STA302/1001 Quiz #2 Solution

1. (a) For convenience, we assign 1 to Statistics, 2 to Mathematics, 3 to Economics, and 4 to Psychology. There are totally $n = 32$ students, for the ith students, define $u_{ij} = \begin{cases} 1, & \text{if } D_i = j, \\ 0, & \text{otherwise.} \end{cases}$ Then the linear model with intercept that contains interaction between library hours (X) and departments (D) is

$$y_i = \eta_{01} + \eta_{11}x_i + \sum_{j=2}^{4}(\eta_{0j}u_{ij} + \eta_{1j}u_{ij}x_i),, \quad i = 1, \ldots, 32.$$

where η_{01} and η_{11} are the intercept and the slope effects of library hours for Statistics (baseline). For $j = 2, 3, 4$, η_{0j} is the difference between intercepts of department at level j and Statistics, η_{1j} is the difference between library hour slope effects of department at level j and Statistics.

Note: If the model has an intercept, only 3 dummy variables are used. The assignments of numerical values to D can be arbitrary, I use $\{1, 2, 3, 4\}$ for convenient definition of dummy variables. The students also may write a model as $E(Y|X, D) = \eta_0 + \eta_1 X + \sum_{j=2}^{4}(\eta_{0j}U_j + \eta_{1j}U_jX)$.

(b) The hypothesis of interest is $H_0 : \eta_{12} = \eta_{13} = \eta_{14} = 0$, i.e., the interaction effects between D and X are not significant. This corresponds to the reduced model with parallel regression lines but different intercept for four departments.

$$y_i = \eta_{01} + \eta_{11}x_i + \sum_{j=2}^{4}\eta_{0j}u_{ij}, \quad i = 1, \ldots, 32.$$

(c) Model III is a model with the same intercept and slope for for departments, i.e., only 2 regression parameters. The degrees of freedom for Model I is $32 - 8 = 24$, for Model II is $32 - 5 = 27$, for Model III is $32 - 2 = 30$.

The test in (b) is comparing Model I (full) v.s. Model II (reduced). The F-value is

$$F = \frac{(RSS_{II} - RSS_I)/(27 - 24)}{RSS_I/27} = \frac{(581 - 565)/3}{565/27} = 0.2549 < F_{0.05,3,24} = 3.01.$$

Do NOT reject H_0, i.e., Model II is adequate.

(d) The null hypothesis comparing Model II and Model III is $H_0 : \eta_{02} = \eta_{03} = \eta_{04} = 0$. The F-value is given by

$$F = \frac{(RSS_{III} - RSS_{II})/(30 - 27)}{RSS_{II}/27} = \frac{(866 - 581)/3}{581/27} = 4.4148 > F_{0.05,3,27} = 2.96.$$

Thus we reject H_0, and Model II cannot be further reduced.

2. (a) Write the model in the centered form, $y_i = \beta_0 + \beta_1 \bar{x} + \beta_1(x_i - \bar{x}) + e_i$, the design matrix is $X = \begin{pmatrix} 1 & x_1 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}$. Since $\sum_{i=1}^n (x_i - \bar{x}) = 0$, we have $X^\top X = \begin{pmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x})^2 \end{pmatrix}$. Then

$$h_{ij} = (1, x_i - \bar{x}) \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{pmatrix} \begin{pmatrix} 1 \\ x_j - \bar{x} \end{pmatrix} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^n (x_k - \bar{x})^2}.$$

Thus $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^n (x_k - \bar{x})^2} \geq \frac{1}{n}$, $\sum_{j=1}^n h_{ij} = 1 + \frac{(x_i - \bar{x})}{\sum_{k=1}^n (x_k - \bar{x})^2} \sum_{j=1}^n (x_j - \bar{x}) = 1$.

Note: the students may also show $\sum_{j=1}^n h_{ij} = 1$ with the following argument given in practice solution (general for multiple linear models): because $HX = X(X^\top X)^{-1} X^\top X = X$, denote the kth column in X by $\boldsymbol{x}_{(k)} = (x_{1k}, \ldots, x_{nk})^\top$, we have $H\boldsymbol{x}_{(k)} = \boldsymbol{x}_{(k)}$. For the linear model with an intercept, the first column is $\mathbf{1}$, thus $H\mathbf{1} = \mathbf{1}$, i.e., the ith equality corresponds to $\sum_{j=1}^n h_{ij} = 1$.

(b) For replicates $x_k = x_i$, we have $h_{ii} = h_{ik}$. Also note $H^2 = H$, i.e., $h_{ii} = \sum_{j=1}^n h_{ij}^2 \geq r_i h_{ii}^2$, thus $h_{ii} \leq 1/r_i$.