625.661 Statistical Models and Regression

Module 3-4 Assignment

H.M. James Hung

Please complete all the following problems.

Make sure that you state the assumptions for your analyses or proof/derivation steps. If applicable, you can use math/stat software to produce statistical results.

1. In a typical multiple linear regression model where x_1 and x_2 are two regressors. The expected value of the response variable y given x_1 and x_2 is denoted by $E(y \mid x_1, x_2)$.

   a) As the value of x_1 increases, the magnitude of change in the value of $E(y \mid x_1, x_2)$ will not depend on the value of x_2 . Write down the multiple linear regression model with assumptions for this scenario.

   Let $\boldsymbol{E(y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2}$

   Then, $\boldsymbol{E(y \mid x_1 + c, x_2) - E(y \mid x_1, x_2) =}$

   $$\boldsymbol{\beta_0 + \beta_1(x_1 + c) + \beta_2 x_2 - (\beta_0 + \beta_1 x_1 + \beta_2 x_2) = \beta_1 c}\ ,$$

   which does not depend on the value of $\boldsymbol{x_2}$ **.**

   b) As the value of x_1 increases, the magnitude of change in the value of $E(y \mid x_1, x_2)$ will depend on the value of x_2 . Write down the multiple linear regression model with assumptions for this scenario.

   Let $\boldsymbol{E(y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma x_1 x_2}$

Then, $E(y \mid x_1 + c, x_2) = \beta_0 + \beta_1(x_1 + c) + \beta_2 x_2 + \gamma(x_1 + c)x_2$

Thus, $E(y \mid x_1 + c, x_2) - E(y \mid x_1, x_2) = \beta_1 c + \gamma c x_2$ **, which depends on the value of** x_2 **.**

2. Do Problem 3.24, page 126 of Textbook

The regression sum of squares corrected for intercept is
$SS_R = \hat{\beta}' X' y - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$.
Now $\hat{\beta}' X' y = y' X (X'X)^{-1} X' y = y' H y$
and $\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy$
Thus,

$$\sum_{i=1}^{n} \hat{y}_i^2 = \hat{y}'\hat{y} = y'H'Hy = y'HHy = y'Hy$$

$$SS_R = \sum_{i=1}^{n} \hat{y}_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2$$

3. Do Problem 3.27, page 127 of Textbook

$$Var(\hat{y}) = Var(X\hat{\beta}) = X Var(\hat{\beta}) X' = \sigma^2 X (X'X)^{-1} X' = \sigma^2 H \ .$$

4. Do Problem 3.38, page 128 of Textbook

We have a typical multiple linear model, $y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$ **, where** x**'s are non-random and the random errors** ε_i **are statistically independent and have mean zero and variance** σ^2**,** i **=1, …,** n**.**

Denote the vector $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$ **. The data matrix X and** y **are defined as given on page 72 of the Textbook.**

Thus, the fitted $\hat{y}_i = \mathbf{x}_i{}'\hat{\beta}$ **, where** $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$ **as in (3.13) on page 73. Hence the variance**

$$Var(\hat{y}_i) = Var(\mathbf{x}_i{}'\hat{\beta}) = \mathbf{x}_i' Var(\hat{\beta}) \mathbf{x}_i = \sigma^2 \mathbf{x}_i' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i$$

$$= \sigma^2 tr(\mathbf{x}_i' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i) = \sigma^2 tr(\mathbf{x}_i' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i) = \sigma^2 tr((\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i') \ .$$

$$\sum_{i=1}^{n} Var(\hat{y}_i) = \sigma^2 \sum_{i=1}^{n} tr((\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i') = \sigma^2 tr((\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i') =$$
$$\sigma^2 tr((\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i') = \sigma^2 tr\left((\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\right) = \sigma^2 tr(\mathbf{I}_p),$$

where $\mathbf{I}_p$ is the identity matrix of dimension p. Since $tr(\mathbf{I}_p) = p$,

$$\sum_{i=1}^{n} Var(\hat{y}_i) = p\sigma^2.$$

5. In a multiple linear regression model, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where y is a response variable, x_1 and x_2 are non-random regressors, and ε is a random error, the parameter β_2 is nonzero. Suppose that n subjects give data on (y, x_1, x_2) to generate the ordinary least-squares (OLS) estimators of all three β parameters in this model. We then fit the same data to the simple linear regression model $y = \beta_0 + \beta_1 x_1 + \varepsilon$.

   a) Create a hypothetical data set and perform regression analysis to compare the OLS estimate of β_1 in the regression model including x_2 with the OLS estimate of β_1 in the regression model excluding x_2. What have you learned?
   We should be able to see that the two OLS estimates are different. Assumptions: In the hypothetical data, the random errors of the regression models are assumed to be independent with mean zero and constant variance.

   b) Discuss with mathematical arguments whether the OLS estimators of β_1 from the two model fittings are equal. If not, discuss with mathematical arguments the condition(s) under which the two OLS estimators of β_1 are equal (Note: β_2 is nonzero).

Assumptions: In the regression models, the random errors are independent with mean zero and constant variance.

In the model that contains intercept, regressors x1, and x2, the least-squares estimator of β_1 will have to be obtained from the following steps. First,

$$\boldsymbol{X'X} = \begin{bmatrix} \boldsymbol{n} & \sum \boldsymbol{x_{1i}} & \sum \boldsymbol{x_{2i}} \\ \sum \boldsymbol{x_{1i}} & \sum \boldsymbol{x_{1i}^2} & \sum \boldsymbol{x_{1i}x_{2i}} \\ \sum \boldsymbol{x_{2i}} & \sum \boldsymbol{x_{1i}x_{2i}} & \sum \boldsymbol{x_{2i}^2} \end{bmatrix} .$$

We can easily see that if $\sum \boldsymbol{x_{1i}} = \boldsymbol{0}$, $\sum \boldsymbol{x_{2i}} = \boldsymbol{0}$, and $\sum \boldsymbol{x_{1i}x_{2i}} = \boldsymbol{0}$, this matrix is a diagonal matrix and it inverse is

$$(\boldsymbol{X'X})^{-1} = \begin{bmatrix} \boldsymbol{1/n} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{1/\sum x_{1i}^2} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{1/\sum x_{2i}^2} \end{bmatrix} .$$

Because $\boldsymbol{X'Y} = \begin{bmatrix} \sum \boldsymbol{y_i} \\ \sum \boldsymbol{x_{1i}y_i} \\ \sum \boldsymbol{x_{2i}y_i} \end{bmatrix}$, under the above three conditions, we have $\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}} = \sum \boldsymbol{x_{1i}y_i} / \sum \boldsymbol{x_{1i}^2}$, which is identical to the least-squares estimator of β_1 from the simple linear regression model $y = \beta_0 + \beta_1 x_1 + \varepsilon$ (i.e., without x2), because $\sum \boldsymbol{x_{1i}} = \boldsymbol{0}$ implies $\overline{\boldsymbol{x}}_{\boldsymbol{1}} = \boldsymbol{0}$. Without these three conditions, the ordinary least-squares estimator of β_1 from the multiple linear regression model may not be equal to that of β_1 from the simple linear regression model $y = \beta_0 + \beta_1 x_1 + \varepsilon$.

6. Use any math/stat software (e.g., *www.*numbergenerator*.org/*randomnumbergenerator) of your choice to find a random number generator to randomly select 22 rows of Table B.3 (page 556) used in Problem 3.5 (page 122) of Textbook and then do (a), (b), (c), (d), (e), (f), (g).