Stat 324 - Day 25
Penalized Regression (Ch. 9)

Last Time: Dealing with multicollinearity
- Model respecification, including principal components analysis (Stat 419)
- Stepwise regression, want to minimize error in model fit plus penalty for number of terms

<table>
<thead>
<tr>
<th></th>
<th>Model fit</th>
<th>Penalty for number of terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cp</td>
<td>SSE/n/MSE</td>
<td>2(p+1)</td>
</tr>
<tr>
<td>AIC</td>
<td>n ln(SSE/n)</td>
<td>2(p+1)</td>
</tr>
<tr>
<td>BIC</td>
<td>n ln(SSE/n)</td>
<td>ln(n)(p + 1)</td>
</tr>
</tbody>
</table>

"Mechanical model selection and modification procedure disguise the substantive implications of modeling decisions. Consequently, these methods generally cannot compensate for weaknesses in the data and are no substitute for judgement and thought." (see online notes)

Example 1: Reconsider the standardized workforce participation data set (cleaned) and least squares model. Another method that was originally developed to handle multicollinearity is now being used more and more for variable selection. It is especially helpful when you have a large number of variables. The general idea of penalized regression (aka shrinkage) is to prevent any individual regression coefficients from being too large. The resulting regression coefficients will be “biased” but we are hoping the increase in bias (↓ in R²) is outweighed by the reduction in variability.

**Bias vs. variability tradeoff**: 
$$\text{MSE} (\hat{\beta}) = E[(\hat{\beta} - \beta)^2] = \text{Var}(\hat{\beta}) + \text{[E}(\hat{\beta} - \beta)]^2.$$

Ridge regression (Hoerl and Kennard, 1970) develops biased estimators by instead of minimizing 
$$\text{SSE} = \sum (y_i - \hat{\beta}_0 - \sum \hat{\beta}_j x_{ij})^2$$
we minimize
$$\sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 = \text{SSE} + \lambda \sum_{j=1}^p \beta_j^2$$

(a) What happens when \( \lambda \) is small? What happens to the slope coefficients as \( \lambda \) gets larger?

A key to this method is finding a value for \( \lambda \). A visual exploration of the role of \( \lambda \) is a ridge trace or trace plot. This plot examines the behavior of the estimated slope coefficients as a function of \( \lambda \). The goal is to find the smallest value of \( \lambda \) so the coefficients are stable, the residual sum of squares is small, and the VIF values are all less than 10.

(b) Create a trace plot (*see online instructions*). Which variables appear to have reasonably large coefficients?

(c) What value is reported for the estimate of \( \lambda \) (*scale* in JMP)?

(d) How do the fitted coefficients (for original predictors) differ from the least squares estimates?
Notes:  
- Ridge regression is helpful for identifying situations with collinear variables (lots of fluctuations when $\lambda$ is small). Can use ridge regression to identify the problem, and if removing variables reduces collinearity then use regular least squares regression on the subset of variables.  
- Ridge regression aims to produce estimates of the coefficients that have less variance than the least squares estimates (so superior in terms of MSE). Forecasts tend to be more accurate.  
- Ridge regression can be run when $p > n$ (and we can’t determine least squares estimates).  
- Ridge regression has a computational advantage over searching through $2^p$ coefficients in Best Subsets.  
- Do need to watch for influence of outliers on estimation of $\lambda$.  
- When $\lambda$ ends up being small, we might as well use least squares estimation.  
- Have to be pretty cautious interpreting p-values, confidence intervals.  

Ridge regression is an example of a broader class of shrinkage methods. A relatively recent method is the lasso (“least absolute shrinkage and selection operator”) which services to find regression coefficients that minimize

$$
\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \text{SSE} + \lambda \sum_{j=1}^{p} |\beta_j|
$$

(e) What is the key difference between the lasso and ridge regression?

As with ridge regression the lasso shrinks the coefficient estimates towards zero. However, whereas ridge regression can never set any coefficients to zero, and will also include all variables in a model it fits, the lasso can force some coefficient estimates to be exactly zero.

(f) Use the lasso with the participation dataset. How does the behavior of this trace plot differ from the previous? Which variables are kept in the model?

Notes
- Another variation is the elastic net which applies both the $\ell_1$ and $\ell_2$ penalties. The $\ell_1$ penalty enables variable selection and the $\ell_2$ penalty improves predictive ability through shrinkage.  
- Ridge regression tends to shrink correlated predictors towards each other, approaching zero at the same rate, giving them a similar slope ($1/k^{\text{th}}$ the size of what any one would get). Lasso is more indifferent to correlated predictors and will tend to pick one and ignore the rest.

*Practice problem, due Wednesday, 9am:*  
(g) Compare ridge regression and the lasso on the credit card balance data set (but not the ID variable). Standardize the quantitative variables first. What do you learn/recommend?