743- Regression and Time Series

Mamikon S. Ginovyan

The Multiple Regression Model-II

Inferences about Parameters

Inferences about Regression Unknown Parameters (b_i).

Statistical inferences about $b_i (i = \overline{0, k})$ are based on the properties of **point estimators** $\hat{b}_i$ for $b_i (i = \overline{0, k})$.

Recall that the model is given by equation:

$$Y = b_0 + b_1 x_1 + \mathbf{L} + b_k x_k + e = X \cdot b + e .$$

Inferences about Parameters

Recall properties of **point estimators:**

1) $E(\hat{b}_i) = b_i.$

2) $Var(\hat{b}_i) = c_{ii}s^2, i = \overline{0,k};$

$$C = (X'X)^{-1} = \|c_{ij}\|_{i,j=\overline{0,k}} .$$

3) $\hat{b}_i \sim N(b_i, c_{ii}s^2).$

4) $s^2 = \hat{s}^2 = \sum_{i=1}^{n} \hat{e}^2 / (n-k-1).$

Inferences about Parameters

It follows from 1) - 4) that the statistic

$$T = \frac{\hat{b}_i - b_i}{s_{\hat{b}_i}}, \qquad \text{where} \quad s_{\hat{b}_i} = s\sqrt{c_{ii}} \qquad (1)$$

has ***t*-distribution** with $(\boldsymbol{n-k-1})$ ***df***.

Inferences about Parameters

- **CI for** b_i.

For given a $(0 \leq a \leq 1)$ a $100(1 - a)\%$ **CI** for b_i, the coefficient of x_i in the regression function, is the interval

$$\hat{b}_i \pm t_{a/2,(n-k-1)} s_{\hat{b}_i}, i = 0, 1, \ldots, k, \qquad (2)$$

where $s_{\hat{b}_i}$ is given by (1) , and

$t_{a/2,(n-k-1)}$ is the $\boldsymbol{\alpha/2}$ upper percentile of ***t*-distribution** with $(\boldsymbol{n-k-1})$ df.

Inferences about Parameters

- **CI for mean** $m_{Y|x_1^0,\mathbf{L},x_k^0} = E[Y|x_i = x_i^0, i = \overline{1,k}].$

Let x_i^0 be a specified value of $x_i, i = \overline{1,k}.$

The **point estimator** for mean

$$m_{Y|x_1^0,\mathbf{L},x_k^0} = b_0 + b_1 x_0 + \mathbf{L} + b_k x_k$$

is the statistic

$$\hat{m}_{Y|x_1^0,\mathbf{L},x_k^0} = \hat{b}_0 + \hat{b}_1 x_0 + \mathbf{L} + \hat{b}_k x_k .$$

Inferences about Parameters

For given **α** , a 100(1 – α) % **CI** for $m_{Y|x_1^0,\mathbf{L},x_k^0}$ is

$$\hat{m}_{Y|x_1^0,\mathbf{L},x_k^0} \pm t_{a/2,(n-k-1)} \text{ x (Estimated SD of } m_{Y|x_1^0,\mathbf{L},x_k^0} \text{)}$$

$$= \hat{y} \pm t_{a/2,(n-k-1)} s_{\hat{Y}} \text{ ,}$$

where

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 x_1{}^0 + \mathbf{L} + \hat{b}_k x_k{}^0 ,$$

and $\hat{y}$ is the observed value (estimate) of $\hat{Y}$.

Inferences about Parameters

- **PI for an individual future value of *Y*.**

For given a $(0 \le a \le 1)$, a $100(1-\alpha)\%$ **Prediction Interval (PI)** for an individual future value of Y when the values of the independent variables are $x_1^0, \mathbf{L}, x_k^0$ is the interval

$$\hat{y} \pm t_{a/2,(n-k-1)} \sqrt{s^2 + s^2_{\hat{Y}}}\,.$$

Inferences about Linear Functions

Inferences about Linear Functions of the Model Parameters. In Multiple-Regression Models.

Model Equation:

$$Y = b_0 + b_1 x_1 + \ldots + b_k x_k + e \qquad (1)$$

$$E[Y] = b_0 + b_1 x_1 + \ldots + b_k x_k .$$

Assume we want to make statistic inferences about the **linear function:**

$$L = a_0 b_0 + a_1 b_1 + \ldots + a_k b_k = \sum_{j=0}^{k} a_k b_k = a' b$$

where $a' = [a_0, a_1, \ldots, a_k].$

Inferences about Linear Functions

We have

$$\hat{L} = \sum_{j=0}^{k} a_k \hat{b}_k = a'\hat{b}.$$

$$E(\hat{L}) = \sum_{j=0}^{k} a_k E(\hat{b}_k) = \sum_{j=0}^{k} a_k b_k = a'b.$$

$$Var(\hat{L}) = Var(a'\hat{b}) = [a'(X'X)^{-1}a]s^2.$$

$$Cov(\hat{b}_i, \hat{b}_j) = c_{ij}s^2, \qquad i, j = 0, 1, \dots, n.$$

where

$$C = (X'X)^{-1} = \left\| c_{ij} \right\|_{ij=\overline{0,n}}.$$

Inferences about Linear Functions

1) 100(1 – α)% CI for $L = a'b$ is

$$a'\hat{b} \pm t_{a/2} \cdot s \cdot \sqrt{a'(X'X)^{-1}a},$$

where

$$a' = [a_0, a_1, \ldots, a_k],$$

$$t_{a/2} = t_{a/2,(n-k-1)},$$

$$s^2 = \frac{SSE}{df} = \frac{SSE}{n-k-1}.$$

Inferences about Linear Functions

2) Assume we have model (1) and we want to predict the value Y^* when

$$x_1 = x_1^*, x_2 = x_2^*, \dots, x_k = x_k^*$$

with

$$\hat{Y}^* = \hat{b}_0 + \hat{b}_1 x_1^* + \mathbf{L} + \hat{b}_k x_k^* = a'\hat{b},$$

where

$$a' = [1, x_1^*, \dots, x_k^*].$$

Inferences about Linear Functions

Then **100(1 – α)%** **PI** for ***y*** when

$$x_1 = x_1^*, x_2 = x_2^*, \ldots, x_k = x_k^*$$

is

$$a'\hat{b} \pm t_{a/2} \cdot s \cdot \sqrt{1 + a'(X'X)^{-1}a}$$

where

$$a' = [1, x_1^*, x_2^* \ldots, x_k^*].$$

Remark: A single regression parameter $b_i (i = \overline{0,k})$
Can be considered as linear combination of all $b_i (i = \overline{0,k})$
if we choose $a' = [a_0, a_1, \ldots, a_k]$ with

$$a_j = \begin{cases} 1, & \text{if } j = i \\ 0, & \text{if } j \neq i. \end{cases}$$

Then $b_i = a'\hat{b}, \quad i = \overline{0,k}.$

Inferences about Parameters

- **Testing hypothesis about** b_i.

To test the **significance** of an independent variable x_i, we test the hypotheses:

$$H_0 : b_i = b_{i0} \quad \text{vs.} \quad H_a : b_i \vee b_{i0} \qquad (1)$$

where $\vee = \{\neq, <, >\}$.

The **Test Statistic** is

$$T = \frac{\hat{b}_i - b_{i0}}{s_{\hat{b}_i}} \sim t(n - k - 1)$$

which under H_0 has ***t*** **- distribution** with (***n – k – 1***) ***df***.

Inferences about Parameters

The **observed value** of **TS** is

$$T_0 = T(obs) = \frac{\hat{b}_i - b_{i0}}{s_{\hat{b}_i}}.$$

Then use standard ***t*** -critical value or ***P*** -value methods to test the hypothesis (1).

Examples

- **Example 1.**

Florida morbidity statistics for the decade ending in 1976 show that infectious hepatitis had the **incidence rates** shown in the accompanying table (in cases per 100,000 population).

Codes (x)	x	y
-9	1967	10.5
-7	1968	18.5
-5	1969	22.6
-3	1970	27.2
-1	1971	31.2
1	1972	33.0
3	1973	44.9
5	1974	49.4
7	1975	35.0
9	1976	27.6

Example 1.

a) **Letting** $\boldsymbol{Y}$ denote the **incidence rate,** and
$\boldsymbol{x}$ denote the **coded year**
(-9 for 1967, -7 for 1968, through 9 for 1976),

fit the model $Y = b_0 + b_1 x + e$.

b) For the same data, fit the model $Y = b_0 + b_1 x + b_2 x^2 + e$.

c) Is there evidence of **quadratic effect** in the relationship between $\boldsymbol{Y}$ and $\boldsymbol{x}$?
(Test $H_0 : b_2 = 0$). Use $\boldsymbol{\alpha}$ **= .10.**

d) Find a **90% confidence interval** for b_2.

Example 1.

e) Find a **98% prediction interval** for the **incidence rate** of infectious hepatitis in **1977.** **Use the quadratic model.**

f) For the **quadratic model** carry out an ***F***-test of $H_0 : b_2 = 0$, using $\boldsymbol{\alpha}$ **= .05.**
Compare the results to that of in **Part (c)**.

g) Test $H_0 : b_1 = b_2 = 0$ at the **5% significance level**.

Example 1.

Solution

(a) Using the model $Y = b_0 + b_1 x + e$, calculate

$$X'X = \begin{bmatrix} 10 & 0 \\ 0 & 330 \end{bmatrix}, \quad X'Y = \begin{bmatrix} 299.9 \\ 458.3 \end{bmatrix},$$

and

$$\hat{b} = (X'X)^{-1}X'Y = \begin{bmatrix} 29.99 \\ 1.39 \end{bmatrix} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \end{bmatrix}.$$

Hence the **least squared line** is

$$\hat{y} = 29.99 + 1.39x.$$

Example 1.

(b) Using the model $Y = b_0 + b_1 x + b_2 x^2 + e$, calculate

$$X'X = \begin{bmatrix} 10 & 0 & 330 \\ 0 & 330 & 0 \\ 330 & 0 & 19338 \end{bmatrix}, \quad X'Y = \begin{bmatrix} 299.9 \\ 458.3 \\ 8220.7 \end{bmatrix},$$

and

$$\hat{b} = (X'X)^{-1} X'Y = \begin{bmatrix} 36.54 \\ 1.59 \\ -.20 \end{bmatrix} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix}.$$

Hence the **least squared line** is

$$\hat{y} = 36.54 + 1.35x - .2x^2.$$

Example 1.

(c) From **Part (b)**, $\hat{b}_2 = -.20$, and $c_{22} = .00012$. Then

$$SSE = Y'Y - \hat{b}'X'Y = 245.69.$$

The hypothesis to be tested is

$$H_0 : b_2 = 0 \quad \text{vs.} \quad H_a : b_2 \neq 0$$

The **observed value** of **test statistic** is

$$t_0 = \frac{\hat{b}_2 - 0}{\sqrt{s^2(c_{22})}} = \frac{-.20}{\sqrt{\left(\frac{SSE}{7}\right)c_{22}}} = \frac{-.20}{.0645} = -3.1.$$

Example 1.

The **Rejection Region** with **α = .10** , and **df = 7** is

$$|T| > 1.895.$$

We **reject** H_0 , since $|t_0| = 3.1 > 1.895.$

Thus, there is evidence of a quadratic effect.

(d) The **90% confidence interval** for b_2 is

$$\hat{b}_2 \pm t_{.05}\sqrt{s^2 c_{22}} = -.20 \pm 1.895(.0645) = -.20 \pm .12$$

or [**- .32 , - .08**].

Example 1.

(e) Recall that For given a $(0 \le a \le 1)$, a $100(1 - \alpha)\%$ **Prediction Interval** for an individual future value of $\boldsymbol{Y}$ when the values of the independent variables are $x_1^0, \mathbf{L}, x_k^0$ is the interval

$$a'\hat{b} \pm t_{a/2} \cdot s \cdot \sqrt{a'(X'X)^{-1}a}$$

where $a' = [1, x_1^0, x_2^0 \ldots, x_k^0].$

If the year is 1977, $x^0 = 11.$ Hence $x_1^0 = x^0 = 11,$ $x_2^0 = \left(x^0\right)^2 = 121,$ and $a' = [1, 11, 121].$
Therefore,

$$\hat{y} = a'\hat{b} = 36.54 + 1.39(11) - .20(121) = 27.63.$$

Example 1.

Also, $SSE = Y'Y - \hat{b}'X'Y = 245.69$ and

$$s^2 = \frac{SSE}{n-k-1} = \frac{245.69}{7} = 35.1.$$

Then the **98%** **prediction interval** for an individual future value of ***Y*** when $x^0 = 11$ is

$$a'\hat{b} \pm t_{a/2} \cdot \sqrt{s^2\left(1 + a'(X'X)^{-1}a\right)}$$

$$= 27.63 \pm 2.998\sqrt{35.1(1+1.3833)}$$

$$= 27.63 \pm 27.42 \quad or \quad [.21, 55.05].$$

Example 1.

(f) For the **complete model,**

$$Y = b_0 + b_1 x + b_2 x^2 + e,$$

by Part (c) we have

$SSE_c = 245.69$ with **df = 7**.

For the **reduced model,**

$$Y = b_0 + b_1 x + e,$$

$$SSE_R = Y'Y - b_1 X'Y = 10208.67 - \begin{bmatrix} 29.99 & 1.3879 \end{bmatrix} \begin{bmatrix} 299.9 \\ 458.3 \end{bmatrix} = 578.6.$$

with **df = 8**.

Example 1.

The **test statistic** for testing $H_0 : b_2 = 0$ is

$$F = \frac{\frac{SSE_R - SSE_c}{8-7}}{\frac{SSE_c}{7}} = \frac{332.91}{\frac{245.69}{7}} = 9.49.$$

The **rejection region** with $\boldsymbol{\alpha} = .05$ is $F > F_{1,7} = 5.59,$

and H_0 is **rejected.**

There is evidence that $b_2 \neq 0$.

These results do agree with the results of Part (c).

Example 1.

(g) For the **reduced model**

$$Y = b_0 + e, \quad SSE_R = \sum (y_i - \bar{y})^2 = 1214.669$$

with ***df*** $= \mathbf{9}$.

Then $$F = \frac{\dfrac{SSE_R - SSE_c}{9-7}}{\dfrac{SSE_c}{7}} = \frac{\dfrac{968.98}{2}}{\dfrac{245.69}{7}} = 13.80.$$

The **rejection region** with $\alpha = .05$ is $F > F_{2,7} = 4.74$, and the null hypothesis, $H_0 : b_2 = b_1 = 0$, is **rejected.**

Example 2.

A response $\boldsymbol{Y}$ is a function of **three** independent variables $\boldsymbol{x_1}$, $\boldsymbol{x_2}$ and $\boldsymbol{x_3}$ that are related as follows:

$$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + e.$$

a) **Fit this model** to the $\boldsymbol{n} = 7$ data points shown in the accompanying table

y	x_1	x_2	x_3
1	-3	5	-1
0	-2	0	1
0	-1	-3	1
1	0	-4	0
2	1	-3	-1
3	2	0	-1
3	3	5	1

Example 2

b) **Predict** $\boldsymbol{Y}$ when $x_1 = 1, \quad x_2 = -3, \quad x_3 = -1$.
Compare with the observed response in the original data.
Why are these two not equal?

c) Do this data present sufficient evidence to indicate that $\boldsymbol{x_3}$ contributes information for the prediction of $\boldsymbol{Y}$?
(Test the hypothesis $H_0 : b_3 = 0,$ using $\boldsymbol{\alpha} = .05$.)

d) **Find** a **95%** **confidence interval** for the expected values of $\boldsymbol{Y}$, given $x_1 = 1, \quad x_2 = -3, \quad x_3 = -1$.

e) **Find** a **95%** **prediction interval** for $\boldsymbol{Y}$, given

$x_1 = 1, \quad x_2 = -3, \quad x_3 = -1$.

Example 2

Solution.

a)

$$X = \begin{bmatrix} 1 & -3 & 5 & -1 \\ 1 & -2 & 0 & 1 \\ 1 & -1 & -3 & 1 \\ 1 & 0 & -4 & 0 \\ 1 & 1 & -3 & -1 \\ 1 & 2 & 0 & -1 \\ 1 & 3 & 5 & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 2 \\ 3 \\ 3 \end{bmatrix}, \quad X'Y = \begin{bmatrix} 10 \\ 14 \\ 10 \\ -3 \end{bmatrix}$$

Example 2

$$X'X = \begin{bmatrix} 7 & 0 & 0 & 0 \\ 0 & 28 & 0 & 0 \\ 0 & 0 & 84 & 0 \\ 0 & 0 & 0 & 6 \end{bmatrix}, \quad (X'X)^{-1} = \begin{bmatrix} 1/7 & 0 & 0 & 0 \\ 0 & 1/28 & 0 & 0 \\ 0 & 0 & 1/84 & 0 \\ 0 & 0 & 0 & 1/6 \end{bmatrix},$$

$$\hat{b} = (X'X)^{-1}X'Y = \begin{bmatrix} 1.4285 \\ .5000 \\ .1190 \\ -.5000 \end{bmatrix} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \end{bmatrix},$$

and the **fitted model** is

$$\hat{y} = 1.4825 + .5000x_1 + .1190x_2 - .5000x_3.$$

Example 2

b) When $x_1 = 1, \; x_2 = -3, \; x_3 = -1,$ that is, $a' = [1, 1, -3, -1].$
the **predicted value** of ***y*** is

$$\begin{aligned} \hat{y} = a'\hat{b} &= (1)1.4825 + (1).5000 + (-3).1190 - (-1).5000 \\ &= 1.4825 + .5000 - .3570 + .5000 \\ &= 2.0715. \end{aligned}$$

whereas the **observed response** at this setting was ***y*** = 2.

- The difference appears because the former is predicted value based on a model fit using all of the data whereas latter is an observed response.

Example 2

c) Calculate

$$SSE = Y'Y - \hat{b}'X'Y = 24 - 23.9757 = .0243$$

and $$s^2 = \frac{SSE}{n-4} = \frac{.0243}{3} = .008.$$

In order to test the hypothesis

$$H_0 : b_3 = 0 \ \text{ vs. } \ H_a : b_3 \neq 0$$

we use the test statistic

$$t = \frac{\hat{b}_3 - b_3}{s\sqrt{c_{44}}} = \frac{-.5000}{\sqrt{.008(1/6)}} = \frac{-.5000}{.0365} = -13.7.$$

The **rejection region**, with $\boldsymbol{\alpha} = .05$ and **df = 3** is

$$|\,t\,| > t_{.025,3} = 3.182,$$

and the null hypothesis is **rejected.**

Example 2

d) We have $a' = [1 \quad 1 \quad -3 \quad -1]$ and $\hat{Y} = a'\hat{b} = \hat{b}_0 + \hat{b}_1 - 3\hat{b}_2 - \hat{b}_3$.

Hence a **95%** **confidence interval** for ***E*(*Y*)** is given by

$$\bar{Y} \pm t_{a/2} S\sqrt{a'(X'X)^{-1}a}, \quad \text{where}$$

$$a'(X'X)^{-1}a = [1 \quad 1 \quad -3 \quad -1]\begin{bmatrix} 1/7 & 0 & 0 & 0 \\ 0 & 1/28 & 0 & 0 \\ 0 & 0 & 1/84 & 0 \\ 0 & 0 & 0 & 1/6 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ -3 \\ -1 \end{bmatrix}$$

$$= [1/7 \quad 1/28 \quad -3/84 \quad -1/6]\begin{bmatrix} 1 \\ 1 \\ -3 \\ -1 \end{bmatrix} = .45238.$$

Example 2

Hence the **95%** **confidence interval** is

$$\bar{Y} \pm t_{a/2} S \sqrt{a'(X'X)^{-1} a}$$
$$= 2.0715 \pm 3.182 \sqrt{.008} \sqrt{.45238}$$
$$= 2.07 \pm .19.$$

e) The **95%** **prediction interval** for ***Y*** is

$$\hat{y} \pm t_{a/2} s \sqrt{1 + a'(X'X)^{-1} a}$$
$$= 2.07 \pm 3.182 \sqrt{.008} \sqrt{1.45238}$$
$$= 2.07 \pm .34.$$

Summary: Regression Analysis

Basic Model-Building Concepts

- Models are used to test changes without actually implementing the changes.
- Can be used to predict outputs based on specified inputs
- Consists of 3 components:
  - Model specification
  - Model fitting
  - Model diagnosis.

Summary: Regression Analysis

Model Specification

- Sometimes referred to as model identification.
- Is a process for establishing the framework for the model.
  - Decide what you want to do and select the dependent variable (***y***).
  - Determine the potential independent variables (***x***) for your model.
  - Gather sample data (observations) for all variables.

Summary: Regression Analysis

Model Building

- Process of actually constructing the equation for the data.
- May include some or all of the independent variables (***x***).
- The goal is to explain the variation in the dependent variable (***y***) with the selected independent variables (***x***).

Summary: Regression Analysis

Model Diagnosis

- Analyzing the quality of the model (perform diagnostic checks).
- Assess the extent to which the assumptions appear to be satisfied.
- If unacceptable, begin the model-building process again.
- Should use the simplest model available to meet needs
  - The goal is to help you make better decisions.