Statistics 600 Problem Set 3
Due in class on Wednesday, November 4th.

1. Suppose we have a $n \times p$ design matrix $X$ in which all covariates are standardized, the correlation of any pair of covariates is $\rho$, and the elements of $X$ are uniformly bounded by a constant $\kappa$. Suppose that both $p$ and $n$ are growing to infinity. Show that if $p/n \to 0$ then $\text{var}(\hat{Y}_i) \to 0$ for each $i$.

2. Suppose we have a sequence of populations

$$Y^{(n)} = \sum_{j=1}^{p_n} \beta_j^{(n)} X_j^{(n)} + \epsilon^{(n)};$$

where $p_n$ is a sequence of integers increasing in $n$, $X_1^{(n)}, \ldots, X_{p_n}^{(n)}$ are mutually orthogonal vectors in $\mathbb{R}^n$ for each $n$, each $X_j^{(n)}$ satisfies $\bar{X}_j^{(n)} = 0$ and $n^{-1} \sum_{i=1}^n X_j^{(n)}(i)^2 = 1$. For each $n$, $\epsilon^{(n)}$ is an iid vector of standard normal random values.

(a) Suppose we use least squares to estimate the $\beta_j^{(n)}$, then form the fitted values $\hat{Y}^{(n)}$. Derive an expression for $n^{-1} E \|\hat{Y}^{(n)} - E[Y^{(n)}]\|^2$. Under what circumstances would this converge to zero as $n \to \infty$?

(b) Now suppose that $\beta_j^{(n)} = \theta^j$ for some $0 < \theta < 1$, $p_n \equiv \infty$, and we form our fitted values using the first $k_n$ variables. Derive an expression for $n^{-1} E \|\hat{Y}^{(n)} - E[Y^{(n)}]\|^2$, and use this expression to derive the optimal $k_n$ for each $n$. Describe in qualitative terms how the optimal $k_n$ scales with $n$.

3. Suppose we have $p$ covariates and an intercept in our model, and we calculate the partial $R^2$ for adding $X_1$ to the model that already contains $X_2, \ldots, X_p$. Show how this partial $R^2$ value can be monotonically related to an F statistic.

4. Using the 2009-2010 NHANES data, develop a model that predicts systolic blood pressure as a function of age and gender. Build a realistic model that accounts for the possibility that the relationship between blood pressure and age is different for men and women, and that this function may not be linear.

(a) Plot the difference between female mean and male mean blood pressure as a function of age, with a 95% simultaneous confidence band.

(b) Imagine that for people over age 60, the probability of a person dying during each year has the form $a + bx$, where $a$ and $b$ are chosen so that the mortality is consistent with figure 1 in this paper:


(Feel free to manually approximate the numbers based on the graph. Use your judgement about which values from the graph to use.)
Devise a way to approximately recover the mean blood pressure values for men between 60 and 70 years of age that would be observed in the absence of mortality due to hypertension.

Note: doing this will require making some approximations that may or may not hold depending on certain conditions that cannot be verified. You don’t need to do anything here that is very complicated. Try to come up with a relatively simple way to do this that at least gives us a rough idea about what might be happening.