Economics 142
Problem Set 9

In this problem set we will analyze the relationship between a new-born infant's birthweight and her "Apgar Score". The Apgar score was devised by Virginia Apgar, an anethesiologist, in 1953. It is based on evaluation of five components (breathing, heart rate, muscle tone, reflexes, and skin color), each of which is scored 0 1 or 2. Babies with scores below 9 or 10 often have medical problems in their first year of life.

The data set apar.csv has 48,470 observations on newborns, with their "5 minute Apgar Score" (apgar5), their birthweight (bthwght) in grams (mean = 3347.035), and their mother's height in inches (height). We are interested in developing a model that can predict the probability of having an Apgar score$\leq 8$.

a) standardize birthweight (to have mean 0 and standard deviation 1). Calling the standardized birthweight variable $x$, fit a series of linear probability models for $y = 1[Apgar \leq 8]$:

- model 1: constant and $x$
- model 2: constant and $x, x^2$
- model 3: constant and $x, x^2, x^3$

....

- model 6: constant and $x, x^2, ...x^6$
- model 7: constant and $x, x^2, x^3, x^4$ plus two other variables
  $Low = 1[x < -2]$, $High = 1[x > 2]$
- model 8: constant and $x, x^2, x^3, x^4$ plus mother's height

Using a RMSE criterion, which model has the best "within sample" fit?

***Hint***: this means that you get the "Root mean squared error" from each regression model and select the model with lowest RMSE.

b) Now we will compare models using an out-of-sample comparison based on 5-fold cross-validation.

(i) Randomly divide your sample into 5 equal-sized "folds"

(ii) For fold $k = 1...5$ conduct the following steps:

- excluding the data for fold $k$, fit model 1, then use that model to predict $y$ for observations in fold $k$

- repeat the previous step for models 2-8

- for each fold you now have 8 predicted probabilities of low Apgar, in each case based on data for the other 4/5 of the sample

(iii) combine the predictions for all 5 folds

(iv) now using an RMSE criterion, which model has the best "out of sample" fit?

(v) Using the model selected in step (iv) plot the probability of a low Apgar score against birthweight