Economics 142
Problem Set 8

In this problem set we will use some methods to develop good forecasts of earnings for people with different undergraduate majors. The data set is the 2010 American Community Survey, which collected data on earnings, education, age, etc, as well as detailed field of study (FOS) for people with BA degrees (or higher).

The file ps8.csv has 54,534 observations on people age 30-40 with exactly a BA degree who had earnings in the 12 months before the survey. codebook.pdf is a codebook. The table at the end shows the means for the main variables. On average, earnings were about 60,000 per year.

The FOS variable is called "FOD1P" and has 173 codes, e.g.
1100=general agriculture
2102=computer science
2304=elementary education
3700=mathematics
5501=economics

At the end of the "descriptives.txt" file is a listing of all the degree codes, the number of observations with the code, which ranges from 1 to over 2,300.

1. Using the first 2 digits of FOD1P only (which we will call "2-digit FOS") , follow the procedure presented in Lecture 20 to develop "shrinkage" estimates of mean earnings by 2-digit FOS. In particular, we will assume the correct model of earnings is

$$y_{vi} = \alpha + \beta_v + \epsilon_{vi}$$

where y_{vi} is log annual earnings of person i in 2-digit FOS v, and β_v is the deviation of mean earnings in field of study v from the grand mean.

a) divide the sample randomly into 2 subsamples: training and test.

b) using the training sample, get the mean earnings for each 2-digit FOS ($\overline{y}_v$), and the number of observations for this mean. Also find the grand mean of earnings ($\overline{y}$).

c) As shown in lecture 20, the best forecast of the mean for a new observation from 2-digit FOS v is:

$$\widehat{y}_v = \theta_v \overline{y} + (1 - \theta_v)\overline{y}_v$$

where

$$\theta_v = \frac{N - N_v}{N - N_v + N_v k}$$

and k is a positive constant. So, consider a range of possible values for k. For each value k form θ_v, use this to predict the mean for the 2-digit FOS, and then construct the mean squared forecast error in the test subsample:

$$MSFE = \frac{1}{N_{Test}} \sum_v \sum_i (y_{vi} - \widehat{y}_v)^2$$

where the sums are taken over observations in the TEST subsample. Plot MSFE for different values of k and find an approximately "best k".

2. Instead of shrinking, in this part of the question we will use cross validation to find an optimal value of the ridge "tuning parameter" λ for a ridge regression model fit to the training sample, using dummy variables for the different 2-digit FOS's. As preparation, read section 6.6.1 in ISLR. When you run the ridge model, turn off the "standardization" default, since the dummies are already (more or less) on the same scale.

a) Consider a range of possible values for λ. For each value λ fit a ridge regression model to the training sample, using dummy variables for the different 2-digit FOS's. Use the estimates from this model to form predictions for earnings in each 2-digit FOS, and then construct the mean squared forecast error *in the test subsample*:

$$MSFE = \frac{1}{N_{Test}} \sum_v \sum_i (y_{vi} - \widehat{y}_v)^2$$

Plot MSFE for different values of $\lambda$and find an approximately "best λ". How does the MSFE from the best-λ ridge model compare to the MSFE from your best shrinkage model?

3. Repeat part 2 using lasso. How does the MSFE from the best-λ lasso model compare to the MSFE from your other models?

4. (Optional). Try using a lasso model to predict earnings using the full set of 174 dummies for all FOS's.