STAT 213: Correlated Predictors in MLR

Colin Reimer Dawson

Last Revised April 4, 2016

The goal of this lab is to explore what happens when we fit a multiple regression model with predictors that are highly correlated with each other.

1. Read the dataset at `http://colinreimerdawson.com/data/TestScores.csv` into R. It contains simulated test data for a `Quiz`, a `Midterm` and a `Final`.

2. Fit two simple linear regression models to predict the final exam score from each of the other two variables. Interpret the t-tests for the slopes and get the corresponding 95% confidence intervals using `confint(MODEL)`, where `MODEL` is whatever you called the corresponding model. Do the other grades predict the final exam score well? Which one does a better job?

3. Now fit the multiple regression model with both variables as predictors. Interpret the t-tests of the individual coefficients.

4. Get 95% confidence intervals for the two predictors using `confint(MODEL)` (where `MODEL` is whatever you called your MLR model). How do these differ from the intervals you got from the SLR models?

5. Construct scatterplots of each pair of variables by doing `plot(DATA)`, where `DATA` is whatever you called the data frame. What jumps out at you? Does this explain the stark contrast between the MLR model and the two SLR models?

6. We can get a **confidence ellipse** for the two coefficients together: Since the predictors are positively correlated, if one coefficient goes up, the other one should go down to compensate. The confidence ellipse reflects this. Make sure you have the `mosaic` package loaded, and do `confidenceEllipse(MODEL, levels = c(0.95, 0.99))` to plot 95% and 99% confidence ellipses, where `MODEL` is again the name of your MLR model.

7. Type the following to get the coefficients of the lines that make up the axes of the ellipse (replace `DATA` and `MODEL`)

```
select(DATA, Midterm, Quiz) %>% cov() %>% eigen()
```

   You should see a 2×2 matrix at the bottom. If this matrix is

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

   then the axes of the ellipse have slopes C/A and D/B, respectively.

8. Use `mutate()` to create two new variables that are combinations of `Midterm` and `Quiz` as follows:

$$\begin{aligned} \texttt{V1} &= \mathrm{A} \cdot \texttt{Midterm} + C \cdot \texttt{Quiz} \\ \texttt{V2} &= \mathrm{B} \cdot \texttt{Midterm} + D \cdot \texttt{Quiz} \end{aligned}$$

9. Repeat Step 5 with the augmented data. What do you notice about the relationship between the two new variables?

10. Fit a new model with $V1$ and $V2$ as predictors. Compare the R^2 values and the t-tests to those from the previous model. What do you notice?

11. Plot a confidence ellipse for the coefficients of this new model as in Step 6. What do you notice?