Multicollinearity and Variance Inflation (Sec 3.10, Ch. 9)

Last Time: Prediction Intervals
- Prediction intervals (individual response) are still wider than confidence intervals (mean response)
- Expect inclusion of useful variables to shorten the intervals (decrease $s$)
  - Do need to watch for hidden extrapolation
  - Some software packages will flag observations with large $h_i$ values
- Standardizing variables is a very useful pre-processing tool
  - Slope = predicted number of SD change in $y$ with a 1 SD increase in $x$
  - More meaningful intercept
  - Can make comparing the magnitude of the slope coefficients more meaningful?

Example 1: The file houses.txt contains selling prices for 20 houses sold in 2008 in a small Midwestern town, as well as size of each house (square feet) and size of the lot (square feet) from Cannon et al.

(a) Run a multiple regression model to predict price from house size and lot size. What do you conclude from the overall F test? What do you conclude from the individual $F$ (or $t$) tests?

Definition: When two or more explanatory variables are highly correlated (called multicollinearity or collinear), this can lead to problems with the model. Some signs of multicollinearity include:
- Coefficients do not have the expected sign;
- Overall F test is significant but none of the individual $t$ tests are.
Some negative consequences of multicollinearity include:
- Coefficients can lose statistical significance and even change sign (especially smaller values);
- Becomes even more difficult to interpret the slopes, can’t really think about “holding all other variables” constant;
- Adding a variable actually increases the SE of other coefficients (loss of precision);
- Use of redundant independent variables wastes time and money on unnecessary data collection.

Detecting Multicollinearity Pairwise correlation coefficients and scatterplots are a good first step, but sometimes the linear dependence is across several variables, not just one. The relationship between the predictor variables can be judged by examining the variance inflation factor ($VIF$).

\[
VIF_i = \frac{1}{1 - R_i^2}
\]

Note: \[\text{var}(\hat{\beta}_i) = \left( \frac{s^2}{(n-1)s_i^2} \right) \left( \frac{1}{1 - R_i^2} \right)\]

where $R_i^2$ is the proportion of the variation in $x_i$ that is explained by the its linear relationship to other explanatory variables in the model. $VIF$ measures the increase in variability of the $i^{th}$ sample regression coefficient due to the linear association of $x_i$ with other predictor variables in the model.

- The VIF will equal 1 if the variable is not linearly related to the other variables.
- It is suggested that a VIF in excess of 5 or 10 or max(10, 1/(1 - $R^2$ (full))) is evidence that multicollinearity may be causing problems in estimation. Another sign is if average of the VIF values is considerably larger than 1.
Dealing with Multicollinearity

- Reduce the number of variables in the model (dropping, combining)...
- Design a good experiment (with orthogonal variables)! [Note: See references for discussion of how all this relates to eigenvectors as well.]
- Collect more data to ensure coverage of entire range of variables
- If goal is estimation and prediction, not horrible to keep them all in, just be careful in interpreting the estimated coefficients (“discount individual coefficients and $t$-tests”) and don’t extrapolate
- With polynomial regression, can standardize the variables first to help reduce collinearity
- Can use ridge regression to produce more stable estimates of the coefficients or principal components to study the structure in the design matrix.

Example 2: The amount of September Arctic sea ice has been examined to gauge climate change (e.g., Pollack, 2011). As far as we know, much of the Arctic Ocean has been covered in floating sea ice year round. Ice melts during the summer but stops in September. The average Arctic sea ice extent in millions of square kilometers for each September from 1979 until 2012 is in ArticSeaIce.txt. Scientists have tried to use such data to predict when there will no longer be artic sea ice.

(a) Fit a quadratic model to predict the sea ice extent from year and year$^2$. Are both the linear and quadratic terms significant? How do you interpret the signs of their coefficients?

(b) Does this model have any issues with multicollinearity? (See online instructions for Variance Inflation Factors) Why is that not surprising? [Hint: Plot year vs. year$^2$]

We could center the variables as before. One step better is to standardize the variables.

(c) Create a new variable standardizing the year variable (See online instructions) and then fit the quadratic model with std year and std year$^2$. Examine the variance inflation factors. Why does standardizing (think centering) help? What is the correlation coefficient of these new variables?

(d) How do we interpret the signs of the coefficients? If we could, how would we interpret the coefficient of standardized year?

(e) How does this differ from entering year and year$^2$ and asking the software to standardize the variables?

Practice Problem to be submitted by Tuesday, 9am

(f) Fit a cubic model using the standardized year variable. Overlay this fitted curve on your scatterplot. How does the behavior differ? What can a cubic model do that a quadratic model cannot? Does that seem relevant in this context? Does this improve the fit of the model? Is the cubic term statistically significant?