Homework 7

1. Consider $y \sim N(X\beta, \sigma^2 I_n)$. An unbiased estimator for σ^2 is $\hat{\sigma}^2 = \frac{1}{n-p}\|y - X\hat{\beta}\|^2$.

   (a) Find the asymptotic distribution of $\sqrt{n-p}(\hat{\sigma}^2 - \sigma^2)$.

   (b) Find a transform g so that $\sqrt{n-p}(g(\hat{\sigma}^2) - g(\sigma^2)) \rightsquigarrow N(0,1)$.

   (c) Using the above asymptotic result to construct an approximate $(1-\alpha)$-confidence interval for σ^2.

2. Suppose $y_i \sim N(\beta_0, \sigma^2)$ independently for $i = 1, ..., n$.

   (a) You fit the mean model and get $\hat{\beta}_0 = \arg\min \sum_{i=1}^n (y_i - \beta_0)^2$. Find $\mathbb{E}(\hat{\beta}_0 - \beta_0)^2$.

   (b) You take the covariates $x_1, ..., x_n$, fit a regression model, and get $(\hat{\beta}_0, \hat{\beta}_1) = \arg\min \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$. Find $\mathbb{E}(\hat{\beta}_0 - \beta_0)^2$. Attention: the data generating process is still $y_i \sim N(\beta_0, \sigma^2)$ independently for $i = 1, ..., n$.

   (c) Since the data is generated by the simpler model $N(\beta_0, \sigma^2)$, when you fit the simpler model in (a) or the more complicated model in (b), both are correct approaches. However, they give different results in terms of estimation accuracy (variance). Discuss your results.

3. Suppose $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ independently for $i = 1, ..., n$.

   (a) You fit the mean model and get $\hat{\beta}_0 = \arg\min \sum_{i=1}^n (y_i - \beta_0)^2$. Find $\mathbb{E}(\hat{\beta}_0 - \beta_0)^2$.

   (b) You fit the regression model, and get $(\hat{\beta}_0, \hat{\beta}_1) = \arg\min \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$. Find $\mathbb{E}(\hat{\beta}_0 - \beta_0)^2$.

   (c) When the data is generated by the more complicated model, using the simpler one in (a) to fit the data may cause problem. Discuss your results.

4. Consider $y \sim N(X\beta, \sigma^2 I_n)$. Test $H_0 : \beta_S = 0$ against its alternative. The notation S is a subset of $\{0, 1, 2, ..., p-1\}$. For example, if $S = \{1, 2, 3\}$, then the null hypothesis becomes $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, which means you want to test whether the first three covariates are significant or not. We also use the notation s for the size of S and X_S is an $n \times s$ submatrix with columns in S taken from $X \in \mathbb{R}^{n\times p}$.

   (a) The LSE under H_0 is given by $\hat{\beta}_{H_0} = (X_{S^c}^T X_{S^c})^{-1} X_{S^c}^T y$. Find the distribution of $\hat{\beta}_{H_0}$ under H_0.

   (b) The fits under H_0 and H_1 are $\hat{y}_{H_0} = X_{S^c}\hat{\beta}_{H_0} = X_{S^c}(X_{S^c}^T X_{S^c})^{-1} X_{S^c}^T y$ and $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y$ with $\hat{\beta}$ being the LSE of the full model. Prove
   $$\|y - \hat{y}_{H_0}\|^2 = \|y - \hat{y}\|^2 + \|\hat{y} - \hat{y}_{H_0}\|^2.$$

   (c) Are $\|y - \hat{y}\|^2$ and $\|\hat{y} - \hat{y}_{H_0}\|^2$ independent? Why?

   (d) Find the distribution of $\|y - \hat{y}_{H_0}\|^2/\sigma^2$ and $\|\hat{y} - \hat{y}_{H_0}\|^2/\sigma^2$ under H_0.

   (e) Construct an F-test for the testing problem.

   (f) Show that if $S = \{1, 2, ..., p-1\}$, we get the results that we have learned in the class.