Department of Statistics, The Chinese University of Hong Kong
STAT 5102 Regression in Practice (Term 1, 2018–19)

Assignment 1 · due on 8th October 2018 (Mon)

1. Consider the simple linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where $E(\epsilon) = 0, Var(\epsilon) = \sigma^2$ and ϵ uncorrelated.

(a) Show that $Cov(\bar{y}, b_1) = 0$, where b_0 and b_1 are the least-squares estimates for β_0 and β_1 respectively with $\bar{y} = n^{-1} \sum_{i=1}^n y_i$.

(b) Using (a), or otherwise, show that $Cov(b_0, b_1) = -\bar{x}\sigma^2 / S_{XX}$, where $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ and $S_{XX} = \sum_{i=1}^n (X_i - \bar{X})^2$.

(c) Do your calculations in (a) and (b) require the normality assumption on ϵ?

2. [Optional] Suppose that we have fit the straight-line regression model $\hat{Y} = b_0 + b_1 X_1$, but the response is affected by a second variable X_2 such that the true regression function is

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

(a) Show that $E(b_1) = \sum_{i=1}^n c_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})$, where $c_i = \dfrac{X_{i1} - \bar{X}_1}{S_{X_1 X_1}}$. Here X_{i1} means the i-th observation of X_1, and $S_{X_1 X_1} = \sum_{i=1}^n (X_{i1} - \bar{X}_{i1})^2$.

(b) Show that the bias of b_1 is $\frac{\sum_{i=1}^n (X_{i1} - \bar{X}_1) X_{i2}}{S_{X_1 X_1}}$, *i.e.* $E(b_1) = \beta_1 + \frac{\beta_2 \sum_{i=1}^n (X_{i1} - \bar{X}_1) X_{i2}}{S_{X_1 X_1}}$.

[Hint: For part (a), make use of the fact that $b_1 = \sum_{i=1}^n c_i Y_i$, then take expectation on both sides and apply the condition that $E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$. For part (b), you can use the result in section 2.1 of supplementary note 1, such as $\sum_{i=1}^n c_i = 0$.]

3. [Height and weight data] The table below (see also `htwt.txt`) gives Ht = height in centimetres and Wt = weight in kilograms for a sample of $n = 10$ 18-year-old girls. Interest is in predicting weight from height.

Ht	169.6	166.8	157.1	181.1	158.4	165.6	166.7	156.5	168.1	165.3
Wt	71.2	58.2	56	64.5	53	52.4	56.8	49.2	55.6	77.8

(a) Produce a scatterplot of Wt on the vertical axis versus Ht on the horizontal axis. On the basis of this plot, is a simple linear regression model capable of modelling the given dataset? Explain briefly.

(b) Evaluate $\bar{X}$, $\bar{Y}$, S_{XX}, S_{YY} and S_{XY} where these quantities are defined in the lecture notes. Compute the estimates of the slope and the intercept for the regression of Y on X. Draw the fitted line on your scatterplot.

(c) Obtain the estimate of σ^2 and find the estimated standard errors of b_0 and b_1. Also, find the estimated covariance between b_0 and b_1. Compute the t-tests for the hypotheses that $\beta_0 = 0$ and that $\beta_1 = 0$ and find the appropriate p-values using two-sided tests.

(d) Obtain the analysis of variance table and F-test for regression. Does the F-test yield the same conclusion due to the t-test on $\beta_1 = 0$ performed in (c)? Do you notice any relation between these two tests?

4. [Old Faithful] The data in the data file `oldfaith.txt` include information about eruptions of Old Faithful Geyser during October 1980. Variables are the Duration in seconds of the current eruption, and the Interval, the time in minutes to the next eruption. The data were collected by volunteers R. Hutchinson. Apart from missing data for the period from midnight to 6 AM, this is a complete record of eruptions for that month.

   Old Faithful Geyser is an important tourist attraction, with up to several thousand people watching it erupt on pleasant summer days. The park service uses data like these to obtain a prediction equation for the time to the next eruption.

   (a) Use simple linear regression methodology to obtain a prediction equation for interval from duration. Summarise your results in a way that might be useful for the nontechnical personnel who staff the Old Faithful Visitor's Centre.

   (b) Construct a 95% confidence interval for $E(\text{interval} \mid \text{duration} = 250)$.

   (c) An individual has just arrived at the end of an eruption that lasted 250 seconds. Give a 95% confidence interval for the time the individual will have to wait for the next eruption.

5. Under what situation will the estimates b_0 and b_1 not be well defined? What is the meaning for this mathematical condition?