PS C236A / Stat C239A
Problem Set 5 - Solutions

1) a) The Mahalanobis distance is defined as:

$$D_m(X_i, X_j) = \left\{(X_i - X_j)^T S^{-1}(X_i - X_j)\right\}^{\frac{1}{2}}$$

Where S^{-1} is the inverse of the sample covariance matrix of X.

A binary variable with probability of success p has variance $p(1-p)$. A variable with $p = \frac{1}{2}$ therefore has variance of $p(1-p) = \frac{1}{4}$, whereas a variable with p near 0 would have variance near 0. Since we take the inverse of the sample covariance matrix, therefore dividing by the variance, a variable with $p = \frac{1}{2}$ will be given less weight than a variable with p near 0 (or, similarly a variable with p near 1). By FOC, we can show that the variance of a binary variable is greatest when $p = \frac{1}{2}$, so a binary variable with $p = \frac{1}{2}$ will be given less weight than any binary variable with $p \neq \frac{1}{2}$

b) Variables with long tails or extreme outliers tend to have inflated variances, and by the same logic as above, any variable with larger variance will be given relatively less weight.

c) We should be concerned. Outliers and long tails do not make a covariate unimportant, so we may not wish to downweight it relative to other covariates. Binary variables that are very rare may not be of overriding importance, so it may not be wise to give them significantly higher weight than binary variables with p closer to $\frac{1}{2}$. However, if it is a rare binary event, then we might want to treat a difference in outcome as worse than a difference in outcome for a covariate where p is closer to $\frac{1}{2}$. Overall, we should be concerned that Mahalanobis distance exhibits these behaviors for variables for which the theory was not designed.

3) See `HW5_Answers.R` for solutions