

PREDICTION INTERVALS

What if we want to estimate $Y|x$ not just $E(Y|x)$?

The only estimator available is

$$\hat{y} = \hat{\eta}_0 + \hat{\eta}_1 x \text{ --}$$

the same estimator we used for $E(Y|x)$, calling it $\hat{E}(Y|x)$.

Intuitively, we should not be able to estimate Y as closely as $E(Y|x)$: Estimating $E(Y|x)$ involves *sampling error* only, but estimating $Y|x$ must take into account the *natural variability of the distribution $Y|x$* as well as sampling error.

[Try drawing a picture to illustrate this.]

The increased variability in estimating Y as compared to estimating $E(Y|x)$ requires us to use a different standard error.

To help avoid confusion, estimating Y is called *prediction*, the estimate is sometimes called y_{pred} rather than \hat{y} (so $y_{\text{pred}} = \hat{\eta}_0 + \hat{\eta}_1 x$), and the associated error is called *prediction error*:

Prediction error: For a *new* observation y chosen from $Y|x$ independently of y_1, \dots, y_n , we define

$$\text{Prediction error} = y - \hat{E}(Y|x)$$

- Draw a picture
- Compare and contrast with the error $e|x$ and the residuals \hat{e}_i
- Prediction error is a random variable -- its value depends on the choice of y_1, \dots, y_n , and y

For fixed x ,

$$E(\text{prediction error}) = E(Y|x - \hat{E}(Y|x)) = \underline{\hspace{10cm}}$$

Also,

$$\begin{aligned} \text{Var}(\text{prediction error}) &= \text{Var}(Y|x - \hat{E}(Y|x) | x_1, \dots, x_n) \\ &= \text{Var}(Y|x, x_1, \dots, x_n) + \text{Var}(\hat{E}(Y|x) | x_1, \dots, x_n) \quad (\text{Why?}) \\ &= \text{Var}(Y|x) + \text{Var}(\hat{E}(Y|x) | x_1, \dots, x_n) \\ &= \sigma^2 + \text{Var}(\hat{E}(Y|x)) \quad \text{for short} \end{aligned}$$

$$= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{SXX} \right)$$

$$\sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SXX} \right)$$

Define: $se(y_{pred|x}) = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SXX}}$

$$= \sqrt{\hat{\sigma}^2 + Var(\hat{E}(Y|x))}$$

Sampling Distribution of Prediction Error:

- $\hat{E}(Y|x)$ is a linear combination of the y_i 's, $\Rightarrow y|x - \hat{E}(Y|x)$ is a linear combination of y and the y_i 's .
- This plus independence and normality assumptions on $y|x$ and the y_i 's $\Rightarrow y|x - \hat{E}(Y|x)$ is normally distributed.
- It can be shown that this implies that

$$\frac{Y|x - \hat{E}(Y|x)}{se(y_{pred|x})} \sim t(n-2).$$

Thus we can use this statistic to calculate a *prediction interval* (or "confidence interval for prediction") for y .

Recall: A 90% *confidence interval* for the conditional mean $E(Y|x)$ is an interval produced by a process which, for 90% of all independent random samples y_1, \dots, y_n taken from $Y|x_1, \dots, Y|x_n$, respectively, yields an interval containing the parameter $E(Y|x)$ (assuming all model assumptions fit).

Compare and contrast: A 90% *prediction interval* (or "confidence interval for prediction") is an interval produced by a process which, for 90% of all independent random samples y_1, \dots, y_n, y taken from $Y|x_1, \dots, Y|x_n, Y|x$, respectively, yields an interval containing the new sampled value y (assuming all the model assumptions fit).

Thus the prediction interval is *not* a confidence interval in the usual sense -- since it is used to estimate a value of a random variable rather than a parameter.