Last time we learned that the prediction we get from regression, i.e. \( \hat{y}(x) = \hat{\alpha} + \hat{\beta} x + \varepsilon \) (no \( \varepsilon \)!) has 2 interpretations:

1) The (conditional) mean of all \( y \), given \( x \).
2) The prediction of a single \( y \), given \( x \).

An estimate of the former requires a C.I. of the true (conditional) mean of \( y \), given \( x \).

The uncertainty in the latter is conveyed through a P.I. for the specific \( y \) value being predicted.

\[
\text{C.I.:} \quad \hat{y}(x) \pm t^* s_e \sqrt{\frac{1}{n} + \frac{(x-x)^2}{S_{xx}}} = \hat{y} \pm s^*_y
\]

2) Prediction Interval (P.I.) for a single \( y \).

Suppose \( y^* \) is Joe’s \( y \) value corresponding to his \( x \)-value, \( x^* \).

Theorem states that \( \hat{y}(x) - y^* \) has a normal distr. with params \( \mu = 0 \), \( \sigma^2 = \sigma^2 \) prediction error.

\[
\text{Prediction error} = \hat{y}(x) - y^* \leq \sigma^*_y
\]
\[ \sigma_{\text{pred. err}}^2 = \text{Var}[\hat{y}(x)] + \text{Var}[\epsilon] \]

Estimate/approximate with \( \hat{\sigma}_\gamma^2 \) (given above)

\[ \hat{\sigma}_\gamma^2 = \frac{\sigma_e^2}{\epsilon} \]

Think! The variance of all \( y \) values at a given \( x \) is \( \sigma_e^2 \).

Estimate/approximate with \( S_e^2 \).

\[ S_{\text{pred. err}}^2 = \hat{\sigma}_\gamma^2 + \frac{S_e^2}{\epsilon} \]

above \( \text{SSE/(n-2)} \)

\[ \text{prediction error} \]

\[ z \rightarrow t \quad t = \frac{\hat{y}(x) - \gamma^*}{S_{\text{pred. err}}} \]

\( \sim t - \text{dist, df=}\ n-2 \)

P.I. for a single \( y \): \( \hat{y} \pm t^* S_{\text{pred. err}} = \hat{y} \pm t^* \sqrt{\hat{\sigma}_\gamma^2 + S_e^2} \)

Compare with C.I. for \( \gamma \) (the conditional mean): \( \hat{y} \pm t^* S_{\gamma} \)

Which one is bigger? P.I. makes sense?
Summary picture:

C.I. ← est. error ← $\hat{y} + s_y^* (x^*)$

P.I. ← pred. error ← $\hat{y} + s_y (x^*)$

Don't forget what these intervals mean:

2 interpretations for C.I.:

1) We are 95% confident that the true (conditional) mean of $y$, given $x$, is in a computed C.I.

2) About 95% of random C.I.s will cover the true conditional mean of $y$, given $x$.

For P.I.: only 1 interpretation (for now)

1) About 95% of these P.I.s will cover an individual's $y$, given his/her $x$.

1') After we are more comfortable with these interpretations we will allow ourselves to also say things like "plausible values of observed $y$, given $x$, are in the observed P.I."

(See example, below)
In summary:

\[ \hat{\text{est. error}} = \hat{y} - y(x) \]

\[ \sigma_{\text{est. error}}^2 = \sigma_y^2 + \sigma_{y(x)}^2 \]

Recall that \( \sigma_y^2 \) means the variance of \( y \) under resampling. \( y(x) \) is the fit to the \( y \) for a given \( x \), and so, its variance under resampling is just \( \sigma_y^2 \).

\[ \sigma_{\text{pred. error}}^2 = \sigma_y^2 + \sigma_e^2 \]

\[ S_{\text{est. error}}^2 = S_y^2 \]

\[ S_{\text{pred. error}}^2 = S_y^2 + S_e^2 \]

C.I.: \( \hat{y} \pm t^* S_e \sqrt{\frac{1 + (x - x)^2}{n - S_{xx}^{-1}}} \)

P.I.: \( \hat{y} \pm t^* \sqrt{\frac{S_y^2 + S_e^2}{n}} \)

\[ S_y^2 \xrightarrow{\text{n} \to \infty} 0 \] \( \xrightarrow{n \to \infty} \) But P.I. does not!

One more comparison: How do C.I. & P.I. vary as \( n \to \infty \)?
In the example from prev. lecture, we found the C.I.:

\[
\hat{y} \pm t^{*} \frac{\hat{\sigma}_{e}}{\sqrt{n}} = \hat{y} \pm t^{*} \frac{S_{e}}{\sqrt{n}} \sqrt{\frac{1}{n} + \frac{(x-x)^{2}}{S_{xx}}}
\]

\[
= 3.445 \pm 2.365 \left(0.0644\right) \sqrt{\frac{1}{9} + \frac{(1.5 - 12.6)^{2}}{0.6}}
\]
\[
\text{df} = 9 - 2 = 7
\]
\[
0.02302 \approx S_{\text{est. err.}} = S_{\hat{\sigma}_{e}}
\]

\[
\hat{y} \pm t^{*} S_{\hat{\sigma}_{e}} = 3.445 \pm 0.0544
\]

\[
(3.39, 3.50)
\]

\[
\frac{\hat{y} \pm t^{*} S_{\hat{\sigma}_{e}}}{\text{for a single case}}
\]

\[
\text{predict oxygen diffusivity when temperature is 1.5 K} \text{ in a way that conveys info about reliability & precision.}
\]

This is asking for a prediction interval:

\[
\hat{y} \pm t^{*} \sqrt{S_{\hat{\sigma}_{e}}^{2} + S_{\hat{\sigma}_{e}}^{2}}
\]

\[
= 3.445 \pm 2.365 \sqrt{(0.02302)^{2} + (0.0644)^{2}}
\]

\[
= 3.445 \pm 0.1617 = (3.28, 3.61)
\]

1) 95% of such PI's will cover single observations of \( y \) at x = 1.5

2) At 95% prediction level, plausible values for a single observation on \( y \), at \( x = 1.5 \), are between 3.28 and 3.61.
We have been talking about inference on $\beta$ (and $\alpha$), the conditional mean of $y$, given $x$, and a single $y_i$ at $x$.

What about multiple regression?

In going from $y = \alpha + \beta x$ \hline (1+1 params) to $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ \hline \hline ($k+1$ params) things generalize in a straightforward way.

Basically df changes: $n-2 \Rightarrow n-(k+1)$ \hline $\Rightarrow \# of \beta$'s.

E.g. $\sigma^2$ is estimated with $\frac{SSE}{[n-(k+1)]} = s_e^2$.

The df for $t$, and p-value, changes too.

But don’t forget that the issues of collinearity and interaction come back again.

But the presence of multiple $\beta$'s presents one more test we can do:

$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$ \hline \hline $\leftrightarrow$ Test of “model utility”. \hline \hline makes sense! In $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ if all $\beta_i = 0$, then none of the predictors are good. \hline \hline i.e. bad model!