STAT 213: R^2, Adjusted R^2, and Parsimony

Colin Reimer Dawson

Last Revised March 10, 2016

Let's do an experiment to examine what happens when we build a sequence of increasingly complex regression models, where some of the predictors have nothing to do with the response, in reality.

First, let's create a fake dataset with ten predictor variables, each with random values generated from a standard Normal distribution.

Enter the following R code to create a dataset with these properties property. If you are running this in a script or a Markdown document, first type `set.seed(SOME_NUMBER)` so that you can get the same results every time you re-run or re-Knit your code.

```
library("mosaic")
n <- 100; k <- 10
Predictors <- do(k) * rnorm(n, mean = 0, sd = 1)
```

Use `head()` to see what this looks like.

Now, let's generate a response variable from a known population model, with some random, Normally distributed residuals. Pick four numbers in the range $[-10, 10]$ for coefficients. Type the following, but replace the "`SOMETHING`"s with the coefficients that you chose.

```
beta_0 = 0; beta_1 = SOMETHING1; beta_2 = SOMETHING2;
beta_3 = SOMETHING_3; beta_4 = SOMETHING_4
FakeData <-
    mutate(Predictors,
           Yhat = beta_0 + beta_1 * V1 + beta_2 * V2 + beta_3 * V3 + beta_4 * V4,
           epsilon = rnorm(n, mean = 0, sd = 0.5),
           Y = Yhat + epsilon)
```

We are defining `Yhat` in terms of the first four predictors, generating independent random Normal residuals, and then defining `Y` as `Yhat` plus the random residuals.

Now, fit a series of ten multiple regression models to this data:

1. $Y = \hat{\beta}_0 + \hat{\beta}_1 V1$
2. $\texttt{Y} = \hat{\beta}_0 + \hat{\beta}_1\texttt{V1} + \hat{\beta}_2\texttt{V2}$

...

9. $\texttt{Y} = \hat{\beta}_0 + \hat{\beta}_1\texttt{V1} + \hat{\beta}_2\texttt{V2} + \hat{\beta}_3\texttt{V3} + \hat{\beta}_4\texttt{V4} + \hat{\beta}_4\texttt{V4} + \hat{\beta}_5\texttt{V5} + \hat{\beta}_6\texttt{V6} + \hat{\beta}_7\texttt{V7} + \hat{\beta}_8\texttt{V8} + \hat{\beta}_9\texttt{V9}$
10. $\texttt{Y} = \hat{\beta}_0 + \hat{\beta}_1\texttt{V1} + \hat{\beta}_2\texttt{V2} + \hat{\beta}_3\texttt{V3} + \hat{\beta}_4\texttt{V4} + \hat{\beta}_4\texttt{V4} + \hat{\beta}_5\texttt{V5} + \hat{\beta}_6\texttt{V6} + \hat{\beta}_7\texttt{V7} + \hat{\beta}_8\texttt{V8} + \hat{\beta}_9\texttt{V9} + \hat{\beta}_{10}\texttt{V10}$

Notice that for the first four models, we are adding predictors that are actually related to the response, according to our known population model. For models 5 through 10, there is no connection between the predictors we are adding and the response.

For your ten models, find the R^2 and the adjusted R^2, and plot them both below as a function of the number of predictors.