INFERENCE ON MEAN RESPONSE LEVELS AND PREDICTION
NORMAL SIMPLE LINEAR REGRESSION MODEL:

\[ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \]

\( x_i, i = 1...n \) fixed (or condition on)
\( \varepsilon_i, i = 1...n \) random errors s.t.
\[ \varepsilon_i \sim N(0, \sigma^2), \forall i \]

independent

Under this scenario, we consider inference for:

\[ E(y \text{ at } x) = \text{mean response} \]
\[ y_{(\text{new})} \text{ at } x = \text{response prediction} \]

Starting point, and pivot for our intervals: the fitted value at \( x \):

\[ \hat{y}(x) = b_0 + b_1 x \]
The fitted values is itself a linear combination of the $y$'s, and with the usual reasoning we obtain:

$$
\hat{y}(x) \sim N \left( E(y \text{ at } x) = \beta_0 + \beta_1 x, \quad \sigma^2 \right)
$$

Unbiased estimate of the expected value of $y$ at $x$.

Variance depends inversely on sample size and on spread of the $x$'s, and directly on the distance between the $x$ under consideration and the center of the $x$ data.
\[ \text{var}(\hat{y}(x)) = s^2 \left[ \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right] \]

and
\[ \frac{\hat{y}(x) - (\beta_0 + \beta_1 x)}{se(\hat{y}(x))} \sim t_{n-2} \]

... basis for confidence interval

1-\(\alpha\) level Confidence Interval for the mean response at \(x\):

\[ \hat{y}(x) \pm t_{n-2} (1 - \alpha / 2)se(\hat{y}(x)) \]
Predicting $y_{\text{new}}$ at $x$: two sources of uncertainty

$$
\hat{y}(x) \sim N(\beta_0 + \beta_1 x, \sigma^2)
$$

... variation in the possible location of the distribution of $y$ at $x$ (estimating $E(y \text{ at } x)$)

$$
y(x) \sim N(\beta_0 + \beta_1 x, \sigma^2)
$$

... variation within the distribution of $y$ at $x$ (adding $\varepsilon \sim N(0, \sigma^2)$)
\[
\text{"error" in } y_{(new)} \text{ independent from those in the data } y\text{'s}
\]

\[
\text{var}(y_{(new)}(x)) = \text{var}(\hat{y}(x)) + \text{var}(\varepsilon) = \sigma^2 \left[ \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right] + \sigma^2
\]

\[
y_{(new)}(x) \sim N \left( \beta_0 + \beta_1 x, \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right] \right)
\]
\[ \text{var}(y_{(\text{new})}(x)) = s^2 \left[ 1 + \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2 \right] \]

\[ s\{\text{pred}\} = s \left[ 1 + \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2 \right]^{1/2} \]

and \[ \frac{y_{\text{new}}(x) - (\beta_0 + \beta_1 x)}{s\{\text{pred}\}} \sim t_{n-2} \]

prediction error (larger than se of the fitted value for \(E(y \text{ at } x)\))

... basis for prediction interval

1-\(\alpha\) level Prediction Interval for \(y_{(\text{new})}\) at \(x\):

\[ \hat{y}(x) \pm t_{n-2} \left( 1 - \frac{\alpha}{2} \right) s\{\text{pred}\} \]
Since the prediction error is larger than the standard error at all $x$’s, the PI is larger than the CI for any $x$ and chosen level $1-\alpha$.

Since both prediction error and standard error increase with the square distance between $x$ and $x$-bar, both PI and CI become broader for any chosen level $1-\alpha$.

The fitted value is the pivot for both PI and CI.
1-\(\alpha\) level Working-Hotelling Confidence band for the entire regression line:

\[
\hat{y}(x) \pm t_{n-2} \left(1 - \frac{\alpha}{2}\right) se(\hat{y}(x))
\]

replace the multiplier

\[
\hat{y}(x) \pm w \, se(\hat{y}(x))
\]

\[
w^2 = 2F_{2,n-2}(1-\alpha)
\]

the band is broader at every \(x\) level, because it takes into account the fact that the inference encompasses the entire line.
Regression Analysis: Second versus First

The regression equation is
Second = 22.5 + 0.755 First

$$\text{Predictor} \quad \text{Coef} \quad \text{SE Coef} \quad T \quad P$$

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>22.47</td>
<td>10.22</td>
<td>2.20</td>
<td>0.036</td>
</tr>
<tr>
<td>First</td>
<td>0.7546</td>
<td>0.1417</td>
<td>5.32</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 11.5131   R-Sq = 49.4%   R-Sq(adj) = 47.7%

Predicted Values for New Observations

New

<table>
<thead>
<tr>
<th>Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95% CI</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75.29</td>
<td>2.07</td>
<td>(71.05, 79.52)</td>
<td>(51.36, 99.21)</td>
</tr>
</tbody>
</table>

Values of Predictors for New Observations

New

<table>
<thead>
<tr>
<th>Obs</th>
<th>First</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>70.0</td>
</tr>
</tbody>
</table>
Regression Analysis: %WhoGrad versus CSAT
The regression equation is
%WhoGrad = 2.4 + 0.0700 CSAT
49 cases used, 7 cases contain missing values
Predictor Coef SE Coef T P
Constant 2.39 10.62 0.22 0.823
CSAT 0.07005 0.01032 6.79 0.000
S = 11.9236 R-Sq = 49.5% R-Sq(adj) = 48.4%

Predicted Values for New Observations
New
Obs  Fit  SE Fit  95% CI  95% PI
 1  65.43  2.08  (61.25, 69.61)  (41.08, 89.78)

Values of Predictors for New Observations
New
Obs  CSAT
 1  900