# Multivariate Regression: Introduction and Ceteris Paribus

EC420 MSU

Justin Kirkpatrick

Last updated January 18, 2021

# This lecture

MICHIGAN STATE UNIVERSITY

**Goal:**

1. Introduce two-variable (multivariate) regression
2. Motivate use of multivariate regression
3. Relate concepts from single variable to multivariate
4. Refine concept of *ceteris paribus*
5. Concept of "partialing out"
6. Extend multivariate from two to $K$ variables
7. Specification errors
    - Irrelevant variables and omitted variables

# What is multivariate regression?

Multivariate regression is the estimation of the PRF:

$$E[Y|X_1, X_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Where we previously had PRF:

$$E[Y|X] = \beta_0 + \beta_1 x$$

---

The SRF for two variables is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

We still have one error term, $u$

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + u_i$$

And we estimate $\hat{\beta} = \{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\}$ the same way.

## Examples

- We want to explain country-level *life expectancy* as a function of *gdp per capita* and *population growth*.

$$LifeExp_i = \beta_0 + \beta_1 gdppc_i + \beta_2 popgrowth_i + u_i$$

- We want to explain *mortality rate* with *number of cigarettes smoked* and *average daily caloric intake*

$$Mortality_i = \beta_0 + \beta_1 cigarettes_i + \beta_2 calories_i + u_i$$

- We want to explain *wage* with *education* and *ability*:

$$Wage_i = \beta_0 + \beta_1 educ_i + \beta_2 ability_i + u_i$$

## Ceteris Paribus - *all else held equal*

$$Wage_i = \beta_0 + \beta_1 educ_i + \beta_2 ability_i + u_i$$

We interpret $\beta_1$ as "the effect of $educ$ on the expectation of $wage$, *all else held equal*"

What other random variables are we holding equal:

- $ability$
- $u$ too!

## This means:

$$\beta_1 = \frac{\Delta Wage}{\Delta educ} \text{ when } \underbrace{\frac{\Delta Wage}{\Delta ability} = \frac{\Delta Wage}{\Delta u} = 0}_{\text{all else held equal}}$$

## And a similar interpretation for $\beta_2$

We interpret $\beta_2$ as "the effect of $\boldsymbol{experience}$ on the expectation of $\boldsymbol{wage}$, *all else held equal*"

## This means:

$$\beta_2 = \frac{\Delta Wage}{\Delta ability} \text{ when } \underbrace{\frac{\Delta Wage}{\Delta educ} = \frac{\Delta Wage}{\Delta u} = 0}_{\text{all else held equal}}$$

---

Does this require that $\frac{\Delta educ}{\Delta experience}$ be zero?

Nope. But we are measuring the effect of one *while holding the other equal to zero.*

## We can interpret $\beta_0$ as:

$\beta_0 = E[wage]$ when $educ$ and $ability$ **and** $u = 0$

## This is because:

$$E[wage|educ, ability] = \beta_0 + \beta_1 educ + \beta_2 ability$$

or, in general notation

$$E[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

A slightly different interpretation, and unique to the $\beta_0$.

## What if we *should* use two variables, but we only use one?

We could run the regression $wage_i = \beta_0 + \beta_1 educ_i + u_i$, but we probably think $ability_i$ also affects wages.

What if we don't observe $ability_i$?

- Just because we don't observe it doesn't mean it isn't affecting $wage_i$
- It *is present in the error term.*
- Let's make a new variable called $\tilde{u} = \delta_1 ability_i + u_i$

$$wage_i = \beta_0 + \beta_1 educ_i + \underbrace{\delta_1 ability_i + u_i}_{\tilde{u}_i}$$

- Note: usually the $\sim$ over a coefficient or variable (like $\tilde{u}$) will indicate it is related to, but different from, the non- $\sim$ version.

- $\delta_1$ is just the effect of $ability$ on $wage$

We can naively write this as a single variable regression:

$$wage_i = \beta_0 + \beta_1 educ_i + \tilde{u}_i$$

## But wait!

- We think $E[ability|educ] > 0$
  - Then this violates the assumption that $E[\tilde{u}|X] = 0$

Because

1. $\frac{\Delta \tilde{u}}{\Delta ability} = \delta_1 \neq 0$
2. $\frac{\Delta ability}{\Delta educ} \neq 0$

$\Rightarrow \frac{\Delta \tilde{u}}{\Delta educ} \neq 0$

## Bias

Recall that we could show $E[\hat{\beta}_1] = \beta_1$ if and only if $E[u|X] = 0$.

Looking at $\tilde{u}_i = \delta_1 ability_i + u_i$, we can see why $E[\tilde{u}|educ] \neq 0$

- Thus, $\beta_1$ in the single-variable regression was **biased**.

# Multivariate regression

Adding in $ability$ as a second variable fixes this:

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 ability_i + u_i$$

because $u_i$ does not change with $educ$ or $ability$.

Now, $E[u|X_1, X_2] = E[u|educ, ability] = 0$

## Multivariate regression allows us to account for the effect of both $ability$ and $educ$

$$wage = \beta_0 + \beta_1 educ + \beta_2 ability + u$$

And we can calculate the change in $wage$ from any change in $ability$ and $educ$ using the SRF:

$$\Delta\widehat{wage} = \widehat{\beta_1}\Delta educ + \widehat{\beta_2}\Delta ability$$

## This also means our assumption is now:

$$E[u|educ, wage] = 0$$

## Which we write in general as:

$$E[u|x_1, x_2] = 0$$

Which is to say that we think we've got everything that could potentially be correlated with $x_1$ and $x_2$ out of the error term.

Since we want to work with some sample data (wage2.dta from Wooldridge), let's replace $ability$ with $experience$, which is in the dataset....

# Multivariate regression

When we estimate $\boldsymbol{wage_i = \beta_0 + \beta_1 educ_i + \beta_2 experience_i + u_i}$, we are fitting a *plane*:

# Multivariate regression

When we estimate $\boldsymbol{wage_i = \beta_0 + \beta_1 educ_i + \beta_2 experience_i + u_i}$, we are fitting a *plane*:

# Multivariate regression

The best fit is no longer a line, but a *plane* with a slope in the $\boldsymbol{educ}$ and the $\boldsymbol{exper}$ axis of $\beta_1$ and $\beta_2$

Fitted values from a regression on the data in the previous slide where $\beta_{educ} = \mathbf{76.22}$ and $\beta_{exper} = \mathbf{17.64}$.

## How do we estimate $\hat{\beta}$?

Remember our two assumptions that let us derive $\hat{\beta}_1$ and $\hat{\beta}_0$:

$$E[u] = 0, \quad E[u|x] = 0$$

Now, we want to estimate $\hat{\beta} = \{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\}$

- And we have two $x$'s: $x_1$ and $x_2$.

$$E[u] = 0, \quad E[u|x_1] = 0, \quad E[u|x_2] = 0$$

We have **three** moment conditions, and three unknowns to estimate. We can do that!

These three moment conditions give us the following to start with:

$$E[y_i - \beta_0 - \beta_1 x_1 - \beta_1 x_2] = E[u] = 0$$

$$E[x_1(y_i - \beta_0 - \beta_1 x_1 - \beta_1 x_2)] = E[x_1 u] = 0$$

$$E[x_2(y_i - \beta_0 - \beta_1 x_1 - \beta_1 x_2)] = E[x_2 u] = 0$$

But don't worry, we won't derive them directly from this, but that's how we would do it.

## Let's talk notation for a second:

- I will use $\beta$ as the coefficients we are estimating ( $\hat{\beta}$ )
- When talking about the right hand side (the covariates), I'll either call them $x_1$, $x_2, \cdots$
  - Or sometimes just using the variable names: $wage = \beta_0 + \beta_1 educ$
  - Or sometimes with subscripts: $y = \beta_0 + \beta_{educ} x_{educ} + u$
- And sometimes, if we want to emphasize that two regressions are wholly different, I will use $\delta$ or $\gamma$ instead of $\beta$
- $u$ and $v$ will represent errors in two different regressions

## Partialing out

Imagine if we had two $x$'s, $x_{temp}$ and $x_{rain}$ that both had an effect on $y$, but were closely related.

We could look at $\beta_{temp}$ in

$$y = \beta_0 + \beta_{temp} x_{temp} + u$$

And $\beta_{rain}$ in

$$y = \beta_0 + \beta_{rain} x_{rain} + u$$

We learned from the previous section that $\beta_{temp}$ is biased when $x_{rain}$ is in the error term $u$, and vice versa.

That is, $\beta_{temp}$ is going to "pick up some of the effect" of $x_{rain}$.

MICHIGAN STATE UNIVERSITY

## Partialing out

So, would $\tilde{\beta}_{temp}$ in the following (correct) specification equal $\beta_{temp}$ from the previous slide?

$$y = \tilde{\beta}_0 + \tilde{\beta}_{temp} x_{temp} + \tilde{\beta}_{rain} x_{rain} + \tilde{u}$$

## No!

Once we include both variables, we will get a different estimate for $\tilde{\beta}$ than before since each effect is isolated (*ceteris paribus*)

So, to calculate the correct $\tilde{\beta}_{temp}$, we need estimates that take the effect of $\tilde{\beta}_{rain}$ into consideration.

This is called **partialing out**.

## One way we can estimate unbiased $\beta_{temp}$ is the following way:

First, estimate the regression of $x_{temp}$ on $x_{rain}$

$$x_{temp} = \delta_0 + \delta_{rain} x_{rain} + v$$

Couple of things:

- $x_{temp}$ is on the left hand side.
- **We are "explaining temperature with rainfall"**
- Using $\delta$ to show that these are different coefficients

That error term, $v$ has an interpretation

- $v$ is the $temp$ that is **not explained by** $rain$
- $\delta_{rain} x_{rain}$ is the $temp$ that **is** explained by $rain$

$v$ is $temp$ that has had $rain$ "partialed out"

## Of course, we have a sample analog for $v$, the SRF residuals:

$$\hat{v} = x_{temp} - (\hat{\delta}_0 + \hat{\delta}_{rain} x_{rain})$$

Remember, $v$ still varies along with $x_{temp}$, but it is not correlated at *all* with $x_{rain}$.

Now, if we want to get the correct value for $\beta_{temp}$ in the full regression:

$$y = \gamma_0 + \gamma_1 \hat{v} + u$$

- We do not use $x_{rain}$ directly.
- We use $\hat{v}$ and leave $x_{rain}$ out.
- $\hat{v}$ is correlated with $x_{temp}$, but not with $x_{rain}$
- Put another way, $\hat{v}$ contains only the part of $x_{temp}$ that is not correlated with $x_{rain}$.

One can show that $\gamma_1 = \tilde{\beta}_{temp}$

## One can show that $\gamma_1 = \tilde{\beta}_{temp}$, the unbiased estimate.

That is, we get the (unbiased) coefficient one would get from regressing

$$\boldsymbol{y} = \tilde{\beta}_0 + \tilde{\beta}_{temp}\boldsymbol{x}_{temp} + \tilde{\beta}_{rain}\boldsymbol{x}_{rain} + \tilde{\boldsymbol{u}}$$

by first "partialing" $\boldsymbol{x}_{rain}$ out of $\boldsymbol{x}_{temp}$ then regressing what is left on $\boldsymbol{y}$.

Similarly, one can do the same for $\boldsymbol{x}_{rain}$:

$$\boldsymbol{x}_{rain} = \boldsymbol{\kappa}_0 + \boldsymbol{\kappa}_1 \boldsymbol{x}_{temp} + \boldsymbol{w}$$

Then use the residuals, $\hat{\boldsymbol{w}}$, in a regression:

$$\boldsymbol{y} = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1 \hat{\boldsymbol{w}} + \boldsymbol{\epsilon}$$

And $\tilde{\beta}_{rain} = \boldsymbol{\alpha}_1$

Since $\hat{\tilde{\beta}}_{rain} = \hat{\alpha}_1 = \frac{Cov(y,\hat{w})}{Var(\hat{w})}$, we can say that $\beta_{rain}$ is the effect of $\boldsymbol{x}_{rain}$ *once we've taken out the effect of* $\boldsymbol{x}_{temp}$ and vice versa.

# Multivariate regression

Since we get the same $\hat{\beta}_1$ if we

- Partial out the effect of $x_2$ and run a single variable regression, or
- Run a two-variable regression

Then we can think of the $\tilde{\beta}_1$ in:

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \tilde{u}$$

As the effect of $x_1$ on $y$ **after partialing** $x_2$ **out of** $x_1$, and vice versa.

## Multivariate regression automatically partials out each of the $x's$.

## Let's compare a simple and multiple regression estimates

- $y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \tilde{u}$
- $y = \beta_0 + \beta_1 x_1 + u$

How will $\tilde{\beta}_1$ differ from $\beta_1$?

It depends on the relationship between $x_2$ and $x_1$

$$x_2 = \delta_0 + \delta_1 x_1 + v$$

If we take $x_2 = \delta_0 + \delta_1 x_1 + v$ and substitute it into the first equation above:

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 (\delta_0 + \delta_1 x_1 + v) + \tilde{u}$$

$$y = \tilde{\beta}_0 + \tilde{\beta}_2 \delta_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 \delta_1 x_1 + \tilde{\beta}_2 v + \tilde{u}$$

$$y = \tilde{\tilde{\beta}}_0 + (\tilde{\beta}_1 + \tilde{\beta}_2 \delta_1) x_1 + \tilde{v}$$

MICHIGAN STATE UNIVERSITY

Therefore it is true that:

- $\hat{\beta}_1 = \hat{\tilde{\beta}}_1 + \hat{\tilde{\beta}}_2 \hat{\delta}_1$
    - In words: to whatever extent $x_1$ and $x_2$ are correlated ( $\delta_1$ ), our naive $\hat{\beta}_1$ will include that correlation.

Knowing this, when would the simple regression estimate $\hat{\beta}_1$ **equal** the multiple regression (multivariate) estimate $\hat{\tilde{\beta}}_1$ ?

## Will this hold empirically?

Will $\hat{\beta}_1$ change when you add in $\hat{\beta}_2$?

| | Model 1 | Model 2 |
|---|---|---|
| (Intercept) | 146.952 | -272.528 |
| | (77.715) | (107.263) |
| educ | 60.214 | 76.216 |
| | (5.695) | (6.297) |
| exper | | 17.638 |
| | | (3.162) |

$$wage = \beta_0 + \beta_1 educ + u$$
$$wage = \tilde{\beta}_0 + \tilde{\beta}_1 educ + \tilde{\beta}_2 exper + \tilde{u}$$

## Will this hold empirically?

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 23.78 | 0.79 | 30.03 | 0 |
| educ | -0.91 | 0.06 | -15.63 | 0 |

$$exper = \delta_0 + \delta_1 educ + u$$

and

$$60.21 = 76.22 + (\underbrace{17.64}_{\hat{\tilde{\beta}}_2} \times \underbrace{-.91}_{\hat{\delta}_1})$$

## We can still use the same formula for R^2

$$R^2 = \frac{SSE}{SST} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

This is because $R^2$ only uses the fit of the whole model, determined by how well $\hat{y}$ fits.

MICHIGAN STATE UNIVERSITY

## Multiple regression is easily extended to many variables.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

And multiple variables, we can extend the partialing out in the following manner:

- $x_1 = \delta_0 + \beta_1 x_2 + \beta_2 x_3 + \cdots + \beta_k x_{k+1} + v$
  - $v$ is the part of $x_1$ that has had $x_2, x_3, \cdots$ partialed out
- $y = \alpha_0 + \alpha_1 \hat{v}$
- $\beta_1$ = $\alpha_1$

You can "partial out" multiple variables, leaving only variation that is uncorrelated with the other variables.

## OLS is easily extended from 2 to >2 variables

## We might be worried about two *specification errors*

- Including an irrelevant variable.
  - Suppose the "true model" is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

  - And we estimate:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

  - Then OLS is still an unbiased estimator, since *unbaisedness* holds regardless of the true value of the parameters, even if $\beta_j = 0$ for some $j$.
  - Including an irrelevant variable will, however, impact the variance of the OLS estimator.

# We might be worried about two *specification errors*

- Omitting a relevant variable.

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + u$$

  - We showed that $\tilde{\beta}_1 = \beta_1$ only when $\beta_2 = 0$ or $\delta_1 = 0$.
  - Size and direction depend on the sign and size of $\beta_2\delta_1$, which depends on the relationship of the omitted variable and the included variable, $x_1$, and the outcome variable, $y$.
  - With multiple regressors, the sign and size may not be clear.
  - We can usually *"sign the bias"* if we
    1. have an idea of what is omitted,
    2. have an idea of how it's correlated with $y$, and
    3. have an idea of how it's correlated with one or more of $x_1, x_2, \cdots, x_k$