STAC67H: Regression Analysis

Fall, 2014

Instructor: Jabed Tomal

Department of Computer and Mathematical Sciences
University of Toronto Scarborough
Toronto, ON
Canada

October 9, 2014

Regression through Origin:

1. Sometimes the regression function is known to be linear and to go through the origin at $(0, 0)$.
2. Example 1: X is units of output and Y is variable cost, so Y is zero by definition when X is zero.
3. Example 2: X is the number of brands of beer stocked in a supermarket and Y is the volume of beer sales in the supermarket.

Regression through Origin:

Model

The normal error model for these cases is the same as regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ except that $\beta_0 = 0$:

$$Y_i = \beta_1 X_i + \epsilon_i$$

where:

1. β_1 is a parameter
2. X_i are known constants
3. ϵ_i are independent $N(0, \sigma^2)$.

Regression through Origin:

The regression function is

$$E(Y_i) = \beta_1 X_i$$

which is a straight line through the origin, with slope β_1.

Regression through Origin:

Inferences

The least squares estimator of β_1 is obtained by minimizing:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2$$

with respect to β_1.

Regression through Origin:

Inferences

The resulting normal equation is:

$$\sum_{i=1}^{n} X_i(Y_i - \beta_1 X_i) = 0$$

leading to the point estimator:

$$b_1 = \frac{\sum X_i Y_i}{\sum X_i^2}$$

Inferences

The maximum likelihood estimator is:

$$\hat{\beta}_1 = \frac{\sum X_i Y_i}{\sum X_i^2}$$

Regression through Origin:

Inferences

The fitted value $\hat{Y}_i$ for the *i*th case is:

$$\hat{Y}_i = b_1 X_i$$

Inferences

The ith residual is defined as the difference between the observed and fitted values:

$$e_i = Y_i - \hat{Y}_i = Y_i - b_1 X_i$$

Is $\sum_{i=1}^{n} e_i$ *zero*?

Regression through Origin:

Inferences

An unbiased estimator of the error variance σ^2 is:

$$s^2 = MSE = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-1} = \frac{\sum e_i^2}{n-1} = \frac{SSE}{n-1}$$

The reason for the denominator $n-1$ is that only one degrees of freedom is lost in estimating the single parameter in the regression function.

Regression through Origin:

Table: Confidence Limits for Regression through Origin.

Estimate of	**Estimated Variance**	**Confidence Limits**
β_1	$s^2\{b_1\} = \frac{MSE}{\sum X_i^2}$	$b_1 \pm ts\{b_1\}$
$E\{Y_h\}$	$s^2\{\hat{Y}_h\} = \frac{X_h^2 MSE}{\sum X_i^2}$	$\hat{Y}_h \pm ts\{\hat{Y}_h\}$
$Y_{h(new)}$	$s^2\{pred\} = MSE\left(1 + \frac{X_h^2}{\sum X_i^2}\right)$	$\hat{Y}_{h(new)} \pm ts\{pred\}$

Here, $t = t(1 - \alpha/2; n - 1)$

Cautions for Using Regression through Origin:

- Here, $\sum e_i \neq 0$. Thus, in a residual plot the residuals will usually not be balanced around the *zero* line.
- Here, $SSE = \sum e_i^2$ may exceed the total sum of squares $\sum(Y_i - \bar{Y})^2$. This can occur when the data form a curvilinear pattern or a linear pattern with an intercept away from the origin.
- The coefficient of determination R^2 has no clear meaning and may turn out to be negative.

Cautions for Using Regression through Origin:

- Evaluate the aptness of your regression model; the regression function may not be linear or the variance of the error terms may not be constant.
- It is generally a safe practice not to use regression-through-the-origin model ($Y_i = \beta_1 X_i + \epsilon_i$) and instead use the intercept regression model ($Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$).

Measurement Errors in Y

- Measurement errors could be present in the response variable Y.
- Consider a study of the relation between the time required to complete a task (Y) and the complexity of the task (X).
- The time to complete the task may not be measured accurately due to inaccurate operation of the stopwatch.

Measurement Errors in Y

- If the random measurement errors on Y are uncorrelated and unbiased, no new problems are created.
- Such random, uncorrelated and unbiased measurement errors on Y are simply absorbed in the model error term ϵ.
- The model error term always reflects the composite effects of a large number of factors not considered in the model, one of which now would be the random variation due to inaccuracy in the process of measuring Y.

Measurement Errors in X

- Measurement errors may be present in the predictor variable X, for instance, when the predictor variable is presure in a tank, temperature in an oven, speed of a production line, or reported age of a person.
- Interested in the relation between employees' piecework earnings and their ages.
- Let X_i be the true age and X_i^* be the reported age of the ith employee. Hence, the measurement error is:

$$\delta_i = X_i^* - X_i$$

Measurement Errors in X

- The regression model under study is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- We replace X_i by the observed X_i^*

$$Y_i = \beta_0 + \beta_1 (X_i^* - \delta_i) + \epsilon_i$$

- We rewrite the regression model as:

$$Y_i = \beta_0 + \beta_1 X_i^* + (\epsilon_i - \beta_1 \delta_i)$$

  One can consider this model as a simple linear regression model with predictor X_i^* and error $(\epsilon_i - \beta_1 \delta_i)$. But, here, X_i^* and $(\epsilon_i - \beta_1 \delta_i)$ are not independent.

Measurement Errors in X

- Let us assume that X_i^* is an unbiased estimator of X_i

$$E(\delta_i) = E(X_i^*) - E(X_i) = 0$$

- As usual, we assume that the error terms ϵ_i have expectation 0

$$E(\epsilon_i) = 0$$

- Let the measurement error δ_i and the model error ϵ_i are uncorrelated

$$\sigma\{\delta_i, \epsilon_i\} = E\{\delta_i, \epsilon_i\} - E\{\delta_i\}E\{\epsilon_i\} = E\{\delta_i, \epsilon_i\} = 0$$

Measurement Errors in X

$$\begin{aligned}
\sigma\{X_i^*, \epsilon_i - \beta_1\delta_i\} &= E\left\{\left[X_i^* - E(X_i^*)\right]\left[(\epsilon_i - \beta_1\delta_i) - E\{\epsilon_i - \beta_1\delta_i\}\right]\right\} \\
&= E\left\{\left(X_i^* - X_i\right)(\epsilon_i - \beta_1\delta_i)\right\} \\
&= E\left\{\delta_i(\epsilon_i - \beta_1\delta_i)\right\} \\
&= E\left\{\delta_i\epsilon_i - \beta_1\delta_i^2\right\} \\
&= -\beta_1\sigma^2\{\delta_i\}
\end{aligned}$$

This covariance is not zero whenever there is a linear regression relation between X and Y.

Measurement Errors in X

- If we assume that the response Y and the predictor X^* follow a bivariate normal distribution, then the conditional distribution $Y_i|X_i^*$, $i = 1, \cdots, n$, are independent normal with mean
- $$E\{Y_i|X_i^*\} = \beta_0^* + \beta_1^* X_i^*$$
- and variance
$$\sigma^2_{Y|X}.$$

Measurement Errors in X

- It can be shown that

$$\beta_1^* = \beta_1 \left[\sigma_X^2/(\sigma_X^2 + \sigma_Y^2)\right]$$

  where σ_X^2 is the variance of X and σ_Y^2 is the variance of Y.
- Hence, the least squares slope estimate from fitting Y and X^* is not an estimate of β_1, but is an estimate of $\beta_1^* \leq \beta_1$.
- If σ_Y^2 is small relative to σ_X^2, then the bias would be small; otherwise the bias may be substantial.

Measurement Errors in X

One of the approaches to deal with measurement errors in X is to use additional variables that are known to be related to the true value of X but not to the errors of measurement δ. Such variables are called *instrumental variables* because they are used as an instrument in studying the relation between X and Y.

Simple Linear Regression Model in Matrix Terms:

- Consider, the normal error regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad ; \; i = 1, \cdots, n$$

- We will not present any new results, but shall only state in matrix terms the results obtained earlier.

Simple Linear Regression Model in Matrix Terms:

- We write the limple linear regression model in matrix terms as following

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times 2}{\mathbf{X}} \underset{2\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}$$

  where

- $$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Simple Linear Regression Model in Matrix Terms:

- $$\underset{n\times 2}{\mathbf{X}} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

- $$\underset{2\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Simple Linear Regression Model in Matrix Terms:

- $$\underset{n\times 1}{\boldsymbol{\epsilon}} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Simple Linear Regression Model in Matrix Terms:

- The vector of expected values of the Y_i observations is

$$\underset{n\times 1}{E(\mathbf{Y})} = \underset{n\times 1}{\mathbf{X}\beta}$$

- The condition $E\{\epsilon_i\} = 0$ in matrix terms is:

$$\underset{n\times 1}{\mathbf{E}\{\epsilon\}} = \underset{n\times 1}{\mathbf{0}}$$

Simple Linear Regression Model in Matrix Terms:

- The condition that the error terms have constant variance σ^2 and that all covariances $\sigma\{\epsilon_i, \epsilon_j\}$ for $i \neq j$ are zero is expressed in matrix terms as following

$$\underset{n\times n}{\text{Var}\{\boldsymbol{\epsilon}\}} = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \underset{n\times n}{\mathbf{I}}$$

Simple Linear Regression Model in Matrix Terms:

- Thus the normal error regression model in matrix terms is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- where, $\boldsymbol{\epsilon}$ is a vector of independent normal random variables with $\mathbf{E}\{\boldsymbol{\epsilon}\} = \mathbf{0}$ and $\text{Var}\{\boldsymbol{\epsilon}\} = \sigma^2\mathbf{I}$.

Least Squares Estimation of Regression Parameters:

- The normal equations in matrix terms are

$$\underset{2\times 2}{\mathbf{X}'\mathbf{X}}\ \underset{2\times 1}{\mathbf{b}} = \underset{2\times 1}{\mathbf{X}'\mathbf{Y}}$$

- where, $\boldsymbol{b}$ is the vector of the least squares regression coefficients:

$$\underset{2\times 1}{\mathbf{b}} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

- The estimated regression coefficients are

$$\underset{2\times 1}{\mathbf{b}} = \underset{2\times 2}{\left(\mathbf{X}'\mathbf{X}\right)}^{-1} \underset{2\times 1}{\mathbf{X}'\mathbf{Y}}$$

Least Squares Estimation of Regression Parameters:

Exercise 5.6 Refer to **Airfreight breakage** Problem 1.21. Using matrix methods, find (1) $\mathbf{Y}'\mathbf{Y}$, (2) $\mathbf{X}'\mathbf{X}$, (3) $\mathbf{X}'\mathbf{Y}$, and (4) $\mathbf{b}$.

1. $$\mathbf{Y}'\mathbf{Y} = 2194$$
2. $$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 10 & 10 \\ 10 & 20 \end{bmatrix}$$
3. $$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 142 \\ 182 \end{bmatrix}$$
4. $$\mathbf{b} = \begin{bmatrix} 10.2 \\ 4.0 \end{bmatrix}$$