Homework 5 6

Due Thursday, 10/1/20

1. (Independence of quadratic forms) Suppose $\boldsymbol{Y} \sim N_n\left(\boldsymbol{\mu}, \sigma^2 I_n\right)$ for some mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times n}$.

   (a) Show that if $\boldsymbol{A}\boldsymbol{B}^T = \mathbf{0}_{n \times n}$, then $\boldsymbol{A}\boldsymbol{Y}$ and $\boldsymbol{B}\boldsymbol{Y}$ are independent.

   (b) Now suppose $\boldsymbol{A}, \boldsymbol{B}$ are symmetric and idempotent matrices, where $\boldsymbol{A}\boldsymbol{B} = \mathbf{0}_{n \times n}$. Use part (a) to show that the quadratic forms $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}$ and $\boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}$ are independent.

2. (General F-tests) Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ with $p \leq n$. Suppose

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_n\left(\mathbf{0}_n, \sigma^2 I_n\right)$$

   for some non-random $\boldsymbol{\beta} \in \mathbb{R}^p$. After thinking about the the above model, you construct a second non-random, full rank design matrix $\boldsymbol{L} \in \mathbb{R}^{n \times s}$, where $s < p$ and $\operatorname{Im}(\boldsymbol{L}) \subset \operatorname{Im}(\boldsymbol{X})$, and have reason to believe that $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{L}\boldsymbol{\gamma}$ for some $\boldsymbol{\gamma} \in \mathbb{R}^s$. You wish to formally test the null hypothesis $H_0 : \mathbb{E}(\boldsymbol{Y}) = \boldsymbol{L}\boldsymbol{\gamma}$ for some $\boldsymbol{\gamma} \in \mathbb{R}^s$.

   (a) Suppose $p = 3$ and $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$. Find the design matrices $\boldsymbol{X}$ and $\boldsymbol{L}$, and show that $\operatorname{Im}(\boldsymbol{L}) \subset \operatorname{Im}(\boldsymbol{X})$, when the null hypothesis is

      (i) $H_0 : \beta_2 = 0$.

      (ii) $H_0 : \beta_1 + \beta_2 = 0$.

   (b) Let $\boldsymbol{H}_X = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$ and $\boldsymbol{H}_L = \boldsymbol{L}\left(\boldsymbol{L}^T\boldsymbol{L}\right)^{-1}\boldsymbol{L}^T$ and define

$$SSE_X = SSE \text{ when the design matrix is } \boldsymbol{X}$$
$$SSE_L = SSE \text{ when the design matrix is } \boldsymbol{L}$$

      and let the F-statistic be

$$f = \frac{(SSE_L - SSE_X)/(p-s)}{SSE_X/(n-p)}.$$

      Show that we can write f as

$$f = \frac{\boldsymbol{Y}^T(\boldsymbol{H}_X - \boldsymbol{H}_L)\boldsymbol{Y}/(p-s)}{\boldsymbol{Y}^T(I_n - \boldsymbol{H}_X)\boldsymbol{Y}/(n-p)}.$$

   (c) Show that $\boldsymbol{H}_X - \boldsymbol{H}_L$ is symmetric and idempotent, and that $(I_n - \boldsymbol{H}_X)(\boldsymbol{H}_X - \boldsymbol{H}_L) = \mathbf{0}_{n \times n}$.

   (d) Use problem 1 to prove that $f \sim F_{(p-s),(n-p)}$ when then null hypothesis $H_0 : \mathbb{E}(\boldsymbol{Y}) = \boldsymbol{L}\boldsymbol{\gamma}$ for some $\boldsymbol{\gamma} \in \mathbb{R}^s$ is true.

3. (Interpretation of R^2) Let $\boldsymbol{Y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be a full rank design matrix and $R^2 = 1 - \frac{SSE}{SSTO}$ be the coefficient of determination from the regression of $\boldsymbol{Y}$ onto $\boldsymbol{X}$. Finally, let $\bar{\hat{Y}} = n^{-1}\mathbf{1}_n^T\hat{\boldsymbol{Y}}$, $\bar{Y} = n^{-1}\mathbf{1}_n^T\boldsymbol{Y}$ and

$$r_{\hat{Y},Y} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{\hat{Y}}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{\hat{Y}}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

be the empirical correlation between $\hat{\boldsymbol{Y}}$ and $\boldsymbol{Y}$. We will assume throughout this problem that $\boldsymbol{X}$ contains the intercept, i.e. $\mathbf{1}_n \in \operatorname{Im}(\boldsymbol{X})$.

   (a) Use the assumption that $\mathbf{1}_n \in \operatorname{Im}(\boldsymbol{X})$ to show that $\bar{\hat{Y}} = \bar{Y}$.

   (b) Use part (a) to show that $\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{\hat{Y}}\right)\left(Y_i - \bar{Y}\right) = \boldsymbol{Y}^T\left(\boldsymbol{H} - n^{-1}\mathbf{1}_n\mathbf{1}_n^T\right)\boldsymbol{Y}$. Will $r_{\hat{Y},Y}$ ever be smaller than 0? Why or why not?

   (c) Show that $R^2 = r_{\hat{Y},Y}^2$.

4. Answer the following questions using the data from "steam_text.txt".

   (a) Suppose you regress steam (Y) onto fat (X_1) and glycerine (X_2).

      (i) Write down the model you are assuming when performing this regression (i.e. what is the mean and variance model). Provide an interpretation for the coefficients in the mean model.

      (ii) In separate plots, plot $\hat{\epsilon}$ as a function of $\hat{\boldsymbol{Y}}$, fat and glycerine. Do you see any evidence that the mean or variance model is incorrect?

      (iii) Consider the null hypothesis that the coefficients for both fat and glycerine are 0. At a significance level of $\alpha = 0.05$, what do you conclude about these coefficients?

      (iv) Plot the variable "temp" against the residuals from this regression. What can you conclude from this plot?

   (b) Now regress steam (Y) onto fat (X_1), glycerine (X_2) and temp (X_3).

      (i) Consider the null hypothesis that the coefficients for both fat and glycerine are 0. At a significance level of $\alpha = 0.05$, what do you conclude about these coefficients?

      (ii) Why are the P values from this test so much smaller than those from part (a)?