Multiple Linear Regression: Inference

EC 320: Introduction to Econometrics

Philip Economides

Winter 2022

Prologue

Review

Suppose that an epidemiologist studies the effect of coffee consumption on cardiovascular health by estimating

$$\text{Health}_i = \beta_1 + \beta_2 \text{Coffee}_i + u_i.$$

1. What do we have to assume to interpret β_2 as the true effect of coffee consumption on health?

2. What omitted variables would bias the estimator of β_2?

3. For each omitted variable, how would you sign the bias?

OLS Variances

OLS Variances

Multiple regression model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$.

The variance of a slope estimator $\hat{\beta}_j$ on an independent variable X_j is

$$\mathbf{Var}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{\left(1 - R_j^2\right) \sum_{i=1}^{n} \left(X_{ji} - \bar{X}_j\right)^2},$$

where R_j^2 is the R^2 from a regression of X_j on the other independent variables and an intercept.

OLS Variances

$$\mathrm{Var}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{\left(1 - R_j^2\right) \sum_{i=1}^{n} \left(X_{ji} - \bar{X}_j\right)^2}$$

Moving parts

1. **Error variance:** As σ^2 increases, $\mathrm{Var}\left(\hat{\beta}_j\right)$ increases.
2. **Total variation in X_j:** As $\sum_{i=1}^{n} \left(X_{ji} - \bar{X}_j\right)^2$ increases, $\mathrm{Var}\left(\hat{\beta}_j\right)$ decreases.
3. **Relationships between independent variables:** As R_j^2 increases, $\mathrm{Var}\left(\hat{\beta}_j\right)$ increases.

Multicollinearity

Suppose that we want to understand the relationship between crime rates and poverty rates in US cities. We could estimate the model

$$\textbf{Crime}_i = \beta_0 + \beta_1 \textbf{Poverty}_i + \beta_2 \textbf{Income}_i + u_i,$$

where $\textbf{Income}_i$ controls for median income in city i.

Before obtaining standard errors and conducting hypothesis tests, we need:

$$\textbf{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\left(1 - R_1^2\right) \sum_{i=1}^n \left(\textbf{Poverty}_i - \overline{\textbf{Poverty}}\right)^2}.$$

R_1^2 is the R^2 from a regression of poverty on median income:

$$\textbf{Poverty}_i = \gamma_0 + \gamma_1 \textbf{Income}_i + v_i.$$

Multicollinearity

Scenario 1: If Income_i explains most of the variation in Poverty_i, then R_1^2 will approach one.

- If R_1^2 is one, then Poverty_i and Income_i are perfectly collinear (violates the *no perfect collinearity* assumption).

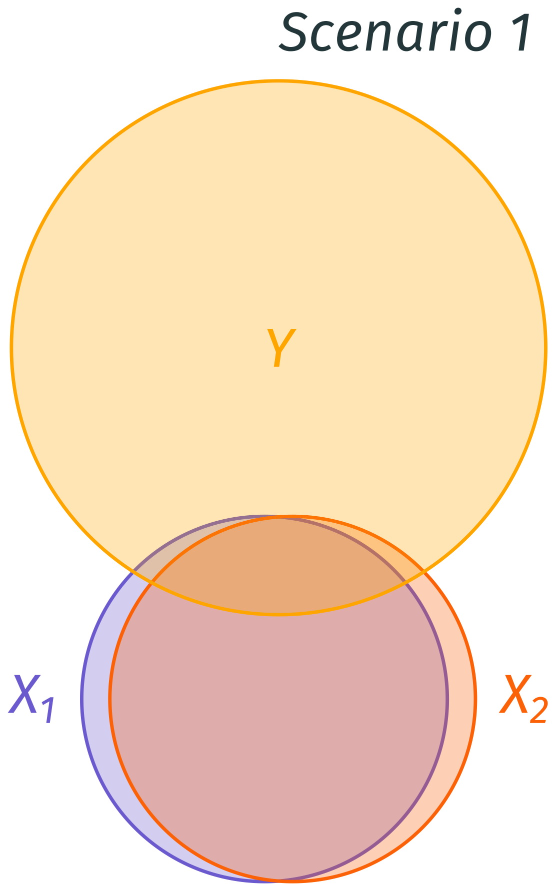
Scenario 2: If Income_i explains none of the variation in Poverty_i, then R_1^2 is zero.

Question: In which scenario is the variance of the poverty coefficient smaller?

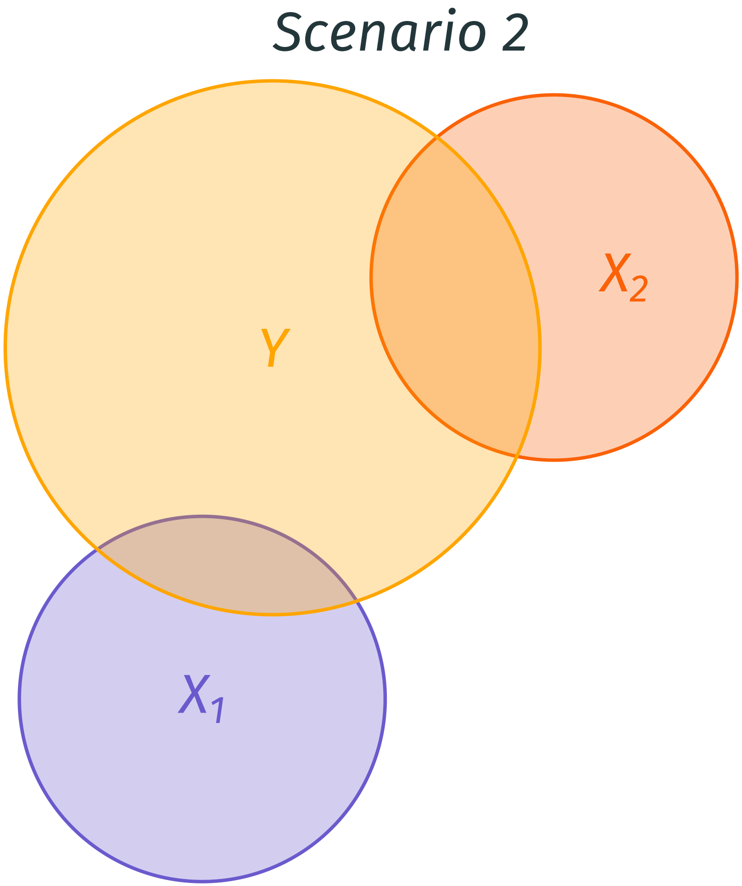
$$\text{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\left(1 - R_1^2\right) \sum_{i=1}^{n} \left(\text{Poverty}_i - \overline{\text{Poverty}}\right)^2}$$

Answer: Scenario 2.

Multicollinearity



Multicollinearity



Multicollinearity

As the relationships between the variables increase, R_j^2 increases.

For high R_j^2, $\mathbf{Var}\left(\hat{\beta}_j\right)$ is large:

$$\mathbf{Var}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{\left(1 - R_j^2\right) \sum_{i=1}^n \left(X_{ji} - \bar{X}_j\right)^2}.$$

This phenomenon is known as **multicollinearity**.

- Some view multicollinearity as a "problem" to be solved.
- Can increase n or drop independent variables that are highly related to the others.
- **Warning:** Dropping variables can generate omitted variable bias.

Multicollinearity

Example: Effect of different types of school spending on high school graduation rates.

$$\begin{aligned}\text{Graduation}_i = \beta_0 &+ \beta_1 \text{Salaries}_i + \beta_2 \text{Athletics}_i \\ &+ \beta_3 \text{Textbooks}_i + \beta_4 \text{Facilities}_i + u_i\end{aligned}$$

- Schools that spend more on teachers also tend to spend more on athletic programs, textbooks, and building maintenance.
- While total spending likely has a statistically significant effect on graduation rates, might not be able to detect statistically significant effects for individual line items.

Potential solutions: Re-define research question to consider the effect of total spending on graduation rates *or* gather more data to decrease OLS variances (*i.e.*, increase n).

Irrelevant Variables

Suppose that the true relationship between birth weight and *in utero* exposure to toxic air pollution is

$$(\textbf{Birth Weight})_i = \beta_0 + \beta_1 \textbf{Pollution}_i + u_i.$$

Suppose that, instead of estimating the "true model," an analyst estimates

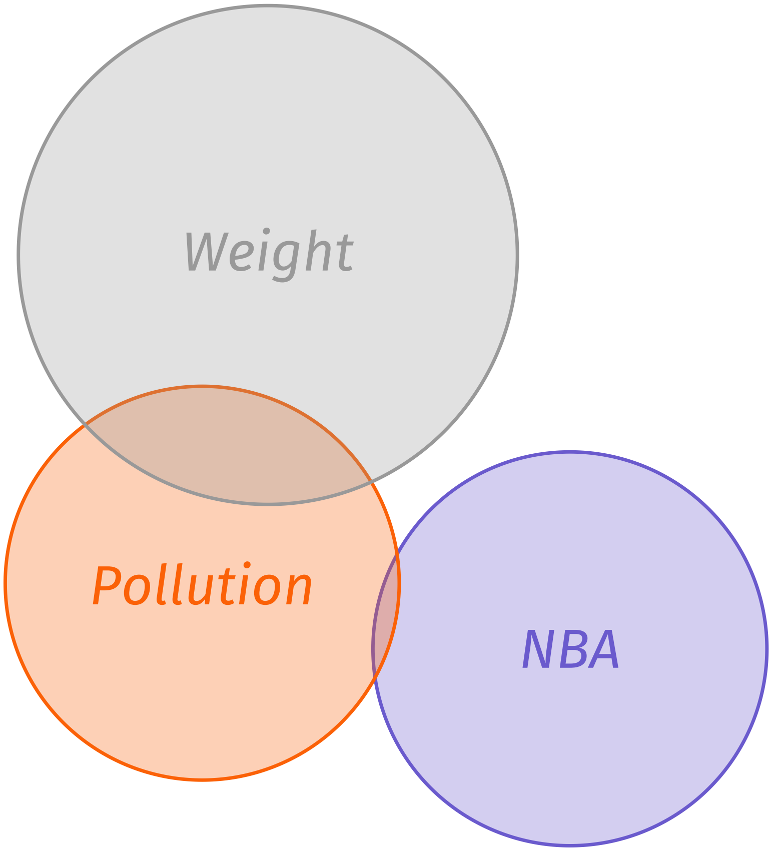
$$(\textbf{Birth Weight})_i = \tilde{\beta}_0 + \tilde{\beta}_1 \textbf{Pollution}_i + \tilde{\beta}_2 \textbf{NBA}_i + u_i,$$

where $\textbf{NBA}_i$ is the record of the nearest NBA team during the season before birth.

One can show that $\mathbb{E}\left(\hat{\tilde{\beta}}_1\right) = \beta_1$ (*i.e.*, $\hat{\tilde{\beta}}_1$ is unbiased).

However, the variances of $\hat{\tilde{\beta}}_1$ and $\hat{\beta}_1$ differ.

Irrelevant Variables



Irrelevant Variables

The variance of $\hat{\beta}_1$ from estimating the "true model" is

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^n \left(\mathrm{Pollution}_i - \overline{\mathrm{Pollution}}\right)^2}.$$

The variance of $\hat{\tilde{\beta}}_1$ from estimating the model with the irrelevant variable is

$$\mathrm{Var}\left(\hat{\tilde{\beta}}_1\right) = \frac{\sigma^2}{(1 - R_1^2) \sum_{i=1}^n \left(\mathrm{Pollution}_i - \overline{\mathrm{Pollution}}\right)^2}.$$

Notice that $\mathrm{Var}\left(\hat{\beta}_1\right) \leq \mathrm{Var}\left(\hat{\tilde{\beta}}_1\right)$.

Including irrelevant control variables can increase OLS variances!

Estimating Error Variance

We cannot observe σ^2, so we must estimate it using the residuals from an estimated regression:

$$s_u^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1}$$

- $k+1$ is the number of parameters (one "slope" for each X variable and an intercept).
- $n-k-1$ = degrees of freedom.
- Using the first 5 OLS assumptions, one can prove that s_u^2 is an unbiased estimator of σ^2.

Standard Errors

The formula for the standard error is the square root of $\mathbf{Var}\left(\hat{\beta}_j\right)$:

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\frac{\boldsymbol{s_u^2}}{(1 - R_j^2)\sum_{i=1}^n (X_{ji} - \bar{X}_j)^2}}.$$

Inference

OLS Classical Assumptions

1. **Linearity:** The population relationship is linear in parameters with an additive error term.
2. **Sample Variation:** There is variation in X.
3. **Exogeneity:** The X variable is exogenous (*i.e.*, $\mathbb{E}(u|X) = 0$).
4. **Homoskedasticity:** The error term has the same variance for each value of the independent variable (*i.e.*, $\mathbf{Var}(u|X) = \sigma^2$).
5. **Non-Autocorrelation:** Any pair of error terms share zero correlation due to having been independently drawn. (*i.e.*, $\mathbb{E}\left(u_i u_j\right) = 0 \; \forall i \text{ s.t. } i \neq j$).
6. **Normality:** The population error term is normally distributed with mean zero and variance σ^2 (*i.e.*, $u \sim N(0, \sigma^2)$)

1-3 imply **unbiasedness.**

1-5 imply **efficiency.**

Normality

With the first five assumptions, normality buys us a **sampling distribution** for $\hat{\beta}_j$:

- $\hat{\beta}_j \sim \text{Normal}\left(\beta_j,\ \text{Var}\left(\hat{\beta}_j\right)\right)$

- $\frac{\hat{\beta}_j - \beta_j}{\sqrt{\text{Var}\left(\hat{\beta}_j\right)}} \sim \text{Normal}(0, 1)$

Common violations: **autocorrelation** and **spatially correlated errors**.

Sampling Distribution

In practice, we can only estimate σ^2, so we use the t distribution:

- $\frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}\left(\hat{\beta}_j\right)} \sim t_{n-k-1} = t_{\mathrm{df}}$.
- Use this to construct t-statistics and conduct hypothesis testing.

Where are the critical values?

- Critical values describe specific quantiles of the t_{df} distribution.
- t_{df} is the entire sampling distribution.

Hypothesis Testing

Conduct a one-sided (right tail) test at the 5% level.

```
lm(read4 ~ lexppp + lunch, data = meap01) %>% tidy()
```

```
#> # A tibble: 3 x 5
#>   term        estimate std.error statistic   p.value
#>   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
#> 1 (Intercept)  -14.0     14.2       -0.989 3.23e-  1
#> 2 lexppp        10.8      1.68       6.45  1.40e- 10
#> 3 lunch         -0.463    0.0136   -33.9   5.72e-196
```

H_0: $\beta_{\text{Spend}} = 0$ *vs.* H_a: $\beta_{\text{Spend}} > 0$

$t_{\text{stat}} = 6.45$ and $t_{0.95,\ 1823-3} = 1.65$

Reject H_0 if $t_{\text{stat}} = 6.45 > t_{0.95,\ 1823-3} = 1.65$.

Statement is true, so we **reject H_0** at the 5% level.

Hypothesis Testing

Conduct a one-sided (left tail) test at the 5% level.

```
lm(read4 ~ lexppp + lunch, data = meap01) %>% tidy()
```

```
#> # A tibble: 3 x 5
#>   term        estimate std.error statistic   p.value
#>   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
#> 1 (Intercept)  -14.0     14.2       -0.989 3.23e-  1
#> 2 lexppp        10.8      1.68       6.45  1.40e- 10
#> 3 lunch         -0.463    0.0136   -33.9   5.72e-196
```

H$_0$: $\beta_{\text{Spend}} = 0$ vs. H$_a$: $\beta_{\text{Spend}} < 0$

$t_{\text{stat}} = 6.45$ and $t_{0.95,\,1823-3} = 1.65$

Reject H$_0$ if $t_{\text{stat}} = 6.45 < -t_{0.95,\,1823-3} = -1.65$.

Statement is false, so we **fail to reject H$_0$** at the 5% level.

Hypothesis Testing

Conduct a two-sided test at the 5% level.

```
lm(read4 ~ lexppp + lunch, data = meap01) %>% tidy()
```

```
#> # A tibble: 3 x 5
#>   term        estimate std.error statistic   p.value
#>   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
#> 1 (Intercept)  -14.0      14.2      -0.989 3.23e-  1
#> 2 lexppp        10.8       1.68      6.45  1.40e- 10
#> 3 lunch         -0.463     0.0136  -33.9   5.72e-196
```

H_0: $\beta_{\text{Spend}} = 0$ *vs.* H_a: $\beta_{\text{Spend}} \neq 0$

$t_{\text{stat}} = 6.45$ and $t_{0.975,\,1823-3} = 1.96$

Reject H_0 if $|t_{\text{stat}}| = |6.45| > t_{0.975,\,1823-3} = 1.96$.

Statement is true, so we **reject H_0** at the 5% level.

Hypothesis Testing

Conduct a two-sided test at the 5% level.

```
lm(read4 ~ lexppp + lunch, data = meap01) %>% tidy()
```

```
#> # A tibble: 3 x 5
#>   term        estimate std.error statistic   p.value
#>   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
#> 1 (Intercept)  -14.0     14.2       -0.989 3.23e-  1
#> 2 lexppp        10.8      1.68       6.45  1.40e- 10
#> 3 lunch         -0.463    0.0136   -33.9   5.72e-196
```

H_0: $\beta_{\textbf{Lunch}} = -1$ vs. H_a: $\beta_{\textbf{Lunch}} \neq -1$

$t_{\text{stat}} = \frac{\hat{\beta}_{\text{Lunch}} - \beta^0_{\text{Lunch}}}{\text{SE}(\hat{\beta}_{\text{Lunch}})} = 39.49$ and $t_{0.975,\, 1823-3} = 1.96$

Reject H_0 if $|t_{\text{stat}}| = |39.49| > t_{0.975,\, 1823-3} = 1.96$.

Statement is true, so we **reject H_0** at the 5% level.

F Tests

***t* tests** allow us to test simple hypotheses involving a single parameter.

- *e.g.*, $\beta_1 = 0$ or $\beta_2 = 1$.

***F* tests** allow us to test hypotheses that involve multiple parameters.

- *e.g.*, $\beta_1 = \beta_2$ or $\beta_3 + \beta_4 = 1$.

F Tests

Example

Economists often say that "money is fungible."

We might want to test whether money received as income actually has the same effect on consumption as money received from tax credits.

$$\textbf{Consumption}_i = \beta_0 + \beta_1 \textbf{Income}_i + \beta_2 \textbf{Credit}_i + u_i$$

F Tests

Example, continued

We can write our null hypothesis as

$$H_0:\ \beta_1 = \beta_2 \iff H_0:\ \beta_1 - \beta_2 = 0$$

Imposing the null hypothesis gives us a **restricted model**

$$\text{Consumption}_i = \beta_0 + \beta_1 \text{Income}_i + \beta_1 \text{Credit}_i + u_i$$

$$\text{Consumption}_i = \beta_0 + \beta_1 \left(\text{Income}_i + \text{Credit}_i\right) + u_i$$

F Tests

Example, continued

To test the null hypothesis $H_o: \ \beta_1 = \beta_2$ against $H_a: \ \beta_1 \neq \beta_2$, we use the F statistic

$$F_{q,\, n-k-1} = \frac{\left(\text{RSS}_r - \text{RSS}_u\right)/q}{\text{RSS}_u/(n-k-1)}$$

which (as its name suggests) follows the F distribution with q numerator degrees of freedom and $n-k-1$ denominator degrees of freedom.

Here, q is the number of restrictions we impose via H_0.

F Tests

Example, continued

The term RSS_r is the sum of squared residuals (RSS) from our **restricted model**

$$\text{Consumption}_i = \beta_0 + \beta_1 \left(\text{Income}_i + \text{Credit}_i\right) + u_i$$

and RSS_u is the sum of squared residuals (RSS) from our **unrestricted model**

$$\text{Consumption}_i = \beta_0 + \beta_1 \text{Income}_i + \beta_2 \text{Credit}_i + u_i$$

F Tests

Finally, we compare our F-statistic to a critical value of F to test the null hypothesis.

If $F > F_{\text{crit}}$, then reject the null hypothesis at the $\alpha \times 100$ percent level.

- Find F_{crit} in a table using the desired significance level, numerator degrees of freedom, and denominator degrees of freedom.

Aside: Why are F-statistics always positive?

F Tests

RSS is usually a large cumbersome number.

Alternative: Calculate the F-statistic using R^2.

$$F = \frac{\left(R_u^2 - R_r^2\right)/q}{(1 - R_u^2)/(n-k-1)}$$

Where does this come from?

- $\text{TSS} = \text{RSS} + \text{ESS}$
- $R^2 = \text{ESS}/\text{TSS}$
- $\text{RSS}_r = \text{TSS}(1 - R_r^2)$
- $\text{RSS}_u = \text{TSS}(1 - R_u^2)$

Application: Hedonic Modeling

Hedonic Modeling

Questions

- How much are home buyers willing to pay for houses with additional bedrooms?
- How much salary are workers willing to give up in exchange for safer working conditions?
- What is the market value of my neighbor's house?

Answers?

Hedonic modeling is a specific application of multiple regression.

- Prices or wages on the left hand side.
- Attributes of a good or a job on the right-hand side.
- Use coefficient estimates and fitted values.

Hedonic Modeling

Example

Using data on home sales, you run a regression and obtain the fitted model

$$\hat{\text{Price}}_i = 75000 + 50 \cdot (\text{Sq. ft.})_i + 16000 \cdot \text{Bedrooms}_i + 10000 \cdot \text{Bathrooms}_i$$

What is the forecasted price of a 1000-square-foot house with 1 bedroom and 1 bathroom?

$$\hat{\text{Price}} = 75000 + 50 \cdot (1000) + 16000 \cdot (1) + 10000 \cdot (1) = 1.51 \times 10^5$$

A homeowner is thinking about adding 1500 square feet to their home with 3 more bedrooms and an additional bathroom. How much extra money could she expect if she completed the addition and sold her home?

$$\Delta\text{Price} = 50 \cdot (1500) + 16000 \cdot (3) + 10000 \cdot (1) = 1.33 \times 10^5$$