Homework 3 – High School Quarterbacks Database

Data Reading and manipulation

A1. What’s the correlation between transferring schools and making a pro roster? What’s the mean, median and standard deviation of both variables?

A2. Create a new variable “transfer\_changeP\_nopro” for QBs who transferred schools or changed positions but never made it to a pro roster. What’s the mean, median and standard deviation?

A3. What’s the correlation between height and weight? What’s the mean, median and standard deviation of both variables?

A4. What’s the correlation between the player’s hometown median income, median home price and percentage of the population with a bachelor’s degree?

Regression Analytics

B1. Run a simple regression with pass attempts as the dependent variable (Y) and the X variables are coach age, coach experience, height, weight, hometown median income, hometown median home price and games played as the independent variables (X)? What does the R-squared statistic mean here? What possible problems are you having with this regression?

B2. Run a simple regression with “transfer\_changeP\_nopro” as the dependent variable (Y) and the X variables are coach experience, height and games played as the independent variables (X)? What does the R-squared statistic mean here? What possible problems are you having with this regression? What is the estimated probability of a HS QB having “transfer\_changeP\_nopro” if the coach has 20 years of experience, the player’s height is 75 inches and the player was in 20 games?

Data Mining

C1. Do your best. Forecast pro\_roster using any of the information here and any combination/transformation of the data you desire.

C2. Do your best. Forecast “transfer\_changeP\_nopro” using any of the information here and any combination/transformation of the data you desire.

C3. Using only coach and institutional behavior do your best to estimate “plays” involving the QB. What’s the best R-squared you can achieve WITHOUT multicollinearity, endogeneity or other issues?

C4. Using only player specific information do your best to estimate “plays” involving the QB. What’s the best R-squared you can achieve WITHOUT multicollinearity, endogeneity or other issues?

C5. Do your best. Forecast “plays” involving the QB using any of the information here and any combination/transformation of the data you desire.