Multiple Regression

AU STAT-615

Emil Hvitfeldt

2021-03-017

Inferences about Regression Parameters

The least squares and maximum likelihood estimators in $\mathbf{b}$ are unbiased.

Meaning

$$E\{\mathbf{b}\} = \boldsymbol{\beta}$$

Inferences about Regression Parameters

The variance-covariance matrix is given by

$$V\{\mathbf{b}\}_{p\times p} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

and

$$S^2\{\mathbf{b}\}_{p\times p} = \mathrm{MSE}(\mathbf{X}^T\mathbf{X})^{-1}$$

Inferences about Regression Parameters

Interval estimation of β_k

$$\frac{b_k - \beta_k}{s\{b_k\}} \sim t(n-p)$$

hence the confidence limits for β_k with $1 - \alpha$ confidence are

$$b_k \pm t\left(1 - \frac{\alpha}{2}; n-p\right) \cdot \{b_k\}$$

Inferences about Regression Parameters

Test

$$H_0 : \beta_k = 0 \quad \text{against} \quad H_\alpha : \beta_k \neq 0$$

The test statistic is

$$t^* = \frac{b_k}{s\{b_k\}}$$

if $|t^*| \leq \dfrac{b_k}{s\{b_k\}}$ then we can conclude H_0 else we conclude H_α

Interval Estimation of $E\{Y_h\}$

For given values of $X_1, \ldots, X_{p-1}$ denoted by $X_{h1}, X_{h2}, \ldots, X_{hp-1}$ we denote $E\{Y_h\}$ by

$$E\{Y_h\} = \mathbf{X}_h^T \mathbf{b}$$

where

$$\mathbf{X}_h = \begin{bmatrix} 1 \\ X_{h1} \\ \vdots \\ X_{hp-1} \end{bmatrix}_{p\times 1} \qquad \text{and} \qquad \underset{1\times 1}{\hat{Y}_h} = \underset{1\times p}{\mathbf{X}_h^T} \cdot \underset{p\times 1}{\mathbf{b}}$$

Interval Estimation of $E\{Y_h\}$

The estimator is unbiased, i.e. $E\{\hat{Y}_h\} = E\{Y_h\}$

and the variance can be stated as follows

$$V\{\hat{Y}_h\} = \sigma^2 \mathbf{X}_h^T \cdot \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \cdot \mathbf{X}_h$$

Interval Estimation of $E\{Y_h\}$

But we have that

$$\sigma^2 \cdot \left(\mathbf{X}^T\mathbf{X}\right)^{-1} = V\mathbf{b}$$

so we get

$$V\{\hat{Y}_h\} = \mathbf{X}_h^T \cdot V\{\mathbf{b}\}\mathbf{X}_h$$

and

$$s^2\{\hat{Y}_h\} = \mathbf{X}_h^T \cdot s^2\{\mathbf{b}\} \cdot \mathbf{X}_h$$

Interval Estimation of $E\{Y_h\}$

The $1 - alpha$ confidence limits for $E\{Y_h\}$ are

$$\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2}; n - p\right) \cdot s\{\hat{Y}_h\}$$

Confidence Region for Regression Surface

The $1 - \alpha$ confidence region for entire regression surface is

$$\hat{Y}_h \pm W \cdot s\{\hat{Y}_h\}$$

with

$$W^2 = p \cdot F(1 - \alpha; \quad p; \quad n - p)$$

Prediction of $Y_{h(new)}$

The $1 - alpha$ confidence limits for $Y_{h(new)}$ are

$$\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2}; n - p\right) \cdot s\{\text{pred}\}$$

with

$$\begin{aligned} s^2\{\text{pred}\} &= \text{MSE} + s^2\{\hat{Y}_h\} \\ &= MSE + MSE \cdot \mathbf{X}_h^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}_h \\ &= MSE \left(1 + \mathbf{X}_h^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}_h\right) \end{aligned}$$

Different Decompositions

We have that

$$SSTO = SSR(X_1) + SSE(X_1)$$

if X_1 is the variable X(main variable). Now since

$$SSE(X_1) = SSR(X_2|X_1) + SSE(X_1, X_2)$$

Then we get

$$SSTO = SSR(X_1) + SSR(X_1|X_2) + SSE(X_1, X_2)$$

Different Decompositions

Now since we also have that

$$SSTO = SSR(X_1, X_2) + SSE(X_1, X_2)$$

We can combine it with

$$SSTO = SSR(X_1) + SSR(X_1|X_2) + SSE(X_1, X_2)$$

that we derived earlier

Different Decompositions

Combing the two gives

$$SSR(X_1, X_2) + SSE(X_1, X_2) = SSR(X_1) + SSR(X_1|X_2) + SSE(X_1, X_2)$$
$$SSR(X_1, X_2) = SSR(X_1) + SSR(X_1|X_2)$$

ANOVA table for three predictors

Source of variation	SS	df
Regression	$SSR(X_1, X_2, X_3)$	3 ($p-1$)
X_1	$SSR(X_1)$	1
$X_2 \mid X_1$	$SSR(X_2 \mid X_1)$	1
$X_3 \mid X_1, X_2$	$SSR(X_3 \mid X_1, X_2)$	1
Error	$SSE(X_1, X_2, X_3)$	$n - 4$
Total	$SSTO$	$n - 1$

Uses of Extra Sums of Squares in Tests for regressiom coefficients

When we wish to test whether the term $\beta_k X_k$ can be dropped from a multiple regression model we are interested in

$$H_0 : \beta_k = 0$$
$$H_\alpha : \beta_k \neq 0$$

Example

In the case where we have X_1, X_2, X_3 and we want to test $\beta_3 = 0$ vs $\beta_3 \neq 0$

we can use

$$SSR(X_3|X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3)$$

Example

Hence we get the test statistic

$$\begin{aligned} F^* &= \frac{SSR(X_3|X_1, X_2)}{1} \div \frac{SSE(X_1, X_2, X_3)}{n-4} \\ &= \frac{MSR(X_3|X_1, X_2)}{MSE(X_1, X_2, X_3)} \end{aligned}$$

which we can think of as a marginal test

Example

When we wish to test whether several terms in the regression model can be dropped at the same time, we can construct a test in a similar way

In the case where we wanted to check if we could remove $\beta_2 X_2$ and $\beta_3 X_3$, we have

$$\begin{aligned} H_0 &: \beta_2 = \beta_3 = 0 \\ H_\alpha &: \text{Not both are zero} \end{aligned}$$

Example

Now the test statistic is

$$
\begin{aligned}
F^* &= \frac{SSR(X_2, X_3|X_1)}{2} \div \frac{SSE(X_1, X_2, X_3)}{n-4} \\
&= \frac{MSR(X_2, X_3|X_1)}{MSE(X_1, X_2, X_3)}
\end{aligned}
$$

where

$$
SSR(X_2, X_3|X_1) = SSR(X_2|X_1) + SSR(X_3|X_1, X_2)
$$