Recall the formula for $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{\sum X_c Y_c}{\sum X_c^2}$$

and that

$$Y_c = \beta_1 X_c + u$$

where β_1 is the true, population β_1. Plug the second equation into the first.

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum X_c Y_c}{\sum X_c^2} \\ &= \frac{\sum X_c (\beta_1 X_c + u)}{\sum X_c^2} \\ &= \frac{\sum X_c^2 \beta_1}{\sum X_c^2} + \frac{\sum X_c u}{\sum X_c^2} \\ &= \beta_1 + \frac{\sum X_c u}{\sum X_c^2}\end{aligned}$$

Let's assume that there are two observations.

$$\begin{aligned}\hat{\beta}_1 &= \beta_1 + \frac{X_{c_1} u_1}{X_{c_1}^2 + X_{c_1}^2} + \frac{X_{c_2} u_2}{X_{c_1}^2 + X_{c_1}^2} \\ &= \beta_1 + \frac{X_{c_1}}{X_{c_1}^2 + X_{c_1}^2} u_1 + \frac{X_{c_2}}{X_{c_1}^2 + X_{c_1}^2} u_2\end{aligned}$$

We shall assume for the moment that every time we collect a sample from the population, the values u change, but the values of X stay the same. For example, suppose we want to know the effect of education on wages. The model is

$$Wage = \beta_0 + \beta_1 Educyears + u$$

We might assume, for example, that in every sample the first person has 12 years of education and the second person has 16 years of education. Thus the X-values are *constant* from sample to sample. We also assume that from sample to sample the u-values *vary*. Thus u remains a *variable*. Recall the rules of variance. In listing the rules, X stands for any variable, and a and b stand for any constants.

$$\begin{aligned}Var(aX) &= a^2 Var(X) \\ Var(a + X) &= Var(X)\end{aligned}$$

and

$$Var(aX + bY) = a^2 var(X) + b^2 var(Y) + 2ab * cov(X, Y)$$

Also recall the definition of correlation.

$$Correl(X, Y) = \frac{cov(X, Y)}{sd(X) \cdot sd(Y)}$$

Therefore

$$cov(X, Y) = correl(X, Y) \cdot sd(X) \cdot sd(Y)$$

and

$$Var(aX + bY) = a^2 var(X) + b^2 var(Y) + 2ab * correl(X, Y) \cdot sd(X) \cdot sd(Y)$$

Let's return to the formula for the variance of $\hat{\beta}_1$ with two observations.

$$\hat{\beta}_1 = \beta_1 + \frac{X_{c_1}}{X_{c_1}^2 + X_{c_1}^2} u_1 + \frac{X_{c_2}}{X_{c_1}^2 + X_{c_1}^2} u_2$$

To simplify the calculations, we shall denote

$$d_1 = \frac{X_{c_1}}{X_{c_1}^2 + X_{c_1}^2}$$
$$d_2 = \frac{X_{c_2}}{X_{c_1}^2 + X_{c_1}^2}$$

Then

$$\hat{\beta}_1 = \beta_1 + d_1 u_1 + d_2 u_2$$

Keep in mind that we are treating X as a constant. Therefore d is a constant. Also keep in mind that β_1, the true, population parameter, is a constant, and therefore has no effect on the variance of $\hat{\beta}_1$. Now let's apply our rule for the weighted sum of variables.

$$Var(aX + bY) = a^2 var(X) + b^2 var(Y) + 2ab * correl(X,Y) \cdot sd(X) \cdot sd(Y)$$
$$Var\left(\beta_1 + d_1 u_1 + d_2 u_2\right) = d_1^2 var\left(u_1\right) + d_2^2 var\left(u_2\right) + 2 d_1 d_2 sd\left(u_1\right) sd\left(u_2\right) correl\left(u_1, u_2\right)$$

Usually errors are correlated in time series models. Since we will be studying cross-sectional models for most of this class, we shall assume that

$$correl\left(u_1, u_2\right) = 0$$

Then

$$var\left(\hat{\beta}_1\right) = d_1^2 var\left(u_1\right) + d_2^2 var\left(u_2\right)$$

Sometimes a further assumption is made.

$$var\left(u_1\right) = var\left(u_2\right)$$

This assumption is of **homoskedasticity**, which means "equal variances". Under this assumption, the formula for $var\left(\hat{\beta}_1\right)$ simplifies further.

$$\begin{aligned}
var\left(\hat{\beta}_1\right) &= d_1^2 var\left(u\right) + d_2^2 var\left(u\right) \\
&= var\left(u\right)\left(d_1^2 + d_2^2\right) \\
&= var\left(u\right)\left[\left(\frac{X_{c_1}}{X_{c_1}^2 + X_{c_1}^2}\right)^2 + \left(\frac{X_{c_2}}{X_{c_1}^2 + X_{c_1}^2}\right)^2\right] \\
&= var\left(u\right)\left[\frac{X_{c_1}^2 + X_{c_2}^2}{\left(X_{c_1}^2 + X_{c_1}^2\right)^2}\right] \\
&= \frac{var\left(u\right)}{\sum X_c^2}
\end{aligned}$$

The remaining question is how to estimate $var\left(u\right)$. The problem is that we don't observe the error terms, u. Therefore we shall estimate $var\left(u\right)$ using the residuals. The formula for this is

$$\widehat{var}\left(u\right) = \frac{\sum \hat{u}^2}{n - k}$$

where $\hat{u}$ is a residual, n is the number of observations, and k is the number of parameters we are estimating. Based on the simple model used so far, we shall estimate two parameters, β_0 and β_1, so $k = 2$.