Prediction Intervals[1]

STA 302 Fall 2020

[1] See last slide for copyright information.

Prediction

- You have a data set, and you fit a regression model.
- Now you obtain a *new* observation, independently sampled from the same population.
- You have $(1, x_{n+1,1}, \ldots, x_{n+1,k})'$. Or $(x_{n+1,1}, \ldots, x_{n+1,k})'$.
- Want to predict y_{n+1}.
- For example, based on the `cars` data, you want to predict litres per kilometer for a Japanese car 4.52 metres long, weighing 1,295 kilograms.
- I wish we could write $(1, x_{n+1,1}, \ldots, x_{n+1,k})' = \mathbf{x}_{n+1}$.
- But we will follow the book's notation and call it $\mathbf{x}_0$.

New observation: $y_0 = \mathbf{x}_0'\boldsymbol{\beta} + \epsilon_0$

- $E(y_0) = \mathbf{x}_0'\boldsymbol{\beta}$.
- Estimate $E(y_0)$ with $\mathbf{x}_0'\widehat{\boldsymbol{\beta}}$.
- That's a reasonable *prediction* of y_0, too.
- But the intervals are different.
- Prediction intervals are not the same as confidence intervals.

Prediction intervals versus confidence intervals

Based on $y_0 = \mathbf{x}_0'\boldsymbol{\beta} + \epsilon_0$

- A confidence interval tries to trap the unknown constant $\mathbf{x}_0'\boldsymbol{\beta}$ with high probability, say $1 - \alpha = 0.95$
- Have $\mathbf{a}'\widehat{\boldsymbol{\beta}} \pm t_{\alpha/2}\sqrt{MSE\,\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}$. Let $\mathbf{a} = \mathbf{x}_0$.
- A prediction interval seeks to trap y_0, a random variable.
- It makes sense that the prediction interval should be wider.
- We will have $\mathbf{x}_0'\widehat{\boldsymbol{\beta}} \pm t_{\alpha/2}\sqrt{MSE\,(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}$.

Theorem

- Assume the usual linear regression model with normal error terms.
- Let $y_0 = \mathbf{x}_0'\boldsymbol{\beta} + \epsilon_0$, where $\epsilon_0 \sim N(0, \sigma^2)$, independently of $\epsilon_1, \ldots, \epsilon_n$.

A $(1-\alpha)100\%$ prediction interval for y_0 is given by

$$\mathbf{x}_0'\widehat{\boldsymbol{\beta}} \pm t_{\alpha/2} \sqrt{MSE\,(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}$$

Proof will use $t = \frac{z}{\sqrt{w/\nu}}$

- Predict y_0 with $\mathbf{x}_0'\widehat{\boldsymbol{\beta}} \sim N(\mathbf{x}_0'\boldsymbol{\beta}, \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)$.
- $y_0 \sim N(\mathbf{x}_0'\boldsymbol{\beta}, \sigma^2)$.
- And y_0 is independent of $\mathbf{x}_0'\widehat{\boldsymbol{\beta}}$. Why?
- Error of prediction: $y_0 - \mathbf{x}_0'\widehat{\boldsymbol{\beta}} \sim N(0, \sigma^2 + \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)$.
- Standardize this to get the z in the numerator of t.

$$z = \frac{y_0 - \mathbf{x}_0'\widehat{\boldsymbol{\beta}}}{\sqrt{\sigma^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}} \sim N(0, 1)$$

- And $w = \frac{SSE}{\sigma^2}$ (what else?)
- z and w are independent. Why?

$t = \frac{z}{\sqrt{w/(n-k-1)}} \sim t(n-k-1)$

With $z = \frac{y_0 - \mathbf{x}_0'\widehat{\boldsymbol{\beta}}}{\sqrt{\sigma^2(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}}$ and $w = \frac{SSE}{\sigma^2}$

$$
\begin{aligned}
t &= \frac{z}{\sqrt{w/\nu}} \\
&= \frac{y_0 - \mathbf{x}_0'\widehat{\boldsymbol{\beta}}}{\sqrt{\sigma^2(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}} \bigg/ \sqrt{\frac{SSE}{\sigma^2}/(n-k-1)} \\
&= \frac{y_0 - \mathbf{x}_0'\widehat{\boldsymbol{\beta}}}{\sqrt{MSE(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}} \sim t(n-k-1)
\end{aligned}
$$

Deriving the prediction interval

$$\begin{aligned} 1-\alpha &= P\{-t_{\alpha/2} < t < t_{\alpha/2}\} \\ &= P\left\{-t_{\alpha/2} < \frac{y_0 - \mathbf{x}_0'\widehat{\boldsymbol{\beta}}}{\sqrt{MSE\,(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}} < t_{\alpha/2}\right\} \\ &\vdots \\ &= P\left\{\mathbf{x}_0'\widehat{\boldsymbol{\beta}} - t_{\alpha/2}\sqrt{MSE\,(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} < y_0 \right. \\ &\qquad \left. < \mathbf{x}_0'\widehat{\boldsymbol{\beta}} + t_{\alpha/2}\sqrt{MSE\,(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}\right\} \end{aligned}$$

Or, $\mathbf{x}_0'\widehat{\boldsymbol{\beta}} \pm t_{\alpha/2}\sqrt{MSE\,(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}$. ■

Copyright Information