Module 5: Sums of Squares and Multicollinearity

Jeffrey Woo

MSDS, University of Virginia

Hypothesis testing in MLR

So far we have seen

- t test: can we drop a predictor from the model while leaving the other predictors in the model?
- ANOVA F test: is our model useful in predicting the response variable?

Notice neither of these tests allow us to assess if we can drop a subset of predictors simultaneously.

NFL Example

From 1976 season (anyone knows what is special with this season?)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.292e+00  1.281e+01  -0.569 0.576312
x1           8.124e-04  2.006e-03   0.405 0.690329
x2           3.631e-03  8.410e-04   4.318 0.000414 ***
x3           1.222e-01  2.590e-01   0.472 0.642750
x4           3.189e-02  4.160e-02   0.767 0.453289
x5           1.511e-05  4.684e-02   0.000 0.999746
x6           1.590e-03  3.248e-03   0.490 0.630338
x7           1.544e-01  1.521e-01   1.015 0.323547
x8          -3.895e-03  2.052e-03  -1.898 0.073793 .
x9          -1.791e-03  1.417e-03  -1.264 0.222490
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.83 on 18 degrees of freedom
Multiple R-squared:  0.8156,    Adjusted R-squared:  0.7234
F-statistic: 8.846 on 9 and 18 DF,  p-value: 5.303e-05
```

The t tests do not inform us that all the predictors, except x_2, can be dropped from the model.

Partial F Test

The partial F test allows us to assess if multiple predictors can be dropped simultaneously from the model. The partial F statistic measures the change in the SS_R (or SS_{res}) with the removal of these predictors from the model.

Sum of Squares

- As long as we have the same response variable, SS_T is constant, regardless of the number and form of predictors used.
- $SS_T = SS_R + SS_{Res}$
- Each time predictors are added to the model, the SS_R increases and the SS_{Res} decreases by the same amount, since SS_T stays constant.

Partial F Test

Goal: is the increase in SS_R significant with the addition of predictor(s)?

Issues with Multicollinearity

When predictors are nearly linear dependent on each other. Issues:

- High variance with estimated coefficients: the estimated coefficient may be very different from the true value.
  - Caution with interpreting estimated coefficients in the usual manner.
  - Estimated coefficients tend to be large.
  - Algebraic sign of coefficients different than what is known theoretically.
  - Adding or removal of one or more data points results in large changes in the estimated regression coefficients.
- Predictions are fine but must be very careful with extrapolation.

Detecting Multicollinearity

- Insignificant t tests for predictors that are known to be useful in predicting the response variable, and significant ANOVA F test.
- High VIFs (exceeds 10).
- High correlation between pairs of predictors.

NFL Example

From 1976 season.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.292e+00  1.281e+01  -0.569 0.576312
x1           8.124e-04  2.006e-03   0.405 0.690329
x2           3.631e-03  8.410e-04   4.318 0.000414 ***
x3           1.222e-01  2.590e-01   0.472 0.642750
x4           3.189e-02  4.160e-02   0.767 0.453289
x5           1.511e-05  4.684e-02   0.000 0.999746
x6           1.590e-03  3.248e-03   0.490 0.630338
x7           1.544e-01  1.521e-01   1.015 0.323547
x8          -3.895e-03  2.052e-03  -1.898 0.073793 .
x9          -1.791e-03  1.417e-03  -1.264 0.222490
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.83 on 18 degrees of freedom
Multiple R-squared:  0.8156,    Adjusted R-squared:  0.7234
F-statistic: 8.846 on 9 and 18 DF,  p-value: 5.303e-05
```

Some Solutions

- Use a subset of predictors (drop some of the predictors that are linearly dependent on each other).
- Dimension reduction methods (principal component analysis).
- Shrinkage methods (ridge regression, lasso).