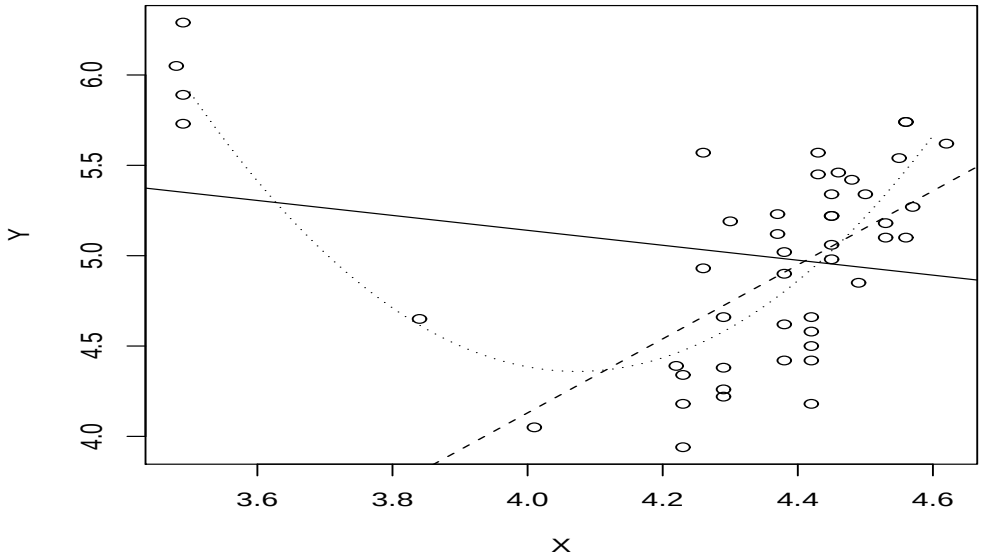
Qualifying Exam: CAS MA 575.

Boston University, Spring 2007

1. The figure below shows a plot of

$$Y = \log(\text{Light Intensity}) \quad \text{versus} \quad X = \log(\text{Surface Temperature}) \ ,$$

based on measurements for 47 stars in a certain star cluster. The goal of the study under which the data were obtained was to characterize the relationship between light intensity and surface temperature.



(a) Consider the solid line and the dotted curve. (Ignore the dashed line for the moment.) The line shows the OLS fit to a linear model of the form $E[Y|X] = \beta_0 + \beta_1 X$, while the curve shows the OLS fit to a quadratic model of the form $E[Y|X] = \beta_0 + \beta_1 X + \gamma X^2$.

Let $\beta = (\beta_0, \beta_1)^T$ and let $\hat{\beta}^{Lin}$ be the OLS estimate of β when the *linear* model is fit to the data. Suppose, however, that the quadratic model is in fact true. Provide a concise expression for $E[\hat{\beta}^{Lin}]$ under the quadratic model. Is $\hat{\beta}^{Lin}$ unbiased or biased? If biased, what is its bias?

(b) Suppose that you wish to test whether a linear model is adequate, or whether a quadratic model is more appropriate. State an appropriate pair of null and alternative hypotheses for this problem. Provide an expression for an appropriate F-statistic, and state the distribution of that statistic under the assumption of normal (i.e., Gaussian) errors. (NOTE: Be careful to specify *precisely* each of the components in your F-statistic.)

(c) Now consider the solid line and the dashed line in the figure. The solid line is, as described above, the OLS fit of a linear model $E[Y|X] = \beta_0 + \beta_1 X$. It is based on data from all 47 stars. The dashed line corresponds to an OLS fit of the same linear model, but without the four data points in the upper left-hand corner of the figure (i.e., without those points for which $X < 3.6$).

Comment on the degree of (i) outlying-ness, (ii) leverage, and (iii) influence of the four points in the upper left-hand corner, based upon the evidence presented by the two lines shown. Justify your answer through appropriate description of the likely values of the statistics t_i, h_{ii}, and D_i. (That is, the outlier t-test value, the hat-matrix entry, and Cook's distance.)

(d) Just based on visual inspection of the plots, comment on the appropriateness of the three models shown. Which would you suggest to the astronomers as being most appropriate? What question(s) might you have for the astronomers?

2. Consider a regression model of the form

$$y_i = \beta x_i + e_i \ ,$$

for $i = 1, \ldots, n$, where

$$E[e_i|x_i] = 0 \qquad \text{and} \qquad \text{Var}(e_i|x_i) = \sigma^2 x_i^2 \ ,$$

for $\sigma^2 > 0$ unknown. This is a model specifying regression through the origin, with nonconstant error variance increasing quadratically in the explanatory variable x_i.

(a) Set up a weighted-least squares criterion for estimation of β i.e., something of the form

$$\text{RSS}(\beta) = \sum_{i=1}^{n} w_i (y_i - \beta x_i)^2 \ ,$$

for appropriate choice of weights w_i. Derive an expression for the value $\hat{\beta}^{WLS}$ that minimizes this criterion. Simplify your expression as much as possible.

(b) Derive an expression for the variance $\text{Var}(\hat{\beta}^{WLS})$ of your estimator $\hat{\beta}^{WLS}$.

(c) Suppose you wish to test the null hypothesis $H_0 : \beta = 0$ against the alternative hypothesis $H_1 : \beta \neq 0$. Provide an expression for an appropriate F-statistic, and state the distribution of that statistic under the assumption of normal (i.e., Gaussian) errors e_i. (NOTE: Be careful to specify *precisely* each of the components in your F-statistic.)

(d) The decision to use a nonconstant variance is itself part of the process of model specification. How might you assess the adequacy of the choice $\text{Var}(e_i|x_i) \propto x_i^2$, given measurements $\{(x_i, y_i)\}_{i=1}^n$.