Inference for Regression Equations

In a beginning course in statistics, most often, the computational formulas for inference in regression settings are simply given to the students. Some attempt is made to illustrate why the components of the formula make sense, but the derivations are "beyond the scope of the course".

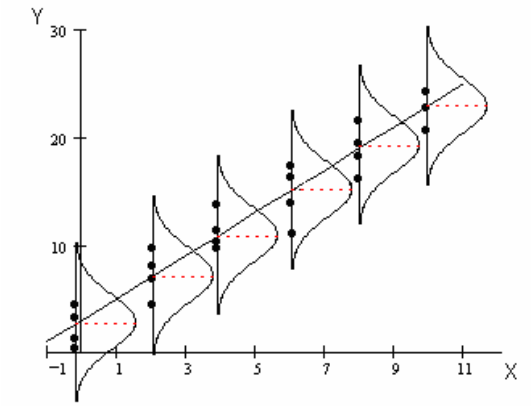
For advanced students, however, these formulas become applications of the expected value theorems studied earlier in the year. To derive the regression inference equations, students must remember that $Var(kX) = k^2 Var(X)$, and, when X and Y are independent, $Var(X+Y) = Var(X) + Var(Y)$ and $Var(XY) = Var(X)Var(Y)$. Finally, $Var(\bar{X}_n) = \frac{Var(X)}{\sqrt{n}}$.

In addition, the Modeling Assumptions for Regression are:

1. There is a normally distributed subpopulation of responses for each value of the explanatory variable. These subpopulations all have a common variance. So, $y \mid x \sim N(\mu_{y|x}, \sigma_e)$.

2. The means of the subpopulations fall on a straight-line function of the explanatory variable. This means that $\mu_{y|x} = \alpha + \beta x$ and that $\hat{y} = a + bx$ estimates the mean response for a given value of the explanatory variable.

Another way to describe this is to say that $Y = \alpha + \beta X + \varepsilon$ with $\varepsilon \sim N(0, \sigma_e)$.



Graphical Representation of regression assumptions

3. The selection of an observation from any of the subpopulations is independent of the selection of any other observation. The values of the explanatory variable are assumed to be fixed. This fixed (and known) value for the independent variable is essential for developing the formulae.

The key to understanding the various standard errors for regression is to realize that the variation of interest comes from the distribution of *y* around $\mu_{y|x}$. This is $\varepsilon \sim N(0, \sigma_e)$.

From our initial work on regression, we saw that $\hat{y} = a + bx$ and $\hat{y} = \bar{y} + b(x - \bar{x})$.

Now, if we let $X_i = x_i - \bar{x}$ and $Y_i = y_i - \bar{y}$, then $b = \frac{\sum_i X_i Y_i}{\sum_i X_i^2}$. All of the regression equations originate with this computational formula for *b*.

To see that this is true, consider $\hat{y} = \overline{y} + b(x - \overline{x})$. In this form, we have a one variable problem. Since we know all the individual values of *x* and *y*, and, consequently, the means $\overline{x}$ and $\overline{y}$, we can use first semester calculus to solve for *b*. Define $S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \left(\overline{y} + b(x_i - \overline{x})\right)\right)^2$. Now, let $X_i = x_i - \overline{x}$ and $Y_i = y_i - \overline{y}$, so $S = \sum_{i=1}^{n} (Y_i - bX_i)^2$. Find the value of *b* that minimizes *S*.

$\frac{dS}{db} = 2\sum_{i=1}^{n} (Y_i - bX_i)(-X_i)$. If $\frac{dS}{db} = 0$, then $\sum_{i=1}^{n} \left(-X_i Y_i + bX_i^2\right) = 0$. Solving for *b*, we find $b\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i$ and $b = \frac{\sum_i X_i Y_i}{\sum_i X_i^2}$.

The Standard Error for the Slope

To compute a confidence interval for β, we need to determine the variance of *b*, using the expected value theorems.

Since $b = \frac{\sum_i X_i Y_i}{\sum_i X_i^2}$, we compute $Var(b) = Var\left(\frac{\sum_i X_i Y_i}{\sum_i X_i^2}\right)$. Since the values of *X* are assumed to be fixed, $\sum X^2$ in the denominator is a constant. So,

$$Var\left(\frac{\sum_i X_i Y_i}{\sum_i X_i^2}\right) = \frac{1}{\left(\sum_i X_i^2\right)^2} Var\left(\sum_i X_i Y_i\right)$$

and $Var\left(\sum_i X_i Y_i\right) = Var\left(X_1 Y_1 + X_2 Y_2 + \cdots + X_n Y_n\right)$. The *X*'s are constants and we are interested in the variation of Y_i for the given X_i, which is the common variance σ_e^2.

So, $Var\left(X_1 Y_1 + X_2 Y_2 + \cdots + X_n Y_n\right) = X_1^2 Var(Y_1) + X_2^2 Var(Y_2) + \cdots + X_n^2 Var(Y_n) = \sigma_e^2 \sum_i X_i^2$.

Putting it all together, we find $Var(b) = Var\left(\frac{\sum_i X_i Y_i}{\sum_i X_i^2}\right) = \sigma_e^2\left(\frac{\sum_i X_i^2}{\left(\sum_i X_i^2\right)^2}\right) = \frac{\sigma_e^2}{\sum_i X_i^2}$. This is often written as $Var(b) = \frac{\sigma_e^2}{\sum_i (x_i - \bar{x})^2}$.

So, the standard error for the slope in regression can be estimated by

$$s_{\hat{b}} = \frac{s_e}{\sqrt{\sum_i (x_i - \bar{x})^2}} \text{ or } s_{\hat{b}} = \frac{s_e}{\sqrt{n-1}\, s_x}.$$

The Standard Error for $\hat{y}$, the Predicted Mean

Confidence intervals for a predicted mean can now be obtained. The standard error can be determined by computing $Var(\hat{y})$. We know that $\hat{y} = \bar{y} + b(x - \bar{x})$, so, as before, using the expected value theorems, we find

$$Var(\hat{y}) = Var\left(\bar{y} + b(x - \bar{x})\right) = Var(\bar{y}) + (x - \bar{x})^2 Var(b),$$

with $Var(b) = \frac{\sigma_e^2}{\sum (x_i - \bar{x})^2}$ and $Var(\bar{y}) = Var\left(\frac{\sum_i y_i}{n}\right) = \frac{1}{n^2} Var\left(\sum_i y_i\right) = \frac{n\sigma_e^2}{n^2} = \frac{\sigma_e^2}{n}$.

So,

$$Var(\hat{y}) = Var\left(\bar{y} + b(x - \bar{x})\right) = \frac{\sigma_e^2}{n} + \frac{\sigma_e^2 (x - \bar{x})^2}{\sum (x_i - \bar{x})^2}.$$

The standard error for predicting a mean response for a given value of *x* can be estimated by $s_{\hat{y}} = s_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$.

The Standard Error for the Intercept

The variance of the intercept *a* can be estimated using the previous formula for the standard error for $\hat{y}$. Since $\hat{y} = a + bx$, the variance of *a* is the variance of $\hat{y}$ when $x = 0$. So, $s_a = s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum (x_i - \bar{x})^2}}$.

The Standard Error for a Predicted Value

Finally, to predict a *y*-value, y_p, for a given *x*, we need to consider two independent errors. We know that *y* is normally distributed around $\mu_{y|x}$ so

$y \mid x \sim N(\mu_{y|x}, \sigma_e)$. Given $\mu_{y|x}$, we can estimate our error in predicting *y*. But, as we have just seen, there is also variation in our predictions of $\mu_{y|x}$. First, we predict $\hat{y}$ taking into account its own variation and then we use that prediction in predicting *y*. So

$$Var(y_p) = Var(\hat{y}) + Var(\varepsilon) = \left(\frac{\sigma_e^2}{n} + \frac{\sigma_e^2 (x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) + (\sigma_e^2) = \sigma_e^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right).$$

The standard error for this prediction can be estimated with

$$s_{y_p} = s_e^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right).$$

Now we have all the equations found in the texts.

Standard error for Slope:	$s_{\hat{b}} = \frac{s_e}{\sqrt{n-1}\, s_x}$
Standard error for Predicted Mean:	$s_{\hat{y}} = s_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$
Standard error for the Intercept:	$s_a = s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum (x_i - \bar{x})^2}}$
Standard error for a Predicted Value:	$s_{y_p} = s_e^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)$

Reference: Kennedy, Joh B. and Adam M. Neville, *Basic Statistical Methods for Engineers and Scientists*, 3rd, Harper and Row, 1986.