12/13 Regression

2011/2011

1 D. The point estimate for the average value of delay in innovation for company of size 60m is given by
$y' = b_0 + b_1(60)$

A 95% CI here has the usual interpretation

If we were to get the delay in innovation time for several companies of size 60m, then we are 95% confident that the true mean/average of delay would lie within the interval

A point estimate for the delay time for a "single" company of size 60m is also $y' = b_0 + b_1(60)$

But a 95% PI refers to the interval ~~within~~ within which the true delay time of that single company of size = 60m will lie. with (95% confidence).

They are different since the variability of delay time for a single value is much greater than the variability of the average

Single observation = mean + chance variation

This is why the PI are wider than 95% CI.

A 60m euro company took: (single instance)
1. 20 weeks PI = (4.95, 24.02)
If we were to test for $H_0: y' = 20$ vs $H_1: y' \neq 20$ then from PI we see that the null hypothesis will not be rejected

20 can be accepted as a good point estimate for the prediction

ii) 35 means
- If we were to test $H_0: Y' = 35$ vs $H_1: Y' \neq 35$ then from PI we can reject H_0
- 35 cannot be accepted as a good estimate for the true prediction for delay time

E. $Y_i \Rightarrow$ observation i
$\hat{Y}_i \Rightarrow$ i^{th} fitted value
$\bar{Y} \Rightarrow$ mean of Y_i for all Y

$SSTO = SSR + SSE$

$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$R\text{-}Sq = \frac{SSR}{SSTO}$

It is the proportion of total variability about the mean $\bar{y}$, as explained by the regression line

2011 - 2012

Q3 a) $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$ $\quad Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

For β_1 $p = 0.018$ which is less than 0.05, reject H_0.
$\Rightarrow$ There is a relationship. Sweetness index and peltin are related as per the data.

c) For pecan content = 250, Sweetness Index = $6.25 - 0.00231(250)$

For reliability we look at $r\text{-}Sq^2$ value = 22.9%. R^2 is really small
$\Rightarrow$ ~~correlation~~ Proportion of total variability about the mean as explained by the regression line = 0.23 implying a bad fit
Unreliable prediction

1 2012 Q3 Rea

Q3 a prove $\sum(y_i-\bar{y})^2 = \sum(\hat{y}_i-\bar{y}+y_i-\hat{y}_i)^2$

$= \sum[(\hat{y}_i-\bar{y})^2 + (y_i-\hat{y}_i)^2 + 2(\hat{y}_i-\bar{y})(y_i-\hat{y}_i)]$

$= \sum(\hat{y}_i-\bar{y})^2 + \sum(y_i-\hat{y}_i)^2$

$+2\sum(\hat{y}_i-\bar{y})\sum(y_i-\hat{y}_i)$

$+2\sum(\hat{y}_i(y_i-\hat{y}_i)-\bar{y}(y_i-\hat{y}_i))$

$+2\sum\hat{y}_i e_i - 2\bar{y}\sum e$

$=0$

$\sum(\hat{y}_i-\bar{y})^2 + \sum(y_i-\hat{y}_i)^2$

$= SSE + SSR$

b Height $=x$ Weight $=y$ $\sum x_i y_i = 43956$

$\sum x_i = 765$ $\sum y_i = 285$ $SSE = 3641$

$\bar{x} = 153$ $\bar{y} = 57$ $n=5$

$\sum x_i^2 = 117403$ $\sum y_i^2 = 16425$

$b_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} \quad \frac{\sum(x_iy_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y})}{\sum(x_i^2+\bar{x}^2-2x_i\bar{x})}$

$\frac{\sum x_iy_i - \bar{y}\sum x_i - \bar{x}\sum y_i + n\bar{x}\bar{y}}{\sum x_i^2 + n\sum x_i\sum x_i - 2\sum x_i\sum x_i}$

$\frac{\sum x_iy_i - \frac{\sum y_i\sum x_i}{n} - \frac{\sum x_i\sum y_i}{n} + \frac{n\sum x_i\sum y_i}{n^2}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$

$\frac{\sum x_iy_i - \frac{\sum x_i\sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$

Sub in $= \frac{43956 - \frac{765(285)}{5}}{117403 - \frac{(765)^2}{5}} \quad \frac{351}{858} = \frac{9}{22} = 0.4090$

- b_1 is an estimate of β_1
- It is the slope of the regression line
- For each unit increase in height, there is a $\frac{9}{22}$ increase in weight.

Intercept = b_0

Calculated by $\bar{y} - b_1 \bar{x}$

$57 - \frac{9}{22}(153) = -\frac{123}{22} = -5.5909$

- This is the value of weight when height = 0
- However we cannot have a negative weight so we disregard this val

ii. $E[y] = -\frac{123}{22} + \frac{9}{22} x_i$

$x_i = 150 \Rightarrow = \frac{-123}{22} + \frac{9}{22}(150)$

$= \frac{1227}{22} = \approx 55.77$ (kg)

Prediction interval = $y' \pm t_{critical}\,(s.e(x'))$

$= \hat{y} \pm t_{0.975,3} \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-2}} \sqrt{1 + \frac{1}{n} + \frac{(x'-\bar{x})^2}{\Sigma x_i^2 - n(\bar{x})^2}}$

$= \frac{1227}{22} \pm 2.353 \sqrt{\frac{(13.641)^2}{3}} \sqrt{1 + \frac{1}{5} \frac{(150-153)^2}{117935 - 5(153)^2}}$

$\frac{9}{858}$

$\frac{1227}{22} \pm (2.353)(3.4837714) \sqrt{1 + \frac{1}{5} + \frac{9}{858}} \sqrt{(1.100222)}$

$\frac{1227}{22} \pm 4.9572097$

(55.749, 105.345)

$\frac{1227}{22} \pm 8.215513326$

(47.56, 63.99) = prediction interval for height = 150 and weight = $\frac{1227}{22}$

3

1 2012 Reg Q3

Prediction interval tell us where we can expect to see the next data point sampled. Assume that the data ready are randomly selected from Gaussian Normal dist.

Collect a sample of data and calculate a prediction model

Then sample one more value from the population

If you do this many times, we expect that the next value will lie within that prediction interval in 95% of the samples

Key point is that the prediction interval tells you about the distribution of values, not the uncertainty in determining the population mean

Prediction interval must account for both the uncertainty in knowing the value of the population mean, plus data scatter.

So a prediction interval is always wider than a confidence interval

- A prediction interval defines a range of values within which a response is likely to fall given a specified value of a predictor

- The uncertainty represented by a prediction interval includes not only the uncertainty (variance) associated with the population mean and the new observation, but the uncertainty associated with the regression parameters

- Because of the uncertainty associated with the population mean and the new observation are independent

of the observed used to fit the model the uncertainty estimate must be combined using root-sum-of squares to yield unary SE

iii. Confidence interval tell us how well we have estimated the mean. Assume data is randomly sampled from normal dist. If we do this many times, and calculate a confidence interval for the mean from each sample, we expect 95% of those intervals to include the true value of population mean. CI tells you about the likely location of the true population parameter

---

A CI is an interval associated with a parameter and is a frequentist concept. The parameter is assumed to be not random but unknown, and the CI interval is computed from the data because data is random, the interval is random. A 95% CI will contain the true parameter with probability 0.95. With a large number of repeated samples, 95% of the intervals would contain the true value

iv. As explored in above answers, the CI will be a much narrower interval

Reasons for in formula.

For CI SE part = $\sqrt{MSE\left[\frac{1}{n}\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]}$

For PI SE part $\sqrt{MSE\left[1+\frac{1}{n}\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]}$

For PI formula the extra one makes $\sqrt{}$ bigger and thus a larger interval as other figures remain the same

2012 PAPER 1 Q3 Regression DAVID WEITBRECHT

Q3 A. Show that SST = SSE + SSR

$$\Sigma (y_i - \bar{y})^2 = \Sigma (\hat{y}_i - \bar{y} + y_i - \hat{y}_i)^2$$

$$= \Sigma [\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i)^2 + 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

$$\Sigma(\hat{y}_i - \bar{y})^2 + \Sigma(y_i - \hat{y}_i)^2 + 2\Sigma(\hat{y}_i - \bar{y}) \cdot \Sigma(y_i - \hat{y}_i)$$

$$\Sigma(\hat{y}_i - \bar{y})^2 + \Sigma(y_i - \hat{y}_i)^2 + 2\Sigma(\hat{y}_i(y_i - \hat{y}_i) - \bar{y}(y_i - \hat{y}_i))$$

$$+ 2\Sigma \hat{y}_i e_i - 2\bar{y}\Sigma e_i$$

$$\neq 0$$

$$\Rightarrow (\hat{y}_i - \bar{y})^2 + \Sigma(y_i - \hat{y}_i)^2 = SST$$

SSE + SSR

bi. Calculate and interpret slope and intercept.

Height = x | Weight = y | $\Sigma x_i y_i = 43956$

$\Sigma x_i = 765$ | $\Sigma y_i = 285$ | SSE = 36.41

$\bar{x} = 153$ | $\bar{y} = 57$ | $n = 5$

$\Sigma x^2 = 117903$ | $\Sigma y_i^2 = 16453$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{\Sigma(x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y})}{\Sigma(x_i^2 + \bar{x}^2 - 2x_i\bar{x})}$$

$$= \frac{\Sigma x y_i - \bar{y}\Sigma x_i - \bar{x}\Sigma y_i + n\bar{x}\bar{y}}{\Sigma x_i^2 + n\frac{\Sigma x_i \Sigma x_i}{n^2} - 2\frac{\Sigma x_i^2}{n}}$$

$$\Sigma x_i y_i - \frac{\Sigma y_i \Sigma x}{n} - \frac{\Sigma x_i \Sigma y}{n} + \frac{n\Sigma x_i \Sigma y_i}{n^2}$$

$$= \frac{\Sigma x y_i - \frac{\Sigma x_i \Sigma y_i}{n}}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}$$

Sub in values $\Rightarrow \frac{43956 - \frac{765(285)}{5}}{117903 - \frac{(765)^2}{5}} = \frac{351}{858} = \frac{9}{22} = 0.4090$

DW

2

- b_1 is an estimate of population parameter β_1
- It is the slope of the estimated regression line
- For each unit increase in height, there is a $\frac{9}{22}$ increase in weight.

Y intercept $= b_0$.
Calculated by $\bar{y} - b_1 \bar{x}$

$$57 - \frac{9}{22}(153) = -\frac{123}{22} = -5.5909$$

- This is the mean weight when height $= 0$.
- This is theoretical as we cannot have a negative weight.

ii. Prediction Interval for height $= 150$cm 95% confidence

Prediction Interval $= \hat{y} \pm t_{critical}\, s.e(x')$

$$= \hat{y} \pm t_{0.975,3} \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}} \sqrt{1 + \frac{1}{n} + \frac{(x'-\bar{x})^2}{\sum x^2 - n(\bar{x})^2}}$$

$$E[y] = -\frac{123}{22} + \frac{9}{22} x_1$$

$$x_1 = 150 \Rightarrow = -\frac{123}{22} + \frac{9}{22}(150) = \frac{1227}{22} \approx 55.77 \text{ kg}$$

$$\Rightarrow \frac{1227}{22} \pm 2.353 \sqrt{\frac{36.41}{5}} \sqrt{1 + \frac{1}{5} + \frac{(150-153)^2}{11790 - 5(153)^2}}$$

$$\frac{1227}{22} \pm 2.353 (3.483771)(1.100221)$$

$$\frac{1227}{22} \pm 8.215513326$$

95 PI $= (47.56, 63.99)$ height $= 150$cm and weight $\frac{1227}{22}$ kg

3 MANG SCI PAPER 1 Q3 2012 Div

Prediction Interval

Predictional interval tells us where we can expect to see the next dot point sampled Tells us about distribution of values, not the uncertainty in determing the populaon mean

Predicting a particular y for a given x. Predicting outcome of a single experiment given X-value

iii. Confidence Interval

Estimate the mean value of y for a given x.

Confidence Interval CI tells us how well we have determined the mean we expect 95% of time that those intervals will include the true value of the populaon value

iv. Difference

It will always be lower, in the formula, PI has an extra 1 under the square root with all the figure) the same meaning it will always be wider

The error in estimating the mean value of y $E[y]$, for a given X say x_p, is the distance between the least squared line and the true line of means $E(y) = \beta_0 + \beta_1 x$

-In contrast, the error $(y_p - \hat{y})$ in predicting some future value of y is the sum of two errors - the error in estimating the mean of y $E(y)$ plus the random error that is a component of the value of y to be predicted

DW

Consequently the error of predicting a particular value of y will always be larger than the error of estimating the mean value of Y for a particular value of X

The further x (x_i) lies from $\bar{X}$, the larger the error of estimation and prediction will be