

# Homework 5

Due Thursday, 10/3/19

1. (Independence of quadratic forms) Suppose  $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$  for some mean vector  $\boldsymbol{\mu} \in \mathbb{R}^n$  and let  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ .
  - (a) Show that if  $\mathbf{AB}^T = \mathbf{0}_{n \times n}$ , then  $\mathbf{AY}$  and  $\mathbf{BY}$  are independent.
  - (b) Now suppose  $\mathbf{A}, \mathbf{B}$  are symmetric and idempotent matrices, where  $\mathbf{AB} = \mathbf{0}_{n \times n}$ . Use part (a) to show that the quadratic forms  $\mathbf{Y}^T \mathbf{A} \mathbf{Y}$  and  $\mathbf{Y}^T \mathbf{B} \mathbf{Y}$  are independent.
2. (General F-tests) Let  $\mathbf{X} \in \mathbb{R}^{n \times p}$  with  $p \leq n$ . Suppose

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

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  $\mathbf{L} \in \mathbb{R}^{n \times s}$ , where  $s < p$  and  $\text{Im}(\mathbf{L}) \subset \text{Im}(\mathbf{X})$ , and have reason to believe that  $\mathbb{E}(\mathbf{Y}) = \mathbf{L}\boldsymbol{\gamma}$  for some  $\boldsymbol{\gamma} \in \mathbb{R}^s$ . You wish to formally test the null hypothesis  $H_0 : \mathbb{E}(\mathbf{Y}) = \mathbf{L}\boldsymbol{\gamma}$  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  $\mathbf{X}$  and  $\mathbf{L}$ , and show that  $\text{Im}(\mathbf{L}) \subset \text{Im}(\mathbf{X})$ , when the null hypothesis is
  - (i)  $H_0 : \beta_2 = 0$ .
  - (ii)  $H_0 : \beta_1 + \beta_2 = 0$ .
- (b) Let  $\mathbf{H}_X = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$  and  $\mathbf{H}_L = \mathbf{L}(\mathbf{L}^T \mathbf{L})^{-1} \mathbf{L}^T$  and define

$$SSE_X = SSE \text{ when the design matrix is } \mathbf{X}$$

$$SSE_L = SSE \text{ when the design matrix is } \mathbf{L}$$

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

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

- (c) Show that  $\mathbf{H}_X - \mathbf{H}_L$  is symmetric and idempotent, and that  $(\mathbf{I}_n - \mathbf{H}_X)(\mathbf{H}_X - \mathbf{H}_L) = \mathbf{0}_{n \times n}$ .
- (d) Use problem 1 to prove that  $f \sim F_{(p-s), (n-p)}$  when the null hypothesis  $H_0 : \mathbb{E}(\mathbf{Y}) = \mathbf{L}\boldsymbol{\gamma}$  for some  $\boldsymbol{\gamma} \in \mathbb{R}^s$  is true.

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

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

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

- (a) Use the assumption that  $\mathbf{1}_n \in \text{Im}(\mathbf{X})$  to show that  $\bar{\hat{Y}} = \bar{Y}$ .
  - (b) Use part (a) to show that  $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})(Y_i - \bar{Y}) = \mathbf{Y}^T (\mathbf{H} - n^{-1} \mathbf{1}_n \mathbf{1}_n^T) \mathbf{Y}$ . Will  $r_{\hat{\mathbf{Y}}, \mathbf{Y}}$  ever be smaller than 0? Why or why not?
  - (c) Show that  $R^2 = r_{\hat{\mathbf{Y}}, \mathbf{Y}}^2$ .
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  $\hat{\mathbf{Y}}$ , 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)?