STA721 **Homework 5**

1. Show that $\mathsf{P}_{\mathbf{X}^T} = (\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^-$ is a projection onto the column space of $\mathbf{X}^T$ where $(\mathbf{X}^T\mathbf{X})^-$ is a generalized inverse. Does this depend on the actual choice of generalized inverse? (explain) Is this an orthogonal projection?

2. Show that for an estimable function $\lambda = \mathbf{X}^T\mathbf{a}$ with $\mathbf{a} \in C(\mathbf{X})$ that $(\mathbf{I} - \mathsf{P}_{\mathbf{X}^T})\boldsymbol{\lambda} = \mathbf{0}$

3. Using the spectral decomposition of $(\mathbf{X}^T\mathbf{X})$ and the Moore-Penrose generalized inverse (see class notes) find a simple expression for $\mathbf{I} - \mathsf{P}_{\mathbf{X}^T}$ in terms of a reduced set of the eigenvectors of $\mathbf{X}^T\mathbf{X}$.

4. If $\mathbf{X}$ is full column rank, does a Best Linear Unbiased Prediction (BLUP) exist for all $\mathbf{x}_* \in \mathbb{R}^{p+1}$ ($\mathbf{x}_* \neq \mathbf{0}$)? Prove or Disprove.

5. (optional) Write a function in R to find the projection $(\mathbf{I} - \mathsf{P}_{\mathbf{X}^T})\boldsymbol{\lambda}$ with the design matrix (with intercept) and lambda (vector or matrix) as input. (post the `R` code on Piazza) Apply your function to the example from class and compare to the conclusions from `epredict`. What sort of tolerance do you need to decide if $(\mathbf{I} - \mathsf{P}_{\mathbf{X}^T})\boldsymbol{\lambda} = \mathbf{0}$?. Extra challenge - have your function return the estimates, SE and confidence intervals!

6. For the Prostate data: create "dummy" or indicator variables for the levels of the gleason scores and add to the dataframe `Prostate$D7 = (gleason == 7)` and show that they are linearly related to the intercept.

7. Fit a linear model of with response `lpsa` including all of the dummy variables and the intercept. What are the coefficients? If you change the order that the dummy variables enter the model formula, what happens to the coefficients? If you force the intercept to be zero (add -1 to the formula) what are the results?

8. Using as.factor(gleason) as a predictor in `lm`, what is the equivalent model formula using dummy variables? See `model.matrix` to extract the design matrix. What are the interpretation for these coefficients? (provide an explanation in a couple of sentences with the actual estimates.)

9. In the model with all dummy variables and the intercept, use the theorem from class to show that each of the individual coefficients are not estimable.

10. (for the energetic student. otherwise optional) The epredict function assumes that the intercept is always included, so any linear combination of $\boldsymbol{\beta}$ always has the intercept added which means we cannot use the function to see if individual $\boldsymbol{\beta}_j$ are estimable via a $\boldsymbol{\lambda} = (0, 0, 1, \ldots 0)^T$. Create a new variable that is a column of ones `Prostate$I = rep(1, n)` and fit the model using the formula `lpsa` $\sim$ `I + D6 + D7 + D8 + D9 -1` where D7 is the dummy variable indication that the gleason score is 7 and -1 drops the column of ones added by default. Create a data frame for predicting that will let you demonstrate with epredict that none of the individual $\boldsymbol{\beta}$ are estimable.