Qualifying Exam: CAS MA 575

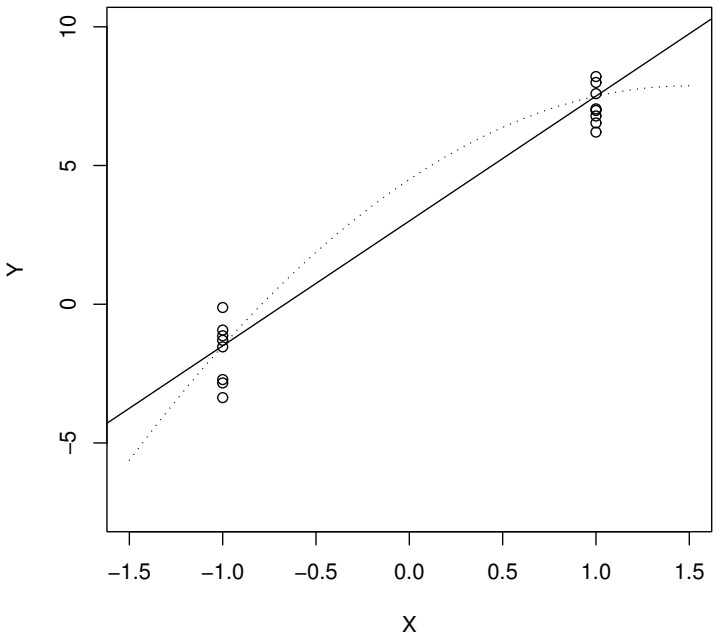
Boston University, Fall 2008

1. Consider the simple linear regression model

$$E[Y|X = x] = \beta_0 + \beta_1 x \ ,$$

and suppose that X takes on only values in the interval $[-1, 1]$. Imagine that we are allowed to select which values of X we observe, and furthermore, that we are confident that all of the standard regression model assumptions hold. Then, given a total 'budget' of n observations, principles of statistical experimental design suggest that it is optimal in estimating $\beta = (\beta_0, \beta_1)^T$ to take $n/2$ measurements $y_1, \ldots, y_{n/2}$ at $x_1 = \cdots = x_{n/2} = -1$ and $n/2$ measurements $y_{n/2+1}, \ldots, y_n$ at $x_{n/2+1} = \cdots = x_n = 1$.

That is, our data consists of $n/2$ replicates at $x = -1$ and $n/2$ replicates at $x = 1$. This scenario is illustrated in the figure below, for $n = 16$, with the linear mean function shown as a solid line.



a. Given measurements of this nature, show that the OLS estimate $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^T$ takes the form

$$\hat{\beta}_0 = \frac{\bar{y}^{(1)} + \bar{y}^{(-1)}}{2} \quad \text{and} \quad \hat{\beta}_1 = \frac{\bar{y}^{(1)} - \bar{y}^{(-1)}}{2} \ ,$$

where

$$\bar{y}^{(-1)} = (n/2)^{-1} \sum_{i=1}^{n/2} y_i \quad \text{and} \quad \bar{y}^{(1)} = (n/2)^{-1} \sum_{i=n/2+1}^{n} y_i$$

are the averages of the $n/2$ replicates at $x = -1$ and $x = 1$, respectively.

b. Provide concise expressions for $\text{Var}(\hat{\beta}_0)$, $\text{Var}(\hat{\beta}_1)$, and $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$.

c. Suppose that we are wrong in assuming a linear model and in fact a quadratic model holds instead, i.e.,

$$E[Y|X = x] = \beta_0 + \beta_1 x + \beta_2 x^2 \ ,$$

as is illustrated by the dotted line in the figure of the previous page. What will be the result of trying to fit this quadratic model to the measurements described above? Is it possible to test whether the linear or the quadratic model is more accurate using this data? If not, what additional measurements would you recommend taking?

2. Assume that the true linear model underlying a set of n measurements is of the form

$$\mathbf{Y} = \mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2 + \mathbf{e} \ ,$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are $n \times p_1$ and $n \times p_2$ matrices, respectively, and the elements e_i in the $n \times 1$ error vector $\mathbf{e}$ are uncorrelated, with mean zero and constant variance σ^2.

Now suppose that we fit an OLS regression on *only* the predictors in $\mathbf{X}_1$. Let

$$\hat{\beta}_1^{(1)} = (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{Y}$$

be the resulting estimate of the coefficient vector β_1 and let

$$\hat{\mathbf{e}}^{(1)} = \mathbf{Y} - \hat{\mathbf{Y}}^{(1)}$$

be the corresponding residual vector, where $\hat{\mathbf{Y}}^{(1)} = \mathbf{X}_1 \hat{\beta}_1^{(1)}$.

a. Show that

$$E(\hat{\beta}_1^{(1)} \,|\, \mathbf{X}_1, \mathbf{X}_2) = \beta_1 + (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{X}_2 \beta_2 \ .$$

Under what conditions on $\mathbf{X}_1$ and $\mathbf{X}_2$ is $\hat{\beta}_1^{(1)}$ unbiased for estimating β_1, i.e., $E(\hat{\beta}_1^{(1)} \,|\, \mathbf{X}_1, \mathbf{X}_2) = \beta_1$?

b. Show that

$$E(\hat{\mathbf{e}}^{(1)} \,|\, \mathbf{X}_1, \mathbf{X}_2) = (I - H^{(1)}) \mathbf{X}_2 \beta_2 \ ,$$

where I is the $n \times n$ identity matrix and $H^{(1)} = \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T$ is the 'hat matrix' for regression on $\mathbf{X}_1$. Under what conditions on $\mathbf{X}_1$ and $\mathbf{X}_2$ are the residuals centered on zero, i.e., $E(\hat{\mathbf{e}}^{(1)} \,|\, \mathbf{X}_1, \mathbf{X}_2) = 0$?

c. Is it possible for both the estimator $\hat{\beta}_1^{(1)}$ to be unbiased and the residuals $\hat{\mathbf{e}}^{(1)}$ to be centered? Is it possible for the estimator to be biased, but the residuals centered? Explain your answers.

(NOTE: You may provide purely mathematical conditions in answering the second parts of problems 2(a) and 2(b), but these must be accompanied by a verbal explanation as well.)