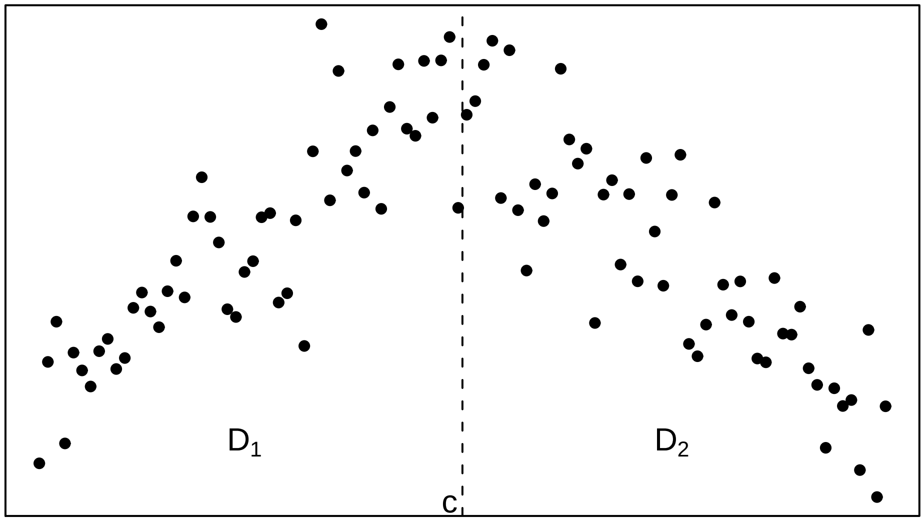
MA 575 – Qualifying Exam

Fall 2009

1. Consider the dataset $\{(x_i, y_i)\}_{i=1}^{2n}$ plotted in the figure below. The dataset is divided in two, with n observations where $x_i < c$ and n observations with $x_i > c$.



To accommodate the change of behavior at $x = c$, you adopt a *change point linear model*:

$$Y_i = \alpha_0 + \gamma_0 X_i + (\delta + \eta X_i) Z_i + e_i, \quad i = 1, \ldots, 2n,$$

where $Z_i = I(X_i > c)$ and $e_i | X_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

(a) What is the interpretation of δ and η?

Now suppose a previous analyst (naively) adopted separate models for dataset 1, $D_1 \doteq \{(x_i, y_i) : x_i < c\}$, and dataset 2, $D_2 \doteq \{(x_i, y_i) : x_i > c\}$:

$$Y_i = \alpha_j + \gamma_j X_i + e_{ij}, \quad e_{ij} | X_i \overset{\text{iid}}{\sim} N(0, \sigma_j^2), X_i \in D_j, j = 1, 2,$$

and found least squares estimates $\hat{\alpha}_j$, $\hat{\gamma}_j$, and $\hat{\sigma}_j^2$ for $j = 1, 2$.

(b) Provide least squares estimates for α_0, γ_0, δ, η, and σ^2, as functions of $\hat{\alpha}_j$, $\hat{\gamma}_j$, and $\hat{\sigma}_j^2$, $j = 1, 2$. Explain your answers considering the design matrices from each model.

The analyst conjectured that the lines from each model, left of c and right of c, would cross at (c, d). The analyst proceeded to compare two F-tests, one for testing the null hypothesis $\alpha_1 + c\gamma_1 = d$ with test statistic F_1, and the other for the null hypothesis $\alpha_2 + c\gamma_2 = d$ with statistic F_2.

(c) Give expressions for F_1 and F_2. What are their distributions under their respective nulls?

(d) Provide an F-test to verify the analyst's conjecture under your change point model (state null hypothesis, specify test statistic and its distribution under the null). What is the relation between your test statistic and F_1 and F_2?

2. To study the linear relation between Y and a set of p predictors X given by

$$Y_i = \mathbf{x}_i^T \beta + e_i, \quad e_i|X \overset{\text{iid}}{\sim} N(0, \sigma^2), \tag{1}$$

a dataset $\{(\mathbf{x}_i, y_i^{(1)}), (\mathbf{x}_i, y_i^{(2)})\}_{i=1}^n$ with two response replicates for each (multivariate) value of $\mathbf{x}_i$ was collected.

(a) Define $\bar{Y} \doteq (Y^{(1)} + Y^{(2)})/2$, the mean response over replicates. Show that the least squares estimate $\hat{\beta}$ from the full model above can be obtained directly by regressing $\bar{Y}$ on X. That is, if

$$\bar{Y}_i = \mathbf{x}_i^T \gamma + \tilde{e}_i, \quad \tilde{e}_i|X \overset{\text{iid}}{\sim} N(0, \sigma_B^2), \tag{2}$$

then $\hat{\gamma} = \hat{\beta}$, the least squares estimates coincide. What is the least squares estimate $\hat{\sigma}_B^2$ for the between-replicates variance σ_B^2?

(b) Further define the within-replicates variance estimate,

$$\hat{\sigma}_W^2 \doteq \frac{\sum_{j=1}^2 \sum_{i=1}^n (y_i^{(j)} - \bar{y}_i)^2}{n}.$$

What is the least squares estimate for σ^2 from model (1) in terms of $\hat{\sigma}_W^2$ and $\hat{\sigma}_B^2$?

(c) Perform a lack-of-fit test: express SS_{LOF}, SS_{PE}, and the test statistic as a function of $\hat{\sigma}_W^2$ and $\hat{\sigma}_B^2$. What is the distribution of the test statistic under the null?

(d) Suppose that $e_i|X \overset{\text{iid}}{\sim} N(0, \sigma^2/w_i)$, w_i *known*. How would you change the assumptions on $\tilde{e}_i$ in model (2) so that $\hat{\beta} = \hat{\gamma}$? Under your changes, what is now the expression for $\hat{\sigma}_B^2$? How would you redefine $\hat{\sigma}_W^2$ to obtain the same LOF test statistic *expression* from the previous item, under the unweighted model? Fully justify your answers.