Assignment 3 — Due Oct 18, 2018

1. Consider the model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ where $\varepsilon_i \sim$ ind $N(0, \sigma_i^2)$ for observations $i = 1, \ldots, n$. Suppose $X_i > 0$ and $\sigma_i^2 = \sigma^2 X_i$. Let $\tilde{Q}(\boldsymbol{\beta}) = \sum_{i=1}^n \sigma_i^{-2}\{Y_i - (\beta_0 + \beta_1 X_i)\}^2$ denote the weighted error sum of squares. Define

   $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$: $n \times 1$ vector of response variables

   $\boldsymbol{X}$: $n \times 2$ design matrix with 1's in the first column and $(X_1, \ldots, X_n)'$ in the second column.

   $\boldsymbol{\beta} = (\beta_0, \beta_1)'$: 2×1 vector of regression coefficients

   (a) Let $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_0, \tilde{\beta}_1)'$ denote the weighted least squares estimates of $\boldsymbol{\beta}$. Derive the distribution (i.e., the type of distribution, mean vector, and the variance-covariance matrix) of $\tilde{\boldsymbol{\beta}}$.

   (b) Despite the weighted variances, derive the ordinary least squares estimates of $\boldsymbol{\beta}$ by minimizing the (unweighted) error sum of squares $Q(\boldsymbol{\beta}) = \sum_{i=1}^n \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$ all in matrix terms.

   (c) Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)'$ denote the (ordinary) least squares estimates of $\boldsymbol{\beta}$ you found in part (b), and let $\widehat{\boldsymbol{e}} = (\widehat{e}_1, \ldots, \widehat{e}_n)' = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ denote the raw residuals under ordinary least squares estimates. Find the distribution of $\widehat{\boldsymbol{\beta}}$ and the variance of $\widehat{e}_i$ for $i = 1, \ldots, n$.

   (d) Draw a connection between $Var(\widehat{\beta}_1)$ and $Var(\tilde{\beta}_1)$.

2. A chemist studied the concentration (Y) of a solution over time (X). Fifteen (15) identical solutions were prepared. The 15 solutions were randomly divided into five sets of three, and the five sets were measured, respectively, after 1, 3, 5, 7, and 9 hours. The data file is `solution.txt`. `online submission`

   (a) Identify ingredients of the experiment/study including population vs. sample, study/experimental units, and whether cause-and-effect relationships can be established.

   (b) (Hand calculation with calculator) Perform a simple linear regression analysis with concentration (Y) as the response variable and time (X) as the explanatory variable. Obtain the regression coefficients estimates along with their standard errors, an unbiased estimate of error variance, and the coefficient of determination R^2.

   (c) (Hand calculation with calculator) Perform a hypothesis test to determine whether there is evidence that the mean concentrations are different for the different hours since the solutions are prepared. Indicate the assumptions underlying the test.

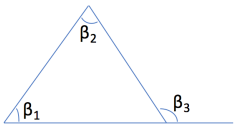
   (d) Indicate the model underlying the regression analysis in (b) and assess the model assumptions by using suitable graphical techniques. How reasonable are the assumptions? What remedial measures are desirable, if any?

   (e) What kind of variable transformations would you recommend to the chemist? Give reasoning.

   (f) Repeat (b) but now with the concentration after the transformation in (e).

   (g) Compare the results in (b) and (f).

3. Suppose we wish to measure the three angles β_1, β_2, and β_3 as depicted in the diagram below.



Elementary geometry shows that $\beta_1 + \beta_2 = \beta_3$. Suppose, as a check on the accuracy of the results, we decide to measure all three angles, with measurement error. Let b_j be the actual measurement for β_j, $j = 1, 2, 3$. Due to measurement error, $b_1 + b_2$ might not be equal to b_3. Assume that the measurement errors are independent and follow normal distribution with mean 0 and variance σ^2. Formulate this as a multiple linear regression model, and derive the least squares estimates $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$, with their standard deviations.

4. Consider the multiple regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots \beta_{p-1} X_{p-1} + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Suppose σ^2 is unknown and consider n independent observations from the model. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ denote the observed responses; let $\boldsymbol{X} = [\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{p-1}]$ denote the $n \times p$ design matrix, where $\boldsymbol{x}_i$ corresponds to the $(i+1)^{th}$ column of $\boldsymbol{X}$; and let $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})'$ denote the vector of regression coefficients. Assume that $\text{rank}(\boldsymbol{X}) = p$. Let $\widehat{\beta}_j$ denote the least squares estimate of β_j, and define $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_{p-1})'$.

   (a) Suppose $\boldsymbol{c} = (c_0, \ldots, c_{p-1})'$ is a vector of known constants. What is the distribution of $\sum_{j=0}^{p-1} c_j \widehat{\beta}_j$? Justify your answer, and find its mean and variance using matrix notation.

   (b) Consider the hypothesis $H_0 : \sum_{j=0}^{p-1} c_j \beta_j = h$ and $H_1 : \sum_{j=0}^{p-1} c_j \beta_j \neq h$, where h is a given constant. Explain how to test theses hypotheses at significance level α. Construct a suitable test statistic, find its distribution, and specify the rejection region.

   (c) Suppose we wish to predict the value of a future observation Y_{n+1}. Let $\boldsymbol{z} = (1, z_1, \ldots, z_{p-1})'$ denote its corresponding vector of predictor variables (i.e., $X_i = z_i$, for $i = 1, \ldots, p-1$), and consider the prediction $\widehat{Y}_{n+1} = \boldsymbol{z}\widehat{\boldsymbol{\beta}}$. Find the distribution of $Y_{n+1} - \widehat{Y}_{n+1}$.

   (d) Show that the MSE (mean square error) of the prediction $\widehat{Y}_{n+1} = \boldsymbol{z}\widehat{\boldsymbol{\beta}}$ is strictly greater than σ^2.

   (e) Given the vector $\boldsymbol{z}$ of predictor variables for the future observation Y_{n+1}, find an interval $\mathcal{I}$ such that $\mathbb{P}(Y_{n+1} \in \mathcal{I}) = 1 - \alpha$.

   (f) Suppose an additional predictor variable X_p is added to the model to obtain

$$Y = \gamma_0 + \gamma_1 X_1 + \cdots \gamma_{p-1} X_{p-1} + \gamma_p X_p + \varepsilon.$$

   Suppose that the augmented design matrix $\widetilde{\boldsymbol{X}} = [\boldsymbol{X}, \boldsymbol{x}_p]$ has rank $p+1$. Let $\widehat{\boldsymbol{x}}_p = a_0 \boldsymbol{x}_0 + \cdots + a_{p-1} \boldsymbol{x}_{p-1}$ denote the least-squares projection of $\boldsymbol{x}_p$ onto the subspace of $\mathbb{R}^n$ spanned by the columns of $\boldsymbol{X}$. Find an expression for the residual vector $\boldsymbol{r}_p = \boldsymbol{x}_p - \widehat{\boldsymbol{x}}_p$ in terms of $\boldsymbol{X}$ and $\boldsymbol{x}_p$.

   (g) Expression the least squares estimates $\widehat{\gamma}_0, \ldots, \widehat{\gamma}_p$ in terms of only $\widehat{\beta}_0, \ldots, \widehat{\beta}_{p-1}, a_0, \ldots, a_{p-1}, \boldsymbol{r}_p$, and $\boldsymbol{Y}$.