 - \frac{1.333}{(33.86+228.38+9.11+1247.6)/940} = 1 - \frac{1.333}{1518.9/940} = 0.825 \\ C_p &= \frac{1247.6}{1230.10/935} - (941 - 2\cdot 5) = 17.303 \quad \surd \\ BIC_p &= 941\cdot\ln(1247.6) - 941\cdot\ln(941) + \ln(941)\cdot 5 = 299.63 \quad \surd \end{aligned}$$

Model 2

$$\begin{aligned} R^2_{adj,p} &= 1 - \frac{1.551}{(33.86+25.30+6.42+1453.37)/940} = 1 - \frac{1.551}{1518.9/940} = 0.96 \quad \surd \\ C_p &= \frac{1453.37}{1230.10/935} - (941 - 2\cdot 4) = 171.71 \\ BIC_p &= 941\cdot\ln(1453.37) - 941\cdot\ln(941) + \ln(941)\cdot 4 = 436.44 \end{aligned}$$

C_p and BIC_p choose model 1, $R^2_{adj,p}$ chooses model 2. I'd probably use model 1. It seems as though `term` is generally more significant than `habit` with a smaller MSE. There is no right answer!