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  +
                    sex + habit + marital + whitemom,
                  data = births14,
                  nvmax = 8) %>%
  tidy()
```

p	(Intercept)	mage	mature	weeks	premie	gained	sex	habit	marital	white
2	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
4	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
5	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
6	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
7	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE
8	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
9	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

p	r.squared	adj.r.squared	BIC	mallows_cp
2	0.2747745	0.2740022	-288.6237	117.185365
3	0.2995942	0.2981008	-314.5450	83.107467
4	0.3181139	0.3159307	-332.9142	58.187380
5	0.3353819	0.3325417	-350.2040	35.086601
6	0.3439540	0.3404457	-355.5728	24.626287
7	0.3503868	0.3462137	-357.9982	17.275620
8	0.3574596	0.3526388	-361.4528	8.994654
9	0.3587717	0.3532676	-356.5294	9.087370

Solution:

There isn't a single right answer!! Always remember, modeling is an art. Seems like any of the best models with at least 4 ($p \geq 5$) variables will be a good balance of information and simplicity. I'll choose the five variable ($p = 6$) model (seems to be the biggest jump in information).

1. The population model:

$$E[\texttt{weight}] = \beta_0 + \beta_1\texttt{mage} + \beta_2\texttt{weeks} + \beta_3\texttt{premie} + \beta_4\texttt{gained} + \beta_5\texttt{sex}$$

2. The sample model:

$$\widehat{\texttt{weight}} = b_0 + b_1\texttt{mage} + b_2\texttt{weeks} + b_3\texttt{premie} + b_4\texttt{gained} + b_5\texttt{sex}$$

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

$$\widehat{\texttt{weight}} = -0.97 + 0.02 \cdot \texttt{mage} + 0.18 \cdot \texttt{weeks} - 0.84 \cdot \texttt{premie} + 0.01 \cdot \texttt{gained} + 0.42 \cdot \texttt{sex}$$

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

```
## # A tibble: 6 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   -0.970    0.798       -1.22 2.24e- 1
## 2 mage           0.0205   0.00587      3.50 4.96e- 4
## 3 weeks          0.185    0.0200       9.23 1.82e-19
## 4 premiepremie  -0.846    0.151       -5.61 2.70e- 8
## 5 gained         0.0113   0.00220      5.12 3.62e- 7
## 6 sexmale        0.422    0.0675       6.25 6.26e-10
```

Note Two things...

- Check out that `mage` isn't included, then is included, then isn't.
- The "best" models are found by comparing every single possible model with a particular value of p and choosing the one with the smalles SSE.