October 18, 2019

Statistics 601 Assignment 2
Due October 18, 2019

Homework Assignment Policy and Guidelines

1. You are expected to complete some homework problems independently (or in a group) without discussing with other students (or other groups) before they are due. You may seek clarification of these problems from your TA and instructor. Such homework problems will be marked `independent work`.

2. Unless otherwise stated in a problem, keep the R code and/or R output for all the relevant problems in a **well organized appendix**. Please streamline and briefly document the R code to be clear and concise.

1. Let $X_1, \ldots, X_n$ be an i.i.d. sample from a uniform distribution $U(0, \theta)$, where $\theta > 0$.
   - Find the method-of-moments estimator of θ.
   - Find the MLE of θ.
   - Suppose we observe $x_1 = 3, x_2 = 5, x_3 = 6$, and $x_4 = 18$. Find the moment-of-moments estimate and MLE of θ. Which estimate is better?

2. An experimenter was interested in the relationship between temperature and heart rate in the common grass frog. The temperature was manipulated in 2-degree increments ranging from 2 to 18 C with heart rates recorded at each temperature in beats per minute. (A different, randomly selected, frog was used at each temperature.)

Temp (X):	2	4	6	8	10	12	14	16	18
Heart rate (Y):	5	11	10	13	22	23	30	28	32

   (a) Draw a scatter plot of heart rate versus temperature, superimposed by the least squares fitted line. Comment on the plot.

   (b) Indicate the model underlying a simple linear regression of heart rate on temperature. Interpret the terms in the context of the study.

   (c) For the model in (b), what is the underlying population or the set of underlying populations?

   (d) (Hand calculation with calculator) Compute the least squares estimates of slope and intercept viewing heart rate as the response variable and temperature as the explanatory variable. Interpret the results in the context of the study.

   (e) (Hand calculation with calculator) At temperature 9 C, estimate the population mean heart rate corresponding to this temperature. Interpret the results in the context of the study.

   (f) (Hand calculation with calculator) At temperature -2 C, estimate the population mean heart rate corresponding to this temperature. Interpret the results in the context of the study.

   (g) (Hand calculation with calculator) Provide a suitable estimate of the population error variance in the model. Interpret the results in the context of the study.

3. Conduct a simulation study to evaluate the sampling distributions of $\hat{\beta}_0$, $\hat{\beta}_1$, MSE (i.e., residual mean square, $\hat{\sigma}^2$) using the least squares method.

(a) Let $\{X_i\} = \{2, 6, 10, 14, 18\}$. Generate one simulated sample of (X_i, Y_i) where $i = 1, 2, \ldots, 5$ under the true model $E(Y|X) = 10 + 4X$ and $\varepsilon \sim N(0, 9)$. That is $\beta_0 = 10, \beta_1 = 4, \sigma^2 = 9$, and $n = 5$. Perform a simple linear regression analysis. Find estimated regression coefficients $\hat{\beta}_0, \hat{\beta}_1$ and residual mean square $\hat{\sigma}^2$. Also, compute a 95% confidence interval for $E(Y_h)$ when $X_h = 7$. Repeat this 100 times.

i. Draw one histogram each for the 100 $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$. What observations can you make about these histograms? How consistent are they with the theoretical results?

ii. What proportion of the 100 confidence intervals for $E(Y_h)$ when $X_h = 7$ cover $E(Y_h|X_h)$?

(b) Repeat (a) but this time with $\{X_i\} = \{2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$. How do the results compare with those in (a)?

(c) Repeat (a) but this time with $\{X_i\} = \{2, 6, 10, 14, 18, 22, 26, 30, 34, 38\}$. How do the results compare with those in (a) and (b)?

(d) Repeat (a) but this time assume that the model $E(Y|X) = 10 + 4X + 0.1X^2$. How do the results compare with those in (a)?

(e) Repeat (a) but this time assume that the model $E(Y|X) = 10 + 4X + 0.2X^2$. How do the results compare with those in (a) and (d)?

4. Continue to work on the frog data in Problem 2. Let β_0 and β_1 denote the intercept and the slope in the simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ where $\varepsilon_i \sim$ iid $N(0, \sigma^2)$ for $i = 1, 2, \ldots, n$. **Unless otherwise stated, use hand calculation with calculator for all parts of this problem.**

(a) Perform a T test for $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$ and construct a 95% confidence interval for β_1.

(b) Assume the normality assumption is not valid, and use the Bootstrapping to obtain 95% percentile interval for the β_1. Then compare your empirical interval with that interval in (a).

(c) Conduct a power analysis for a study in the future where the test of interest will be $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$. Assume the same set of $X = 2, 4, \ldots, 16, 18$. Also, assume $\alpha = 0.05$ and $\sigma = 2.5$. Provide the rejection region and compute the power at $\beta_1 = 0, \pm 0.5$. **Use computer** to compute the power at $\beta_1 = \pm 1.0, \pm 1.5$.

(d) Repeat (c) but replace with $X = 4, 8, 12, 16$. **Use computer** to draw two power curves, one each for (c) and (d), and compare the results.

(e) Perform a T test for $H_0 : \beta_0 = 0$ vs $H_A : \beta_0 \neq 0$ and construct a 95% confidence interval for β_0.

(f) Provide a 95% confidence interval for the error variance σ^2.

(g) Provide an appropriate 95% interval for the population mean heart rate at temperature 9 C.

(h) Provide an appropriate 95% interval for the heart rate of a frog to be drawn at random such that the temperature will be manipulated to be -2 C.

(i) Provide the coefficient of determination R^2.

(j) Provide the sample correlation coefficient $\hat{\rho}$ and construct a 95% confidence interval for ρ. In addition, describe the model underlying the inference in this part.

5. 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 $Var(\tilde{\beta}_1)$.