Confidence Intervals for Predictions

(consider the simple linear regression model)

\[ y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, 2, \ldots, N. \]

1. Suppose we are interested in predicting the expected value of \( y \) given \( X = X_0 \), that is,

\[ E(y | X = X_0) = \beta_0 + \beta_1 X_0 = \mu_{y|x_0}. \]

The unbiased predictor of \( E(y | X = X_0) \) is

\[ \hat{E}(y | X = X_0) = \hat{\beta}_0 + \hat{\beta}_1 X_0. \]

But what is the variance of this prediction?

\[
\text{Var}[\hat{E}(y | X = X_0)] = \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 X_0)
\]

\[
= \text{Var}(\hat{\beta}_0) + X_0^2 \text{Var}(\hat{\beta}_1) \\
+ 2X_0 \text{Cov}(\hat{\beta}_0, \hat{\beta}_1)
\]
\[
\sigma^2 \frac{\sum_{i=1}^{N} x_i^2}{N \sum_{i=1}^{N} (x_i - \bar{x})^2} + x_0^2 \frac{\sigma^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} - 2x_0 \bar{x} \frac{\sigma^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}
\]

\[
= \sigma^2 \left( \frac{1}{N} \sum_{i=1}^{N} x_i^2 + x_0^2 - 2x_0 \bar{x} \right)
\]

\[
= \sigma^2 \left[ \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 + (x_0 - \bar{x})^2 \right]
\]

\[
= \sigma^2 \left( \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \right)
\]

We might abbreviate our notation somewhat by denoting

\[
\sigma^2 (\tilde{y}, x_0) = \sigma^2 \left( \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \right).
\]

That is,
That is the variance of \( \hat{\beta}_0 + \hat{\beta}_1 x_0 \) in predicting \( E(y | x = x_0) \) is denoted \( \sigma^2 \).

As we can see, the accuracy in predicting \( E(y | x = x_0) \) increases as \( N \) increases but decreases as the distance between \( x_0 \) and \( \bar{x} \) increases. (It goes without saying, the larger the error variance, \( \sigma^2 \), the less precise the prediction.)

The standard error of the prediction of \( E(y | x = x_0) \) is given by

\[
se(E(y | x_0)) = \hat{\sigma} \left( \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right)^{1/2}
\]

where \( \hat{\sigma} = \sqrt{\frac{SSE}{N-k-1}} \).
Now if we assume normality of the errors \( \varepsilon \) in the regression model, we can invent the below statistic to get a \((1-\alpha)\)% confidence interval for \( E(Y \mid X = X_0) = \mu_y \mid x_0 \). Note that the random variable

\[
t_v = \frac{\hat{\beta}_0 + \hat{\beta}_1 X_0 - E(Y \mid X = X_0)}{se(\hat{\mu}_y \mid x_0)}
\]

has a \( t \)-distribution with \( v = N-k-1 \) degrees of freedom. Thus,

\[
Pr(-t_{v, \alpha/2} < \frac{\hat{\beta}_0 + \hat{\beta}_1 X_0 - E(Y \mid X = X_0)}{se(\hat{\mu}_y \mid x_0)} < t_{v, \alpha/2}) = 1 - \alpha
\]

Inverting the above expression gives us the \((1-\alpha)\)% confidence interval for \( E(Y \mid X = X_0) \), namely

\[
Pr(\hat{\beta}_0 + \hat{\beta}_1 X_0 - t_{v, \alpha/2} \cdot se(\hat{\mu}_y \mid x_0) < E(Y \mid X = X_0) < \hat{\beta}_0 + \hat{\beta}_1 X_0 + t_{v, \alpha/2} \cdot se(\hat{\mu}_y \mid x_0)) = 1 - \alpha.
\]
Obviously, given the variance-covariance matrix of the OLS estimators \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), we can get the standard error of the predicted \( \hat{\beta}_0 + \hat{\beta}_1 x_0 \) in predicting \( E(y | x = x_0) = \beta_0 + \beta_1 x_0 \) as

\[
se(\hat{y}_i | x_0) = \left[ \text{Var}(\hat{\beta}_0) + x_0^2 \text{Var}(\hat{\beta}_1) + 2x_0 \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \right]^{1/2}.
\]

But is there a way of tricking the computer into getting \( se(\hat{y}_i | x_0) \) directly? The answer is yes! Let

\[
\Theta_0 = \beta_0 + \beta_1 x_0.
\]

Then rewrite the original single linear regression model as

\[
y = \beta_0 + \beta_1 x - \Theta_0 + \Theta_0 + u
\]

\[
= \Theta_0 + \beta_1 (x - x_0) + u \quad (1).
\]
Estimating equation (1) by OLS we have

\[ \hat{y} = \hat{\theta}_0 + \hat{\beta}_1 (X - X_0) \quad (2) \]

Then in (2) \( \hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0 = E(Y | X = X_0) \), as we desire and \( \text{se}(\hat{\theta}_0) = \text{se}(E(Y | X = X_0)) \) that we need for forming the \( (1 - \alpha) \) confidence interval for the prediction of \( E(Y | X = X_0) \). The confidence intervals looks like

\[ \hat{\theta}_0 \pm t_{\alpha/2, \text{se}(E(Y | X = X_0))} \cdot \text{se}(E(Y | X = X_0)) \]
2. Consider the problem of predicting an individual observation, say $y_0$, when $X = X_0$.

The prediction error is then

$$\hat{\epsilon}_0 = (y_0 - \hat{\beta}_1 X_0)$$

$$= (\hat{\beta}_0 + \hat{\beta}_1 X_0 + u_0 - \hat{\beta}_1 X_0)$$

$$= (\hat{\beta}_0 + \hat{\beta}_1 X_0 - \hat{\beta}_1 X_0 + u_0)$$

$$= u_0 + u_0.$$

But $\text{Var}(\hat{\epsilon}_0) = \text{Var}(u_0) + \text{Var}(u_0) + 2\text{Cov}(u_0, u_0)$

$$= \text{Var}(u_0) + \text{Var}(u_0)$$

Since $\text{Cov}(u_0, u_0) = 0$, this follows since $\hat{\beta}_0$ and $\hat{\beta}_1$ are independent of $u_0$. Now $\text{Var}(u_0) = \text{Var}(E(y \mid X = X_0))$ and we have

$$\text{Var}(\hat{\epsilon}_0) = \text{Var}(\hat{\epsilon}_0) = \text{Var}(E(y \mid X = X_0)) + \sigma^2$$

$$= \sigma^2 \left[ 1 + \frac{1}{N} + \frac{(X_0 - \bar{X})^2}{\sigma^2 (x_i - \bar{X})^2} \right]$$

Notice the extra term!
The corresponding standard error in predicting $y_o$ is then
\[ \sqrt{\text{Var}(\hat{y}_o)} = \sqrt{\text{Var}(E(y|X=X_0)) + \hat{\sigma}^2} = \text{so}(\hat{y}_o) \]
we can get \[ \text{Var}(E(y|X=X_0)) = [\text{so}(\hat{\theta}_0)]^2 \]
as previously derived from (2) on page 6.
Also \[ \hat{\sigma}^2 = \frac{\text{SSR}}{(n-k-1)} \] or (standard error of regression) \[ \hat{y}_o \] It follows that the $(1-\alpha)\%$ confidence for predicting $y_o$ when $X=X_0$ is

\[ \Pr(\hat{y}_o - t_{n-k-1,\frac{\alpha}{2}} \cdot \text{so}(\hat{y}_o) < y_o < \hat{y}_0 + t_{n-k-1,\frac{\alpha}{2}} \cdot \text{so}(\hat{y}_0)) = 1-\alpha. \]

Notice that the $(1-\alpha)\%$ confidence intervals for predicting $y_o$ are wider than the corresponding confidence intervals for predicting $E(y|X=X_0)$.
This is represented in the following graph.
solid lines = \((1 - \alpha)\) confidence intervals for 
\(E(Y | X = x_0)\)

dashed lines = \((1 - \alpha)\) confidence intervals for \(\hat{y}_0\)

The bottom line is that it is more difficult to predict \(y_0\) than \(E(Y | X = x_0)\).
3. Suppose that we do not know $X_0$ but instead have an estimate of $X_0$, say $\hat{X}_0$. Let

$$\hat{X}_0 = X_0 + \varepsilon$$

where $\varepsilon$ is a zero mean error and independent of the regression error $\epsilon$.

The first point to note is that

$$\tilde{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 \hat{X}_0$$

is unbiased forecast of $y_0$. This follows because the forecast error $\tilde{e}_0 = \tilde{y}_0 - y_0$ has a zero expectation:

$$E(\tilde{y}_0 - y_0) = E(\hat{\beta}_0 + \hat{\beta}_1 \hat{X}_0 - \beta_0 - \beta_1 X_0 - y_0)$$

$$= E(\hat{\beta}_0 - \beta_0) + E(\hat{\beta}_1 (\hat{X}_0 - X_0)) - E(y_0)$$

$$= 0 + E(\hat{\beta}_1 \varepsilon) - 0$$

$$= E(\hat{\beta}_1)E(\varepsilon) = \hat{\beta}_1 \cdot 0 = 0.$$ 

Here we have used the fact that $\hat{\beta}_1$ and $\varepsilon$ are independent random variables.
Deriving the variance of the forecast error $\tilde{\epsilon}_0$ is more complex. Note that

\[
E(\tilde{\epsilon}_0^2) = E((\hat{y}_0 - y_0)^2) = E((\hat{\beta}_0 + \hat{\beta}_1 \hat{x}_0 - \hat{\beta}_0 - \beta_0 - \beta_1 x_0)^2
\]
\[
= E[(\hat{\beta}_0 - \beta_0)^2 + (\hat{\beta}_1 \hat{x}_0 - \beta_1 x_0)^2 + \epsilon_0^2]
\]
\[
= E[(\hat{\beta}_0 - \beta_0)^2] + E[(\hat{\beta}_1 \hat{x}_0 - \beta_1 x_0)^2] + E(\epsilon_0^2)
\]
\[
+ 2E[(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 \hat{x}_0 - \beta_1 x_0)]
\]

(3)

First let us focus on the second term of (3). 

\[
E(\hat{\beta}_1 \hat{x}_0 - \beta_1 x_0)^2 = E[\hat{\beta}_1 (\hat{x}_0 - x_0) + x_0 (\hat{\beta}_1 - \beta_1)]^2
\]

(since $\hat{\beta}_1 \hat{x}_0 - \beta_1 x_0 = \hat{\beta}_1 (\hat{x}_0 - x_0) + x_0 (\hat{\beta}_1 - \beta_1)$)

\[
= E[\hat{\beta}_1^2 (\hat{x}_0 - x_0)^2] + x_0^2 E(\hat{\beta}_1 - \beta_1)^2
\]
\[
+ 2E[\hat{\beta}_1 (\hat{x}_0 - x_0) x_0 (\hat{\beta}_1 - \beta_1)]
\]
\[
= \left[\beta_0^2 + \text{Var}(\hat{\beta}_1)\right] \epsilon_0^2 + x_0^2 \text{Var}(\hat{\beta}_1)
\]

where \(E[\hat{\beta}_1 (\hat{x}_0 - x_0) x_0 (\hat{\beta}_1 - \beta_1)] = E[\hat{\beta}_1 x_0 (\hat{\beta}_1 - \beta_1) \epsilon_0]
\]
\[
= E[\hat{\beta}_1 x_0 (\hat{\beta}_1 - \beta_1)] \cdot E(\epsilon_0) = 0
\]
\[ E \left[ \beta_1^2 (X_0 - X) \right] = E (\beta_1^2 \varepsilon^2) = E (\beta_1^2) E (\varepsilon^2) \]
\[ = (\text{Var}(\beta_1) + \beta_1^2) \sigma_\varepsilon^2 \]
(since \( \text{Var}(\beta_1) = E (\beta_1^2) - (E (\beta_1))^2 = E (\beta_1^2) - \beta_1^2 \))

Now let us turn to the last term of (3)

\[ E (\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1 X_0) \]
\[ = E \left[ (\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1 X_0) + X_0 (\hat{\beta}_1 - \beta_1) \right] \]
\[ = E (\hat{\beta}_0 - \beta_0) \hat{\beta}_1 + X_0 E (\hat{\beta}_0 - \beta_0) X (\hat{\beta}_1 - \beta_1) \]
\[ = E (\hat{\beta}_0 - \beta_0) \hat{\beta}_1 + X_0 \text{ Cov}(\hat{\beta}_0, \hat{\beta}_1) \]
\[ = X_0 \text{ Cov}(\beta, \hat{\beta}_1) \]

Therefore

\[ \sigma_0^2 = \frac{1}{E (\varepsilon^2)} \]
\[ = \text{Var}(\hat{\beta}_0) + [\beta_1^2 + \text{Var}(\hat{\beta}_1)] \sigma_\varepsilon^2 + X_0^2 \text{Var}(\hat{\beta}_1) \]
\[ + \sigma_\varepsilon^2 + 2X_0 \text{ Cov}(\hat{\beta}_0, \hat{\beta}_1) \]

In terms of our least-squares estimators, the formula of the variance of the error in predicting \( y \) by means of \( \hat{\beta}_0 + \hat{\beta}_1 X_0 \) is
\[ \sigma_{g_0}^2 = \sigma^2 \left[ 1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2 + \sigma_x^2}{\sum (x_0 - \bar{x})^2} \right] + \beta_1^2 \sigma_x^2 \]

This last formula makes it obvious that when \( x_0 \) itself is uncertain, the variance of the forecast error is increased. Therefore, the \((1 - \alpha)\) confidence intervals associated with the predicted value \( \tilde{g}_0 \) will be larger than similar confidence intervals when \( x_0 \) is known. Unfortunately, it is quite difficult to describe the confidence intervals for \( g_0 \) using \( \tilde{g}_0 \) because \( \tilde{g}_0 \) is not normally distributed, since it \((\tilde{g}_0)\) involves the sum of products of normally distributed variables that is not normally distributed. Therefore, we cannot invent a t-random variable to get a prediction confidence interval \( g_0 \) as we previously have done for the previous cases.

1. Calculate the \((1-\alpha)\)% confidence intervals associated with the forecasts that would be obtained were we to select \(x_0\) to be two standard deviations (\(2\delta\)) higher than \(\hat{x}_0\), namely \(\hat{x}_* = \hat{x}_0 + 2\delta\), and two standard deviations lower than \(\hat{x}_0\), namely \(\hat{x}** = \hat{x}_0 - 2\delta\), i.e. confidence intervals associated with \(\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 (x_*)\) and \(\hat{y}** = \hat{\beta}_0 + \hat{\beta}_1 (x**)\).

2. The final prediction interval is taken to be the union of the two confidence intervals, i.e. it contains all the values of \(\hat{y}\) common to both confidence intervals.
A graphical representation of this procedure is presented below:

\[
\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.
\]

First prediction interval:

\[
\hat{x}_0 - 2\hat{\sigma}_e \leq \hat{x}_0 \leq \hat{x}_0 + 2\hat{\sigma}_e
\]

Approximate prediction interval:

\[
\hat{x}_0 - 2\hat{\sigma}_e \leq \hat{x}_0 \leq \hat{x}_0 + 2\hat{\sigma}_e
\]

Second prediction interval:

\[
\hat{x}_0 - 2\hat{\sigma}_e \leq \hat{x}_0 \leq \hat{x}_0 + 2\hat{\sigma}_e
\]
see examples 6.1 and 6.6A for examples of constructing prediction intervals for \( \hat{Y}(X = x_0) \) and \( y_0 \) in a multiple regression problem. Also see the SAS program gpa2.sas that is posted on the course website.

**Residual Analysis**

See pp. 206-207 for a discussion of how residuals from estimated regression models can be used to identify under-valued assets (see the house price example) or (law schools that provide the highest value added.

**Predicting \( y \) when \( \log(y) \) is the Dependent Variable**

Consider the model

\[
\log(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon.
\]

We can predict \( \log(y) \) by

\[
\hat{\log(y)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k.
\]
To get a predicted value of $y$ we can use the formula:

$$
\hat{y} = \exp{\left(\hat{\alpha}/2\right)} \exp{\left(\log(y)\right)}
$$

where

$$
\hat{\alpha} = \sqrt{\frac{SSR}{N-k-1}}.
$$

For an example of how this formula is utilized see the SAS program \texttt{ees12.sas} that is on the website for the course.

Choosing between two regression models, one with $y$ as the dependent variable and the other having the dependent variable $\log(y)$.

Consider the two regression models:

$$
\begin{align*}
 y &= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \quad (4) \\
 \log(y) &= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \quad (5)
\end{align*}
$$

First calculate the $R^2$ of model (4) by computing

$$
R_{y^2}^2 = 1 - \frac{SSR}{SSR} = \frac{SSE}{SSR}.
$$
Second, apply OLS to equation (5) and get the fitted values

\[ \hat{\log(y)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k. \]

Then convert these \( \log(y) \) fitted values to fitted values of \( y \) by using the formula

\[ \tilde{y} = \exp(\hat{\sigma}/2) \exp(\hat{\log(y)}). \]

Finally, calculate the \( R^2 \) of the log model as

\[
R^2_{\text{log}(y)} = \left[ \text{corr}(y, \tilde{y}) \right]^2
= \left[ \frac{\sum_i (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum_i (y_i - \bar{y})^2 \sum_i (\tilde{y}_i - \bar{\tilde{y}})^2}} \right]^2
= \frac{\sum_i (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})^2}{\sum_i (y_i - \bar{y})^2 \sum_i (\tilde{y}_i - \bar{\tilde{y}})^2}.
\]

Then choose the log model if \( R^2_{\text{log}(y)} > R^2_y \), otherwise, choose the model with \( y \) as the dependent variable if \( R^2_{\text{log}(y)} < R^2_y \).
Alternatively, $R^2_{\text{log(y)}}$ could be computed as

$$R^2_{\text{log(y)}} = 1 - \frac{SSR_{\text{log(y)}}}{SS_T}$$

where $SSR_{\text{log(y)}} = \sum_{i=1}^{N} (y_i - \bar{y}_i)^2$ and $SS_T = \sum_{i=1}^{N} (y_i - \bar{y})^2$.

Actually, in this exercise the regressors in Eqs. (4) and (5) need not be exactly the same variables but you do have to have the same number of variables per equation to make the comparison. See the SAS program `ceo2.sas` for a comparison of the regression models of examples (6.7) and (6.8) on pp. 209-210 in the Wooldridge textbook.