ISyE 7406: Hints to Problem 2 of HW#1

Problem 2. Consider a simple linear regression model $Y = \beta_0 + \beta_1 x + \epsilon$. Suppose that we choose m different values of the independent variables x_i's, and each choice of x_i is duplicated, yielding k independent observations $Y_{i_1}, Y_{i_2}, \cdots, Y_{i_k}$. Is it true that the least squares estimates of the intercept and slope can be found by doing a regression of the mean responses, $\bar{Y}_i = (Y_{i_1} + Y_{i_2} + \cdots + Y_{i_k})/k$, on the x_i's? Why or why not? Explain.

Remarks: As in Problem 1, there are two kinds of linear regressions: one is based on a total of $n = mk$ "raw" observations (Y_i, x_i)'s, and the other is based on the m "average" observations $(\bar{Y}_i, x_i)$'s. If you have difficulty to investigate the general k case, it will be OK to consider $k = 2$ case!

Hints: You may have difficulty to do this "theoretical" question, which is the extension of Problem #1. Below we will give the hints for the case $k = 2$.

First, you need to understand the simple linear regression. Please note that while I did not discuss it specifically in class, I assume you learned it from your undergrad class, or you can easily derive it from the general linear regression results we discussed in class. For your information, I provide a brief review below.

In the simple linear regression, we assume we observe n observations (Y_i, X_i), and we want to fit the model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$. As in the lecture note, to estimate β_0 and β_1, the method of least squares is to find b_0 and b_1 that minimizes

$$SS_{err} = \sum_{i=1}^{n} [Y_i - (b_0 + b_1 x_i)]^2. \tag{1}$$

and

$$\hat{\sigma}^2 = \frac{SS_{err}}{n-2}$$

You can write them in the matrix notation of $Y_{n\times 1} = X_{n\times 2}\beta_{2\times 1} + \epsilon_{n\times 1}$, where

$$Y_{n\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{pmatrix}; \qquad X_{n\times 2} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{pmatrix}; \qquad \beta_{2\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix};$$

You can use our general results (or directly minimize SS_{err} by taking derivatives with respect to b_0 and b_1) to find the point estimates have a simple form:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}, \tag{2}$$

where $\bar{x} = (x_1 + \cdots + x_n)/n$ and $\bar{Y} = (Y_1 + \cdots + Y_n)/n$. The corresponding $100(1-\alpha)\%$ confidence intervals will be

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \quad \text{and} \quad \hat{\beta}_0 \pm t_{\alpha/2, n-2}\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}. \tag{3}$$

Now let us go back to our problem for $k = 2$. We want to fit the simple linear regression to two datasets. (1) The full data set has $2m$ observations, and the observations are

$$\begin{gathered} (Y_{11}, x_1) \\ (Y_{12}, x_1) \\ (Y_{21}, x_2) \\ (Y_{22}, x_2) \\ \cdots \\ (Y_{m1}, x_m) \\ (Y_{m2}, x_m). \end{gathered}$$

(2) The "average" data set has m observations, and the observations are

$$\begin{array}{lll} (\bar{Y}_1, x_1) & \text{with} & \bar{Y}_1 = (Y_{11} + Y_{12})/2 \\ (\bar{Y}_2, x_2) & \text{with} & \bar{Y}_2 = (Y_{21} + Y_{22})/2 \\ & & \cdots \\ (\bar{Y}_m, x_m) & \text{with} & \bar{Y}_m = (Y_{m1} + Y_{m2})/2 \end{array}$$

The question asks you whether the point estimates and confidence intervals of β_0 and β_1 are the same or not when fitting the simple linear regression to these two different data sets.

To provide a further hint, let us consider equation (1) for these two data sets.

For the full data set of $n = 2m$ observations, (1) can be rewritten as

$$SS_{err,1} = \sum_{i=1}^{m} \Big(y_{i_1} - (b_0 + b_1 x_i) \Big)^2 + \sum_{i=1}^{m} \Big(y_{i_2} - (b_0 + b_1 x_i) \Big)^2,$$

which has $2m$ terms.

Meanwhile, for the "average" data set, we only have $n = m$ observations, and (1) can be rewritten as

$$SS_{err,2} = \sum_{i=1}^{m} \Big(\bar{y}_i - (b_0 + b_1 x_i) \Big)^2,$$

which only has m terms.

Do these two datasets or approaches lead to the same solution of (b_0, b_1), i.e., the same point estimates of β_0 or β_1 in (2)?

As for the confidence intervals for β_0 or β_1 in (3), most of you realize that the sample size n are different for these two different approaches,and thus the $t_{\alpha/2,n-2}$ values are different. However, how about other terms in (3)? E.g., the $\hat{\sigma}$ values in these two different approaches?