Omitted Variable Bias Problem

Imagine we want to estimate the effect of education on wage. The simple model we would estimate is:

$$wage_i = \beta_0 + \beta_1 educ_i + \epsilon_i \tag{1}$$

To estimate β_1 we would calculate a covariance between wage and education:

$$Cov(wage_i; educ_i) = Cov(\beta_0 + \beta_1 educ_i + \epsilon_i; educ_i)$$

We can split it in three terms:

$$Cov(wage_i; educ_i) = Cov(\beta_0, educ_i) + Cov(\beta_1 educ_i; educ_i) + Cov(\epsilon_i, educ_i)$$

Since a covariance between a constant and a variable is zero $[cov(B, x_i) = 0]$ and a covariance of a variable with itself is a variance $[cov(x_i, x_i) = var(x_i)]$, we can simplify:

$$Cov(wage_i; educ_i) = \beta_1 Var(educ_i) + Cov(\epsilon_i, educ_i)$$

Deviding the whole expression by the variance of education we get:

$$\frac{Cov(wage_i; educ_i)}{Var(educ_i)} = \beta_1 + \frac{Cov(\epsilon_i, educ_i)}{Var(educ_i)}$$

If the OLS assumptions hold and $E[\epsilon_i | educ_i] = 0$ then $Cov(\epsilon_i, educ_i) = 0$ and we get that

$$\beta_1 = \frac{Cov(wage_i; educ_i)}{Var(educ_i)}$$

Now suppose epsilon contains experience so that the real models is:

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 exp_i + u_i, \quad where \quad E[u_i | educ_i, exp_i] = 0 \tag{2}$$

If we do not include experience in our regression and still run (1), what we get is the following:

$$\frac{Cov(wage_i; educ_i)}{Var(educ_i)} = \beta_1 + \frac{Cov(\epsilon_i, educ_i)}{Var(educ_i)} = \beta_1 + \frac{Cov(\beta_2 exp_i + u_i, educ_i)}{Var(educ_i)}$$

Since education and u_i are uncorrelated, what we get is:

$$\frac{Cov(wage_i; educ_i)}{Var(educ_i)} = \beta_1 + \beta_2 \frac{Cov(exp_i, educ_i)}{Var(educ_i)} \tag{3}$$

$\beta_2 \frac{Cov(exp_i, educ_i)}{Var(educ_i)}$ is the omitted variable bias. Notice that it depends on two components, β_2 and $\frac{Cov(exp_i, educ_i)}{Var(educ_i)}$ which is a coefficient of regressing experience on education. Therefore the OVB only shows up if: $\beta_2 \neq 0$ AND $\frac{Cov(exp_i, educ_i)}{Var(educ_i)} \neq 0$. That is the omitted variable is related to both outcome and the included regressor.