Deriving & Understanding the Variance Formulas

Max H. Farrell
BUS 41100
August 28, 2015

The purpose of this handout is to derive the variance formulas that we discussed in class and show why take the form they do. In class we went over the formulas at an intuitive level, and discussed which features of the data impacted them and in what ways. But why were those the right formulas? This handout tries to answer that question, which may also shed some light on how the formulas work.

One important issue will be the forming of prediction intervals, and the reason we care about $\mathbb{V}[e_f]$ versus $\mathbb{V}[\hat{Y}_f]$. The main punchline here is this: a prediction interval is a range of likely values for Y_f, not $\mathbb{E}[Y|X = X_f] = \beta_0 + \beta_1 X_f$.

Throughout, we will only look at simple linear regression. The formulas are easier to understand, and the intuition is entirely the same. (The formulas are the same too, just reinterpreting X as vectors or matrixes as appropriate.) Also, just like in class we will do everything *conditional* on X. That means we will treat the values $X_1, X_2, \ldots, X_n$ as fixed numbers; they are not random. This doesn't change any of the intuition either.

Contents

1 Notation

Remember that the model is

$$Y = \beta_0 + \beta_1 X + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Note that $\mathbb{V}[Y_i|X_i] = \sigma^2$.

The sample variances of X and Y are:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \text{and} \qquad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

The sample covariance is

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})Y_i.$$

The sample correlation is:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

2 Week 2: Sampling Distributions, Derivation 1

In this section I give a fairly simple derivation of the sampling distribution of the slope estimate b_1. Similarly derivations apply to the intercept estimate, b_0, a forecast, $\hat{Y}_f$, and anything from multiple linear regression.

Recall from week 1 that

$$b_1 = \frac{\text{corr}(X,Y)}{\text{var}(X)}.$$

Using this, plugging in the definition of $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, we get

$$\begin{aligned} b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} &= \frac{\sum_{i=1}^n (X_i - \bar{X})\left\{\beta_0 + \beta_1 X_i + \varepsilon_i\right\}}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \beta_0 \frac{\sum_{i=1}^n (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} + \beta_1 \frac{\sum_{i=1}^n (X_i - \bar{X})X_i}{\sum_{i=1}^n (X_i - \bar{X})^2} + \frac{\sum_{i=1}^n \left\{(X_i - \bar{X})\varepsilon_i\right\}}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= 0 + \beta_1 + \frac{\sum_{i=1}^n \left\{(X_i - \bar{X})\varepsilon_i\right\}}{\sum_{i=1}^n (X_i - \bar{X})^2}. \end{aligned}$$

The last equality comes from these two facts:

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0 \qquad \sum_{i=1}^{n}(X_i - \bar{X})X_i = \sum_{i=1}^{n}(X_i - \bar{X})^2,$$

which you can verify by direct calculation (just like you did on homework zero!).

So we have shown that

$$b_1 = \beta_1 + \frac{\sum_{i=1}^n \left\{(X_i - \bar{X})\varepsilon_i\right\}}{(n-1)s_x^2}.$$

The first term is just β_1, which is just some fixed number (even though we don't know it). The second term is Normally distributed, either by the Central Limit Theorem or because the ε_i are

assumed to be Normal. Therefore b_1 is also Normally distributed, but *what* Normal distribution? A Normal distribution is characterized by the mean and variance. The mean is easy: the second term has mean zero, because the ε_i have mean zero, and so the mean of b_1 is just β_1. To compute the variance, remember that β_1 is just a number, so it has no variance, and that we are treating the X_i as fixed numbers. Therefore:

$$\mathbb{V}[b_1] = \mathbb{V}\left[\frac{\sum_{i=1}^n \left\{(X_i - \bar{X})\varepsilon_i\right\}}{(n-1)s_x^2}\right] = \frac{1}{\left((n-1)s_x^2\right)^2}\sum_{i=1}^n \left\{(X_i - \bar{X})^2 \mathbb{V}\left[\varepsilon_i\right]\right\}.$$

Now, we use the assumption that the ε_i have **constant** variance, σ^2. Then we can pull it out of the summation, and we get the result from class:

$$\mathbb{V}[b_1] = \sigma^2 \frac{1}{\left((n-1)s_x^2\right)^2}\sum_{i=1}^n \left\{(X_i - \bar{X})^2\right\} = \sigma^2 \frac{1}{\left((n-1)s_x^2\right)^2}(n-1)s_x^2 = \sigma^2 \frac{1}{(n-1)s_x^2}.$$

We have thus shown that

$$b_1 \sim \mathcal{N}\left(\beta_1, \sigma^2 \frac{1}{(n-1)s_x^2}\right) \tag{1}$$

This exactly matches the result from class and it matches Equation (6) below.

2.1 Nonconstant variance

To see what happens with the variance is **not** constant, return to the penultimate step above:

$$\mathbb{V}[b_1] = \frac{1}{\left((n-1)s_x^2\right)^2}\sum_{i=1}^n \left\{(X_i - \bar{X})^2 \mathbb{V}\left[\varepsilon_i\right]\right\}.$$

Suppose every ε_i can have a different variance; call it σ_i^2. Then we **can't** pull anything out of summation! We just get:

$$\mathbb{V}[b_1] = \frac{1}{\left((n-1)s_x^2\right)^2}\sum_{i=1}^n \left\{(X_i - \bar{X})^2 \sigma_i^2\right\},$$

and therefore

$$b_1 \sim \mathcal{N}\left(\beta_1, \frac{\sum_{i=1}^n \left\{(X_i - \bar{X})^2 \sigma_i^2\right\}}{\left((n-1)s_x^2\right)^2}\right).$$

To deal with this, we either need to do a variance-stabilizing transformation or do heteroskedasticity robust inference; both of which we discuss in Week 4.

3 Week 2: Sampling Distributions, Derivation 2

Here I carefully derive all the variance formulas for b_0, b_1, and $\hat{Y}$ using a weighted-average representation that will prove very useful. This is the first thing that is introduced, in the next subsection.

3.1 Weighted Average Formulas for b_0, b_1, and $\hat{Y}$

We will show that b_0, b_1, and $\hat{Y}$ are all just weighted averages of the outcomes Y_i. This kind of makes sense in the following way: in regression our aim is to extract a "general, on-average" trend for Y given X. Even more precisely, we are estimating the conditonal expectation, and since expectations are just averages, it makes sense that our estimators are just averages.

This is more than just a nice coincidence, it's an important idea in terms of the type of estimation we're doing and how we get the results we do.

Begin with the slope coefficient, b_1. Recall from class that $b_1 = r_{xy}s_y/s_x$. Let's write out exactly what that means, and re-write the formula:

$$b_1 = r_{xy}\frac{s_y}{s_x} = \frac{s_{xy}}{s_x s_y}\frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2} = \frac{s_{xy}}{s_x^2} = \frac{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})Y_i}{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Now, we can recognize that this is really just a weighted sum of the values $Y_i, i = 1, \ldots, n$:

$$b_1 = \sum_{i=1}^n W_i Y_i, \qquad \text{where} \qquad W_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}. \tag{2}$$

This will be a very useful formula. It is also important to note that

$$\sum_{i=1}^n W_i = \frac{\sum_{i=1}^n (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n X_i - n\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n X_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = 0. \tag{3}$$

We can do the same thing for the intercept now too:

$$b_0 = \bar{Y} - b_1\bar{X} = \frac{1}{n}\sum_{i=1}^n Y_i - \sum_{i=1}^n W_i \bar{X} Y_i = \sum_{i=1}^n \left(\frac{1}{n} - W_i\bar{X}\right) Y_i. \tag{4}$$

And for a prediction for some new value X_f. Remember that we just read the prediction off the least squares line: $\hat{Y}_f = b_0 + b_1 X_f$. Therefore:

$$\hat{Y}_f = \sum_{i=1}^n \left(\frac{1}{n} - W_i\bar{X}\right) Y_i + \sum_{i=1}^n W_i Y_i X_f = \sum_{i=1}^n \left(\frac{1}{n} + W_i[X_f - \bar{X}]\right) Y_i \tag{5}$$

3.2 Variance Derivations for b_0 and b_1

It is now really simple to compute the variance of the intercept and slope coefficient. Remember that the X_i are fixed numbers and that the Y_i are independent of each other. Using the weighted sum formula (2):

$$\mathbb{V}[b_1] = \mathbb{V}\left[\sum_{i=1}^n W_i Y_i\right] = \sum_{i=1}^n \mathbb{V}\left[W_i Y_i\right] = \sum_{i=1}^n W_i^2 \mathbb{V}\left[Y_i\right] = \sigma^2 \sum_{i=1}^n W_i^2 = \sigma^2 \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^2}.$$

Canceling in the numerator and denominator we have

$$\mathbb{V}[b_1] = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sigma^2}{(n-1)s_x^2}. \tag{6}$$

This exactly matches the result from class and it matches Equation (1) above.

The intercept works almost exactly the same way. First, expand as follows.

$$\mathbb{V}[b_0] = \sigma^2 \sum_{i=1}^n \left(\frac{1}{n} - W_i\bar{X}\right)^2 = \sigma^2 \sum_{i=1}^n \left\{\frac{1}{n^2} + W_i^2\bar{X}^2 - 2\bar{X}W_i\right\} = \sigma^2 \left(\sum_{i=1}^n \frac{1}{n^2} + \bar{X}^2 \sum_{i=1}^n W_i^2 - 2\bar{X}\sum_{i=1}^n W_i\right\}.$$

The first term is just σ^2/n. For the second, the same canceling in the numerator and denominator of $\sum_{i=1}^n W_i^2$ that we did just above shows that this term is going to be $\sigma^2\bar{X}^2/(\sum_{i=1}^n (X_i - \bar{X})^2) = \sigma^2\bar{X}^2/[(n-1)s_x^2]$. Finally, the third term is zero because (3) tells us that $\sum_{i=1}^n W_i = 0$. Putting this together we get

$$\mathbb{V}[b_0] = \frac{\sigma^2}{n} + \frac{\sigma^2\bar{X}^2}{(n-1)s_x^2}. \tag{7}$$

We found the same thing in class, but sticking to the original formula $b_0 = \bar{Y} + b_1\bar{X}$. From this, it follows that $\mathbb{V}[b_0] = \mathbb{V}[\bar{Y}] + \bar{X}^2\mathbb{V}[b_1] + 2\bar{X}\text{Cov}(\bar{Y}, b_1)$ (remember $\bar{X}$ is fixed because we are conditioning on all the X_i). The first two terms are the same as in (7), and the third we argued intuitively was zero because the average didn't tell us anything about the slope, just shifted the line up or down in parallel. We have now formalized that idea.

3.3 Variance Derivations for $\hat{Y}_f$ and Prediction Intervals

First, let us derive $\mathbb{V}[\hat{Y}_f]$ and $\mathbb{V}[e_f]$, and then talk about what a prediction interval is.

From (5) and using the exact same tricks as in the previous subsection:

$$\mathbb{V}[\hat{Y}_f] = \sum_{i=1}^n \left(\frac{1}{n} + W_i[X_f - \bar{X}]\right)^2 \sigma^2 = \sum_{i=1}^n \frac{\sigma^2}{n^2} + \sigma^2(X_f - \bar{X})^2 \sum_{i=1}^n W_i^2 + 2\frac{\sigma^2(X_f - \bar{X})}{n}\sum_{i=1}^n W_i.$$

The first term is σ^2/n. The second term, after the same canceling in the numerator and denominator of $\sum_{i=1}^n W_i^2$ that we did above, becomes $\sigma^2(X_f - \bar{X})^2/[(n-1)s_x^2]$. As before, the third term is zero because of (3). Therefore

$$\mathbb{V}[\hat{Y}_f] = \sigma^2\left(\frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2}\right). \tag{8}$$

The estimator of this is called s_{fit}^2 in class. In class, we derived this straight from the prediction $\hat{Y}_f = b_0 + b_1X_f$, and got the exact same answer. There we had to use the covariance of b_0 and b_1, which can be derived using these same tricks again.

Finally, because the "new" observation X_f and Y_f is independent of all the others, and because $e_f = Y_f - \hat{Y}_f$, we have:

$$\mathbb{V}[e_f] = \mathbb{V}[Y_f] + \mathbb{V}[\hat{Y}_f] - \cancel{2\text{Cov}(Y_f, \hat{Y}_f)} = \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2}\right). \tag{9}$$

The estimator of this is called s^2_{pred} in class.

3.4 Understanding Prediction Intervals, Which are for Y_f

What is a prediction interval and why do we need$\mathbb{V}[e_f]$? A prediction is just like a confidence interval, it's *range of likely values*. But likely values of what? This is the key: it's a range of likely values for Y_f.

We think of 95% prediction/confidence intervals as being of the form:

> something $\pm$ 2 $\times$ (the std err of that something).

In fact, this is exactly how we form confidence intervals for β_1:

$$\mathbb{P}\left[\beta_1 \in \left(b_1 \pm 2 \times \sqrt{\mathbb{V}[b_1]}\right)\right] = 0.95.$$

Why? Because b_1 is our estimator of β_1, and since β_1 is a fixed (but unknown) number, $\mathbb{V}[b_1] = \mathbb{V}[b_1 - \beta_1]$. That is, the variance of our estimator is *the same as* the variance of the error we make.

What happens if we try to apply the same logic to $\hat{Y}_f$? Well, $\hat{Y}_f$ is our estimator of Y_f, but Y_f is *not a fixed value*! It's random, and we haven't observed it. So it is *not true* that $\mathbb{V}[\hat{Y}_f] = \mathbb{V}[Y_f - \hat{Y}_f]$.

So what we really want is a range of likely values for Y_f that is centered at $\hat{Y}_f$. That is our prediction interval, and it is formed as

$$\hat{Y}_f \pm 2 \times \sqrt{\mathbb{V}[e_f]}.$$

What would happen if we instead used $\hat{Y}_f \pm 2 \times \sqrt{\mathbb{V}[\hat{Y}_f]}$? This is a 95% confidence interval for $\mathbb{E}[Y|X = X_f] = \beta_0 + \beta_1 X_f$, that is, the average value for Y given that X is X_f. But we don't want a range of likely values for for the average value of Y_f, we want one for the actual of Y_f. The actual value (that we haven't observed yet) is the average value with the idiosyncratic shock up or down, and it's the uncertainty of this shock that we capture with the extra variability.

4 Week 4: Variance of e_j and leverage

[Note: in class we looked at the subscript i point, but here I'm changing that to j to not get confused with other notation, like $\sum_{i=1}^n$.]

To derive the variance of a given residual e_j, we have to re-walk the steps that we did for $\mathbb{V}[e_f]$, but with one **crucial**, yet **subtle difference**: for e_f, the "new" observation X_f and Y_f is **independent**

of the data used to form the guess $\hat{Y}_f$, but now, e_j is based on one of the points already in our data set.

Go back to Equation (9): the covariance term will **not cancel**, and we now have

$$\mathbb{V}[e_j] = \mathbb{V}[Y_j] + \mathbb{V}[\hat{Y}_j] - 2\text{Cov}(Y_j, \hat{Y}_j). \tag{10}$$

We already know $\mathbb{V}[Y_j] = \sigma^2$ from the definition of the model, and from Equation (8) (applied to Y_j instead of Y_f) we know that

$$\mathbb{V}[\hat{Y}_j] = \sigma^2 \left(\frac{1}{n} + \frac{(X_j - \bar{X})^2}{(n-1)s_x^2} \right). \tag{11}$$

So we just need to figure out the covariance term. Using the weighted average formula for $\hat{Y}_j$ from Equation (5) (applied to Y_j instead of Y_f), we find that

$$\begin{aligned} -2\text{Cov}(Y_j, \hat{Y}_j) &= -2\text{Cov}\left(Y_j \ , \ \sum_{i=1}^{n} \left(\frac{1}{n} + W_i[X_j - \bar{X}] \right) Y_i \right) \\ &= -2 \sum_{i=1}^{n} \left(\frac{1}{n} + W_i[X_j - \bar{X}] \right) \text{Cov}\left(Y_j \ , \ Y_i \right). \end{aligned}$$

But because all the observations are independent, $\text{Cov}(Y_j, Y_i) = 0$ *unless* $i = j$, and then you get $\text{Cov}(Y_j, Y_j) = \mathbb{V}[Y_j] = \sigma^2$. So only *one* term of the $\sum_{i=1}^{n}$ is left, and we get

$$-2\text{Cov}(Y_j, \hat{Y}_j) = -2 \left(\frac{1}{n} + W_j[X_j - \bar{X}] \right) \sigma^2 = -2 \left(\frac{1}{n} + \frac{(X_j - \bar{X})^2}{(n-1)s_x^2} \right) \sigma^2 \tag{12}$$

where the second equality just plugs in the definition of W_j. Plugging Equations (11) and (12) into (10), we get the final answer:

$$\begin{aligned} \mathbb{V}[e_j] = \mathbb{V}[Y_j] + \mathbb{V}[\hat{Y}_j] - 2\text{Cov}(Y_j, \hat{Y}_j) &= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(X_j - \bar{X})^2}{(n-1)s_x^2} \right) - 2 \left(\frac{1}{n} + \frac{(X_j - \bar{X})^2}{(n-1)s_x^2} \right) \sigma^2 \\ &= \sigma^2 \left[1 - \left(\frac{1}{n} + \frac{(X_j - \bar{X})^2}{(n-1)s_x^2} \right) \right], \end{aligned}$$

which is exactly the formula we have from week 4.