Inference

AU STAT-615

Emil Hvitfeldt

2021-02-10

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 $N(0, \sigma^2)$

Prediction Interval for New Observations

Objective:

Prediction of new observation Y corresponding to a given level X of the predictor variable

The new observation on Y to be predicted is viewed as the result of a new trial independent of the trials on which the regression analysis is based.

Notation

Let X_h be the level of X for new and new observation on Y as $Y_{h(new)}$

Goal:

Predict an individual outcome drawn from the distribution of Y

Prediction Interval for New Observations

In the previous case, we were estimating $E\{Y_h\}$ by $\hat{Y}_h$

Our best guess for a new observation is still $\hat{Y}_h$. The estimated mean is still the best prediction we can make

The difference is in the amount of variability

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

Amount of variability

This means that we have

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

Which we can estimate with $\sigma^2 = MSE$

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

Mean response at X_h

We want to estimate $E\{Y_h\}$

The point estimate is $\hat{Y}_h$ and the variance is

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

The $1 - alpha$ percent confidence interval will be

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

New observation at X_h

We want to predict $Yh(new)$ drawn from Y

$$V\{Y_{h(new)} - \hat{Y}_h\} = V\{Y_{h(new)}\} + V\{\hat{Y}_h\}$$

The $1 - alpha$ percent confidence interval will be

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

Note: This will be wider than the mean response CI

IT accounts for both the uncertainty in knowing the value of the population mean + data scattering

Confidence Band for Regression Line

Goal:

Obtain a confidence band for the entire regression line $E\{Y\} = \beta_0 + \beta_1 X$

Why?

It helps us to determine the appropriation of a fitted regression function

Confidence Band for Regression Line

We get

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

where $W^2 = 2F(1-\alpha; 2, n-2)$

Confidence Band for Regression Line

In the case of a simple linear regression, it is equivalent to another test, the F test for the significance of the regression

This equivalence is true only for simple linear regression

Confidence Band for Regression Line

Let's start with $Y_i - \bar{Y}$ which measures the deviation of an observation from the sample mean

We have that

$$Y_i - \bar{Y} = Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}$$

- Total deviation
- Deviation around fitted regression line
- Deviation of fitted regression value around mean

Confidence Band for Regression Line

It can be shown that the sums of these squared deviations have the same relationship

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2$$

ANOVA Table

We have that $p = 2$

Source of variation	Sum of Squared	Degrees of Freedom	Mean Square	F
Regression	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	$p - 1 \to 1$	$\frac{SSR}{1} = MSR$	$\frac{MSR}{MSE}$
Error	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$n - p = 0 \to n - 2$	$MSE = \frac{SSE}{n-2}$	
Total	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$n - p + p - 1 = n - 1$		

ANOVA Table

Why $p - 1$?

This corresponds to the fact that we specify a line by two points

We will prove this in multiple regression

Significance of the regression test

Idea:

> Comparison of MSR and MSE is useful for testing whether or not $\beta_1 = 0$. If MSE and MSE are of the same order of magnitude, this would suggest that $\beta_1 = 0$

If MSR is substantially greater than MSE, this would suggest that $\beta_1 \neq 0$

Significance of the regression test

We note that

$$E\{MSE\} = \sigma^2$$

and

$$E\{MSR\} = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$$

if $\beta_1 = 0$ then we get that $E\{MSE\} = E\{MSR\}$

So it makes sense to compare them by

$$F^* = \frac{MSR}{MSE}$$

Significance of the regression test

We can set up the hypotheses

$$H_0 : \beta_1 = 0 \qquad H_\alpha : \beta_1 \neq 0$$

If $F^* \leq F(1-\alpha; 1, n-2)$ when we conclude H_0

If $F^* > F(1-\alpha; 1, n-2)$ when we conclude H_α

Note: F^* is always positive

Example

The time it takes to transmit a file always depends on the file size. Suppose you transmitted 30 files with an average size of 126 Kbytes and a standard deviation of 35 Kbytes.

The average transmittance time was 0.04 seconds with a standard deviation of 0.01 seconds. The correlation coefficient between the time and size was 0.86.

In other words, we are given that $n = 30$, $s\{X\} = 35$, $s\{Y\} = 0.01$, and $r = 0.86$.

Example

We are given that $n = 30$, $s\{X\} = 35$, $s\{Y\} = 0.01$, and $r = 0.86$

a) Compute the total, regression, and error sum of squares.

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = (n-1)\cdot V\{Y\} = 29 \cdot 0.01^2 = 0.0029$$

Example

We also have that

$$r^2 = \frac{SSR}{SST}$$

thus

$$SSR = r^2 \cdot SST = 0.86^2 \cdot 0.0029 = 0.00214$$

and

$$SSE = SST - SSR = 0.00076$$

Example

b) Compute the ANOVA Table

Sum of squared	DF	mean sq	F
SSR	1	0.00214	$\frac{MSR}{MSE} = 79.3$
SSE	$n - 2 = 28$	0.000027	
SST	$n - 1 = 29$		

Example

c) Use the F-statistic to test the significance of our regression model that related transmission time to the size of the file. State H_0 and H_1 and draw conclusion for $1 - \alpha = 0.95$.

$$H_0 : \beta_1 = 0 \qquad H_\alpha : \beta_1 \neq 0$$

We have $F^* = 79.3$, and $F(1-\alpha; 1, 28) = 4.17$

we have that $F^* > F(1-\alpha; 1, 28)$

so we reject H_0. The slope is significant. There is evidence of a linear relation between X and Y

Example

d) Coefficient of determination

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

The coefficient of determination is interpreted as the proportion of observed variation in Y that can be explained by the simple linear regression model

Here $R^2 = \dfrac{0.00214}{0.0029} = 0.738$

It means that 73.8% of the total variation of transmission times is explained solely by the file size