PREDICTION INTERVALS

We want to estimate Y|x not just E(Y|x)

The only estimator available for Y|x:

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

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

Estimating E(Y|x) involved *sampling error*. Estimating Y|x involves the *natural variability in the random variable Y|x* as well as sampling error -- so we will need a different (larger) standard error.

Terminology:

- Estimating Y is called *prediction*
- The estimate is sometimes called y_{pred} rather than $\hat{y}$ (so $y_{pred} = \hat{\eta}_0 + \hat{\eta}_1 x$)
- The associated error is called *prediction error*:

Prediction error: For a *new* (or additional) observation y chosen from Y|x <u>independently</u> of $y_1, \ldots, y_n$, we define

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

- It's a random variable -- its value depends on the choice of $y_1, \ldots, y_n$, and y.
- Picture:
- Compare and contrast with error e|x and the residuals $\hat{e}_i$

For fixed x (still assuming fixed $x_1, \ldots, x_n$),

E(prediction error) = E(Y|x - $\hat{E}(Y|x)$)

=

Also (for fixed x),

Var(prediction error)

= Var($Y|x - \hat{E}(Y|x)| x_1, \dots, x_n$)

= Var($Y|x, x_1, \dots, x_n$) + Var($\hat{E}(Y|x)|x, x_1, \dots, x_n$))

(Why?)

= Var($Y|x$) + Var($\hat{E}(Y|x)| x_1, \dots, x_n$))

= σ^2 + Var($\hat{E}(Y|x)$) for short

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

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

Define:

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

$$= \sqrt{\hat{\sigma}^2 + V\hat{a}r(\hat{E}(Y\,|\,x))}$$

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

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

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 $y_1, \ldots, y_n$ taken from $Y|x_1, \ldots, Y|x_n$, respectively, yields an interval containing the parameter E(Y|x) (assuming all the 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 $y_1, \ldots, y_n$, y taken from $Y|x_1, \ldots, Y|x_n$, 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.