625.661 Statistical Models and Regression

Module 1-2 Assignment

H.M. James Hung

Please complete all the problems.

Do not use any math/stat software to produce statistical results, but you can use any math/stat software to generate the percentile of normal, t, F, chi-square distribution or do basic mathematical calculations. State assumptions in your analyses or analytic derivations.

1. In a simple linear regression analysis, n *independent* paired data (y_1, x_1), …., (y_n, x_n) are fitted to the model

$$y_i = \beta_1(x_i - \mu) + \varepsilon_i \ , \ \ i = 1, \dots, n \, ,$$

where x is the only non-random independent variable (or so-called regressor), μ is a known real number, and ε is the random error that has mean zero and unknown constant variance σ^2. Before the data for $(y\,, x)$ are available, we need to construct estimators for the parameters.

a) Construct the ordinary least squares (OLS) estimator of β_1 .

The OLS estimator $\widehat{\beta}_1$ is to minimize $\sum_{i=1}^{n} (y_i - \beta_1(x_i - \mu))^2$.

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\mu)y_i}{\sum_{j=1}^{n}(x_i-\mu)^2} = \sum_{i=1}^{n} d_i \, y_i$$

where $d_i = (x_i - \mu)/\sum_{j=1}^{n}(x_i - \mu)^2$ **.**

The OLSE is a linear combination of y_i .

b) Construct the variance of the OLS estimator of β_1 in a) and construct an <u>unbiased</u> estimator of this variance.

$Var(\widehat{\boldsymbol{\beta}}_1) = Var(\sum_{i=1}^{n} d_i\, y_i) = \sigma^2 \sum_{i=1}^{n} d_i^2$ **because y's are statistically independent.**

σ^2 **can be unbiased estimated by** $\widehat{\sigma}^2 = \frac{1}{(n-1)}\sum_{i=1}^{n} \left(y_i - \widehat{\beta}_1(x_i-\mu)\right)^2$**,**
following the same arguments as in (2.16) – (2.19) in the Textbook.
That is, let $e_i = y_i - \widehat{y}_i$ **.**

$SS_T = \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n}(\widehat{y}_i + [y_i - \widehat{y}_i])^2 = \sum_{i=1}^{n}(\widehat{y}_i + e_i)^2 =$
$\sum_{i=1}^{n} \widehat{y}_i^2 + \sum_{i=1}^{n} e_i^2$ **, because** $\sum_{i=1}^{n} \widehat{y}_i e_i = 0$ **.**

$E(SS_T) = \sum_{i=1}^{n} E(y_i^2) = \sum_{i=1}^{n}\{V(y_i) + [E(y_i)]^2\} = n\sigma^2 +$
$\beta_1^2 \sum_{i=1}^{n}(x_i-\mu)^2$

$\sum_{i=1}^{n} \widehat{y}_i^2 = \widehat{\beta}_1^2 \sum_{i=1}^{n}(x_i-\mu)^2$

Now $E(\widehat{\beta}_1^2) = V(\widehat{\beta}_1) + [E(\widehat{\beta}_1)]^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\mu)^2} + \beta_1^2$
Thus, $E\left(\sum_{i=1}^{n} \widehat{y}_i^2\right) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(x_i-\mu)^2$

Therefore, $E\{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2\} = (n-1)\sigma^2$**.**

That is, $E(\widehat{\sigma}^2) = \sigma^2$**.**

Thus $Var(\widehat{\beta}_1)$ **can be unbiasedly estimated by**
$\widehat{V}(\widehat{\beta}_1) = \widehat{\sigma}^2 \sum_{i=1}^{n} d_i^2 = \frac{\widehat{\sigma}^2}{\sum_{j=1}^{n}(x_i-\mu)^2}$.

2. In Problem 1, add the intercept term β_0 to the model. Then do a) and b).

Let $z = x - \mu$. The linear model becomes $y = \beta_0 + \beta_1 z + \varepsilon$.

Then, using what we learn from the Text or module 1 video, we have OLS estimators:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})y_i}{\sum_{i=1}^{n}(z_i - \bar{z})^2} \quad , \quad \textbf{because } \bar{z} = \bar{x} - \mu \,.$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{z}$$

$$Var(\hat{\beta}_1) = Var\left(\sum_{i=1}^{n} d_i\, y_i\right) = \sigma^2 \sum_{i=1}^{n} d_i^2$$

where $d_i = (z_i - \bar{z}) / \sum_{i=1}^{n}(z_i - \bar{z})^2$

σ^2 **can be unbiased estimated by** $\hat{\sigma}^2 = \frac{1}{(n-2)} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 z_i\right)^2$

Thus $Var(\hat{\beta}_1)$ can be unbiasedly estimated by

$$\hat{V}(\hat{\beta}_1) = \hat{\sigma}^2 \sum_{i=1}^{n} d_i^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2}$$

$$\hat{V}(\hat{\beta}_0) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{z}^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2}\right)$$

3. In Problem 1, the μ is a real number but the value is <u>unknown</u>. Please do a) and b).

The model $y_i = \beta_1(x_i - \mu) + \varepsilon_i$, $i = 1, \ldots, n$, **can be expressed as**

$$y_i = -\beta_1 \mu + \beta_1 x_i + \varepsilon_i \ , \quad i = 1, \ldots, n \,,$$

Because μ is unknown, define β_0 which is also unknown and a free parameter as μ . This resulted model is the same regression model as that covered in the Textbook.

4. Consider a regression model $y = \beta_0 + \beta_1 x + \epsilon$, where x is a non-random regressor. Discuss whether the ordinary least-squares estimator of the

slope β_1 is always unbiased and whether it always has the smallest variance than **any** estimator of β_1, irrespectively of what the value of β_0 is. State assumptions in your discussion. Be careful about the word "**any**".

The OLS estimator of a regression coefficient in simple linear regression model is a linear combination of the y observations and it is unbiased for that regression coefficient. That is,

$$\widehat{\boldsymbol{\beta}}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sum_{i=1}^{n} \boldsymbol{c_i y_i} \quad \text{, where } \boldsymbol{c_i} = (\boldsymbol{x_i}-\bar{\boldsymbol{x}})/\sum_{i=1}^{n}(\boldsymbol{x_i}-\boldsymbol{x})^2$$

$$\widehat{\boldsymbol{\beta}}_0 = \bar{\boldsymbol{y}} - \widehat{\boldsymbol{\beta}}_1\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y_i} - \bar{\boldsymbol{x}}\sum_{i=1}^{n} \boldsymbol{c_i y_i} = \sum_{i=1}^{n} w_i \boldsymbol{y_i}\,,$$

where $w_i = \frac{1}{n} - \bar{\boldsymbol{x}}\boldsymbol{c_i}$

The essence of G-M Theorem is that for that regression coefficient, among all possible <u>unbiased</u> estimators which are <u>linear combinations of the y observations</u>, the OLS estimator is the "best" estimator in the sense of smallest variance (i.e., the most precise). However, there may be an estimator that is not a linear combination of the y observations or not unbiased for that regression coefficient but has a smaller variance than the OLS estimator. Of course, the critical assumptions are: 1) the regression model is correct, 2) the variance of y is constant across subjects or items.

5. Use any math/stat software (e.g., www.**numbergenerator**.org/**randomnumbergenerator**) of your choice to find a random number generator to randomly select 15 rows of Table for Problem 2.18 (page 63-64) of Textbook and then do (a), (b), (c), (d). **State assumptions for all steps in your analyses.**