3.08 Simple Regression: CI and PI for predicted values

In this video you'll learn how to interpret a **confidence interval** used to determine a range of plausible values for a **predicted population mean**. You'll also learn how to interpret a **prediction interval** used to determine plausible values for a **predicted response** for an *individual*. We'll see - roughly - how these intervals are constructed and how they differ from each other.

Take the example where we predicted popularity of cat videos - measured as number of video views - using the cat's age as the predictor.
Suppose we want to know what the range of plausible predicted scores is when a cat is one year old. The width of this range will tell us something about the accuracy of our prediction.

To determine this range we need to calculate a confidence interval, just like we did with the regression coefficient. However, calculating such an interval for a predicted score is a bit more complicated.
With predicted scores we have to be explicit about *what* we're predicting exactly. Are we talking about the predicted mean score for the *group of all* one-year-old cats in the population, or are we talking about the predicted score for an *individual* one-year-old cat?

Why is there a difference? Well estimating a mean will be more precise; a group mean is more stable than individual response scores. If we draw a sample repeatedly and calculate the mean, extreme scores will generally be offset by less extreme scores, stabilizing the mean. If we draw just one case repeatedly we'll see more variability than in the mean.
This is why we have two separate formulas for calculating an interval for a predicted *mean* population score and a predicted *individual* score.

Let's see how these intervals are constructed to better understand the difference between them. Please note that we'll look at an *approximation* of the formulas because the calculation is kind of complicated. Normally you would use software to calculate the intervals.

First we'll look at the interval for predictions of the **mean response** in the population for a particular predictor value. We refer to this as the **confidence interval** for a predicted population mean.

The *approximation* of the formula is the predicted value obtained from our sample - for a one-year-old cat - plus and minus the margin of

error: $CI_{\mu_y}: \hat{y} \pm 2 \cdot \frac{s_{res}}{\sqrt{n}}$. As we've seen earlier, the margin of error is the t-value associated with the selected confidence level and degrees of freedom, times the standard error.
We'll consider only a ninety-five percent ***confidence interval*** and approximate t with the value two.
We'll approximate the standard error with the residual standard deviation, divided by the square root of n: $se \approx \frac{s_{res}}{\sqrt{n}}$. If the sample size is large then the margin of error will be relatively small, resulting in a narrow interval.
If you don't recognize the residual standard deviation: it's the square root of the residual sum of squares divided by n minus two: $\sqrt{\frac{SS_{res}}{(n-2)}}$. It represents the variability around the regression line.

Suppose that in our example the predicted value for one-year-old cats is 41.95, the residual standard deviation is 3.42, and the sample size n is 5. Then the margin of error is 2 times 3.42 divided by the square root of 5, equals 3.06. So the **confidence interval** around the predicted value of 41.95 for one-year-old cats ranges from 38.89 to 45.01.

The second interval, used for predictions of individual response values is referred to as the **prediction interval** for a predicted response. Suppose we wanted to know what the range of plausible predicted popularity scores is for an *individual* one-year-old cat.

Again, the *approximation* of the formula is the predicted value obtained from our sample for a one-year-old cat, plus and minus the margin of error:
$PI_{y_i}: \hat{y}_i \pm 2 \cdot s_{res}$.
Notice I've added sub-i to the predicted value to indicate we are talking about the predicted response for an individual case here.

The predicted response value for a one-year-old cat is the same whether you predict a mean score or an individual score. You can see that the only real difference between the two formulas lies in the standard error, which reflects the precision of our estimation of the predicted score.

For the ***prediction interval*** the standard error is approximately equal to the standard deviation of the residuals: $se \approx s_{res}$. This means that the **prediction interval** will be wider, or less precise than the **confidence interval**, since we're no longer dividing by the square root of n.
In our example the margin of error for the prediction interval is 2 times 3.42. So the **prediction interval** around the predicted value of 41.95 for an individual one-year-old cat ranges from 35.11 to 48.79.

As you can see, the prediction interval is wider - less precise - since individual scores will show larger variability than means of scores.

Remember that in order for any conclusions or interpretations based on these intervals to be valid, all the assumptions of linear regression must hold. If the predictor and response variable aren't linearly related, or if the residuals aren't homoscedastic then the intervals might underestimate or overestimate the precision of the prediction.