WU #14 - Model Selection

Math 158 - Jo Hardin

Tuesday 3/22/2022

Name: ____

Names of people you worked with: ____

Consider the regression model handouts concerning the birth weight data.

Write down two versions of the same model:

1. The population model representing the variables which you've selected to use in the final model.
2. The sample model representing the same variables (which you've selected to use in the final model).

```
leaps::regsubsets(weight ~ mage + mature + weeks + premie + gained + lowbirthweight +
                    habit + marital,
                  data = births14,
                  nvmax = 6) %>%
  tidy()
```

p	(Intercept)	mage	mature	weeks	premie	gained	lowbirthweight	habit	marital
2	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
3	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
4	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE
5	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE
6	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE
7	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE

p	r.squared	adj.r.squared	BIC	mallows_cp
2	0.4143466	0.4137229	-489.7666	177.888962
3	0.4844974	0.4833983	-602.9780	46.345096
4	0.4990304	0.4974264	-623.0407	20.679185
5	0.5072677	0.5051620	-631.7950	6.998072
6	0.5087428	0.5061158	-627.7694	6.189978
7	0.5093806	0.5062289	-622.1450	6.975775

Solution:

There isn't a single right answer. Always remember, modeling is an art. Seems like maybe a model with 4 or 5 ($p = 5$ or 6) variables will be a good balance of information and simplicity. I'll choose the four variable ($p = 5$) model (seems to be the biggest jump in information).

1. The population model:

$$E[\texttt{weight}] = \beta_0 + \beta_1\texttt{mage} + \beta_2\texttt{weeks} + \beta_3\texttt{gained} + \beta_4\texttt{lowbirthweight}$$

2. The sample model:

$$\widehat{\texttt{weight}} = b_0 + b_1\texttt{mage} + b_2\texttt{weeks} + b_3\texttt{gained} + b_4\texttt{lowbirthweight}$$

Which can also be written as (after running the model in R)

$$\widehat{\texttt{weight}} = -1.67 + 0.02 \cdot \texttt{mage} + 0.15 \cdot \texttt{weeks} + 0.01 \cdot \texttt{gained} + 2.42 \cdot \texttt{lowbirthweight}$$

```r
lm(weight ~ mage + weeks + gained + lowbirthweight, data = births14) %>%
  tidy()
```

```
## # A tibble: 5 x 5
##   term                  estimate std.error statistic  p.value
##   <chr>                    <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)            -1.67     0.493       -3.38 7.49e- 4
## 2 mage                    0.0201   0.00507      3.96 8.21e- 5
## 3 weeks                   0.149    0.0131      11.4  4.44e-28
## 4 gained                  0.0103   0.00191      5.43 7.29e- 8
## 5 lowbirthweightnot low   2.42     0.121       20.0  2.50e-74
```