Homework 5 (due May 23)

In your solution include your name and the homework number. Please staple your pages together. When solving the problems below, give detailed derivations in order to receive credit.

1. *(Projection Matrix 1.)* A symmetric matrix $P \in \mathbb{R}^{n\times n}$ is a projection matrix if $P^2 = P$.
   (a) Find $(I_n - P)^2$ and $(I_n - P)P$.
   (b) Assume $\text{rank}(P) = r$, and then find all eigenvalues of P.
2. *(Projection Matrix 2.)*A symmetric matrix $P \in \mathbb{R}^{n\times n}$ is a projection matrix if $P^2 = P$.
   (a) For any $X \in \mathbb{R}^{n\times p}$ such that $(X^T X)^{-1}$ exists. Show $X(X^T X)^{-1}X^T$ is a projection matrix.
   (b) Let $\mathbf{1} \in \mathbb{R}^{n\times 1}$ be a column vector of all ones. Show $n^{-1}\mathbf{1}\mathbf{1}^T$ is a projection matrix.
   (c) If both P_1 and P_2 are projection matrices. Assume $P_1P_2 = 0$. Show $P_1 + P_2$ is a projection matrix.
3. *(Projection Matrix 3.)* A symmetric matrix $P \in \mathbb{R}^{n\times n}$ is a projection matrix if $P^2 = P$. Assume $\text{rank}(P) = r$. For $z \sim N(0, I_n)$, use eigenvalue decomposition to find the distribution of $z^T P z$.
4. *(Variance Bias Trade-off.)* For any estimator $\hat{\theta}$, prove $\mathbb{E}(\hat{\theta} - \theta)^2 = \mathsf{Var}(\hat{\theta}) + (\mathbb{E}\hat{\theta} - \theta)^2$.
5. *(Variance Estimation 1.)* For the linear model $y \sim N(X\beta, \sigma^2 I_n)$, recall that $\hat{y} = Hy$ with $H = X(X^T X)^{-1}X^T$. The residue is defined as $\hat{e} = (I - H)y$. In the class, we derived that $\|\hat{e}\|^2/\sigma^2 \sim \chi^2_{n-p}$. Use this fact to answer the following questions:
   (a) Define $\hat{\sigma}^2 = \frac{1}{n-p}\|\hat{e}\|^2$. What is $\mathbb{E}\hat{\sigma}^2$?
   (b) What is $\mathbb{E}(\hat{\sigma}^2 - \sigma^2)^2$?
   (c) Define $\tilde{\sigma}^2 = \frac{1}{n}\|\hat{e}\|^2$. What is $\mathbb{E}(\tilde{\sigma}^2 - \sigma^2)^2$?
   (d) Consider $\sigma_c^2 = c\|\hat{e}\|^2$. Find the c such that $\mathbb{E}(\sigma_c^2 - \sigma^2)^2$ is the smallest.
   (e) If you want to estimate σ^2, which estimator would you choose? Explain.
6. *(Variance Estimation 2.)* Consider the linear model $y \sim N(X\beta, \sigma^2 I_n)$.
   (a) Find the joint MLE of (β, σ^2), denoted as $(\hat{\beta}, \hat{\sigma}^2)$.
   (b) Construct a pivotal of σ^2 using $\hat{\sigma}^2$. Find an exact 95% confidence interval of σ^2.

7. *(Data Analysis.)* Download NewHaven.txt from chalk. Set up a working directory on your own computer, and read the data into R. Write a report of data analysis that addresses the following items. The report should be printed and submitted together with the homework. No need to include the code.

   (a) Summarize the whole data set.

   (b) Pick up a subset of rows that you want to study. For example. you can study all houses or all condos. You can also study houses that are not too expensive. Whatever subset of rows you pick, you need to justify your choice with some understandings of the data set.

   (c) Pick up at least five variables. Explain what they are, and fit a linear model to predict current values.

   (d) Analyze the linear model result. Use the function cooks.distance to find outliers. Use google search to check why these are outliers. Fit the model after removing those outliers.

   (e) You may want to repeat the last step. Write a nice paragraph with a clear conclusion for your findings. Try to include many nice plots in your report so that it is easy to read.