Inference

AU STAT-615

Emil Hvitfeldt

2021-02-03

Normal error regression model

For this lecture, we assume that the **normal error regression model** is applicable

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

where

- β_0 and β_1 are parameters
- X_i are known constants
- ε_i are independent $N(0, \sigma^2)$

Inference concerning β_1

Example:

> Study relationship between sales Y and advertising expenditures X

We are generally interested in getting an estimate of β_1

Knowledge of β_1 provides information as to how many additional sales, on average, are generated by an additional amount of advertising expenditure

If any

Tests

Sometimes we set up tests concerning β_1 that we want to answer

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

When $\beta_1 = 0$ then there is no linear association between Y and X.

Sampling distribution of β_1

Before discussing the inference concerning β_1 we need the sampling distribution of b_1

where b_1 is the point estimate of β_1.

Sampling distribution of β_1

The sampling distribution of β_1 refers to the different values of b_1 that would be obtained with repeated sampling.

b_1 is a linear combination of Y_i and some constants

Y_i is normally distributed

This leads to

b_1 being normally distributed

Sampling distribution of β_1

We saw last week (and in 1.10a) that the point estimate of b_1 is:

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

For a normal error regression we get

$$E\{b_1\} = \beta_1 \text{ and } V\{b_1\} = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Normality of b_1

Claim:

> b_1 is a linear combination of Y_i

Thus since Y_i are independently normally distributed and that a linear combination of independent normal random variables are normally distributed, then we have that b_1 is also normally distributed

We now need to show that b_1 is a linear combination of Y_i.

We start with

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

it follows that

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^{n}(X_i - \bar{X})Y_i - \sum_{i=1}^{n}(X_i - \bar{X})\bar{Y} \\ &= \sum_{i=1}^{n}(X_i - \bar{X})Y_i\end{aligned}$$

Normality of b_1

We finally get

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

thus $b_1 = \sum_{i=1}^{n} k_i Y_i$ where $k_i = \dfrac{X_i - \bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$

Mean

We can start with

$$E\{b_1\} = E\left\{\sum_{i=1}^{n} k_i Y_i\right\} = \sum_{i=1}^{n} k_i E\{Y_i\} = \sum_{i=1}^{n} k_i(\beta_0 + \beta_1 X_i)$$

$$= \beta_0 \sum_{i=1}^{n} k_i + \beta_1 \sum_{i=1}^{n} k_i X_i = \beta_1$$

only if $\sum_{i=1}^{n} k_i = 0$ and $\sum_{i=1}^{n} k_i X_i = 1$.

Check if $\sum_{i=1}^{n} k_i = 0$:

$$\begin{aligned}
\sum_{i=1}^{n} k_i &= \sum_{i=1}^{n} \frac{X_i - \bar{X}}{(X_i - \bar{X})^2} \\
&= \sum_{i=1}^{n} \frac{1}{(X_i - \bar{X})^2} \cdot \sum_{i=1}^{n} (X_i - \bar{X}) \\
&= \sum_{i=1}^{n} \frac{1}{(X_i - \bar{X})^2} \cdot 0 = 0
\end{aligned}$$

Check if $\sum_{i=1}^{n} k_i X_i = 1$:

$$\begin{aligned}\sum_{i=1}^{n} k_i X_i &= \sum_{i=1}^{n} \frac{X_i - \bar{X}}{(X_i - \bar{X})^2} X_i \\ &= \sum_{i=1}^{n} \frac{1}{(X_i - \bar{X})^2} \cdot \sum_{i=1}^{n} (X_i - \bar{X}) X_i\end{aligned}$$

Check if $\sum_{i=1}^{n} k_i X_i = 1$:

$$\sum_{i=1}^{n} k_i X_i = \sum_{i=1}^{n} \frac{X_i - \bar{X}}{(X_i - \bar{X})^2} X_i$$
$$= \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \cdot \sum_{i=1}^{n}(X_i - \bar{X})X_i$$

if $\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \bar{X})X_i$ then $\sum_{i=1}^{n} k_i X_i = 1$

check if if $\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \bar{X})X_i$

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\
&= \sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}\bar{X} \\
&= \sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}\frac{\sum X_i}{n} \\
&= \sum_{i=1}^{n} X_i^2 - \bar{X}\sum_{i=1}^{n} X_i \\
&= \sum_{i=1}^{n}(X_i - \bar{X})X_i
\end{aligned}$$

Variance of b_1

$$V\{b_1\} = V\left\{\sum_{i=1}^{n} k_i Y_i\right\} = \sum_{i=1}^{n} k_i^2 V\{Y_i\} = \sum_{i=1}^{n} k_i^2 \cdot \sigma^2$$

$$= \sigma^2 \sum_{i=1}^{n} k_i^2 = \sigma^2 \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Variance of b_1

$$\begin{aligned}
\sum_{i=1}^{n} k_i^2 &= \sum_{i=1}^{n} \left[\frac{X_i - \bar{X}}{(X_i - \bar{X})^2} \right]^2 \\
&= \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\left[(X_i - \bar{X})^2\right]^2} \\
&= \frac{1}{\left[\sum_{i=1}^{n} (X_i - \bar{X})^2\right]^2} \cdot \sum_{i=1}^{n} (X_i - \bar{X})^2 \\
&= \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2}
\end{aligned}$$

Estimated Variance

We can now estimate the variance of the sampling distribution of b_1

$$V\{b_1\} = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

we can replace the parameter σ^2 with MSE which we know is the unbiased estimator of σ^2.

$$s^2\{b_1\} = \frac{MSE}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Review of related distributions

Let Y be a random variable that follows a normal distribution with $E\{Y\} = \mu$ and $V\{Y\} = \sigma^2$

- The standard normal random is $Z = \dfrac{Y - \mu}{\sigma} \rightarrow Z \sim N(0, 1)$

- Let $Y_1, Y_2, \ldots, Y_n$ be independent normal, then we have that $a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$ is normally distributed with mean $\sum a_i E\{Y_i\}$ and variance $\sum a_i^2 V\{Y_i\}$

Review of related distributions

- Let $Z_1, Z_2, \ldots, Z_v$ be independent standard normal. A **chi-square** random variable is defined as

$$\chi^2(v) = Z_1^2 + Z_2^2 + \cdots + Z_v^2$$

where v is called the degrees of freedom (df)

and we have that $E\{\chi^2(v)\} = v$

Review of related distributions

- For Z and $\chi^2(v)$ we can define the t distribution as

$$t(v) = \frac{Z}{\left[\frac{\chi^2(v)}{v}\right]^{1/2}}$$

with mean $E\{t(v)\} = 0$

Interval estimation

for interval estimation, we need the t-distribution

If we let $Y_1, \dots, Y_n$ observations of $Y \sim n(0, 1)$

then we get with

$$\bar{Y} = \frac{\sum X_i}{n} \quad \text{and} \quad s = \left[\frac{\sum (Y_i - \bar{Y})^2}{n-1}\right]^{1/2} \quad \text{and} \quad s\{\bar{Y}\} = \frac{s}{\sqrt{n}}$$

We have that $\frac{\bar{Y} - \mu}{s\{\bar{Y}\}}$ is t-distributed with n-1 degrees of freedom.

Interval estimation

the confidence limits for μ with confidence $1-\alpha$ are

$$\bar{Y} \pm t\left(1-\frac{\alpha}{2}; n-1\right) s\{\bar{Y}\}$$

Confidence interval for β_1

We have to similarly work for the confidence interval for β_1.

We need t find the distribution of $\frac{b_1 - \beta_1}{s\{b_1\}}$

Like previously if Y_i come form the same normal population, then $\frac{\bar{Y} - \mu}{s\{\bar{Y}\}}$ follows a t distribution with $n - 1$ degrees of freedom

The degrees of freedom is $n - 1$ because only one parameter is needed to be estimated

Confidence interval for β_1

for the regression model, we need to estimate two parameters, thus we need $df = n - 2$

In addition b_1 is a linear combination of Y_i therefore $\dfrac{b_1 - \beta_1}{s\{b_1\}}$ is t distributed with $n - 2$ degrees of freedom

Confidence interval for β_1

We note that the confidence interval for $\bar{Y}$ and b_1 are very similar

$$\bar{Y} \pm t\left(1 - \frac{\alpha}{2}; n-1\right) s\{\bar{Y}\}$$

$$b_1 \pm t\left(1 - \frac{\alpha}{2}; n-2\right) s\{b_1\}$$

Tests concerning β_1

Test statistics (TS) for testing means often takes the form

$$TS = \frac{EST - HYP}{SE}$$

- estimate for parameter
- hypothesized value of parameter
- standard error

Tests concerning β_1

So for

$$H_0 : \beta_1 = \beta_{10}$$

$$H_1 : \beta_1 \neq \beta_{10}$$

We use test statistic

$$t = \frac{b_1 - \beta_{10}}{\sqrt{s^2\{b_1\}}} = \frac{b_1 - \beta_{10}}{s\{b_1\}}$$

where t is t-distributed with $n-2$ degrees of freedom and $s^2\{b_1\} = \dfrac{MSE}{\sum(X_i - \bar{X})^2}$

Inference concerning β_0

This is a more limited scope since not all models are in scope when $X = 0$

Recall that $b_0 = \bar{Y} - b_1\bar{X}$ and

$$E\{b_0\} = \beta_0 \quad \text{and} \quad V\{b_0\} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]$$

We can get an estimator of $V\{b_0\}$ by replacing σ^2 with MSE

$$s^2\{b_0\} = MSE \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]$$

Sampling distribution of $(b_0 - \beta_0)/s\{b_0\}$

The sampling distribution of $\frac{(b_0 - \beta_0)}{s\{b_0\}}$ can be be set up in a similar fashion to how the sampling distribution of $\frac{(b_1 - \beta_1)}{s\{b_1\}}$ was set up.

We have that $\frac{(b_0 - \beta_0)}{s\{b_0\}}$ is t-distributed with $n - 2$ degrees of freedom

Confidence interval for β_0

The confidence interval for β_0 is similarly set up in the same way as β_1 and they are

$$b_0 \pm t(1 - \frac{\alpha}{2}; n-2)s\{b_0\}$$

Hypothesis tests

For

$$H_0 : \beta_0 = 0$$

$$H_1 : \beta_0 \neq 0$$

the test statistic is

$$t = \frac{b_0 - \beta_0}{\sqrt{MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]}}$$

Interval estimation of $E\{Y_h\}$

Let X_h denote the level of X for which we wish to estimate the mean response

The point estimator $\hat{Y}_h$ of $E\{Y_h\}$ is given by

$$\hat{Y}_h = b_0 + b_1 X_h$$

Normality

The normality of the sampling distribution of $\hat{Y}_h$ follows directly from the fact that $\hat{Y}_h$ is a linear combination of the observation Y_i.

Mean

We have

$$E\{\hat{Y}_h\} = E\{b_0 + b_1 X_h\} = b_0 + b_1 X_h$$

since $\hat{Y}_h$ is a unbiased estimate of $E\{Y_h\}$

Variance

$$V\{\hat{Y}_h\} = \sigma_2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]$$

Note: The variability of the sampling distribution of $\hat{Y}_h$ is affected by how far X_h is from $\bar{X}$ since we have $(X_h - \bar{X})^2$

Confidence interval

We define

$$\frac{\hat{Y}_h - E\{Y_h\}}{s\{\hat{Y}_h\}}$$

which is t-distributed with $n - 2$ degrees of freedom, and the corresponding confidence interval is

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