# Multiple Linear Regression: Estimation

## EC 320: Introduction to Econometrics

Tami Ren

Summer 2022

# Multiple Linear Regression

# Multiple Linear Regression

## More explanatory variables

**Simple linear regression** features one outcome variable and one explanatory variable:

$$Y_i = \beta_0 + \beta_1 X_i + u_i.$$

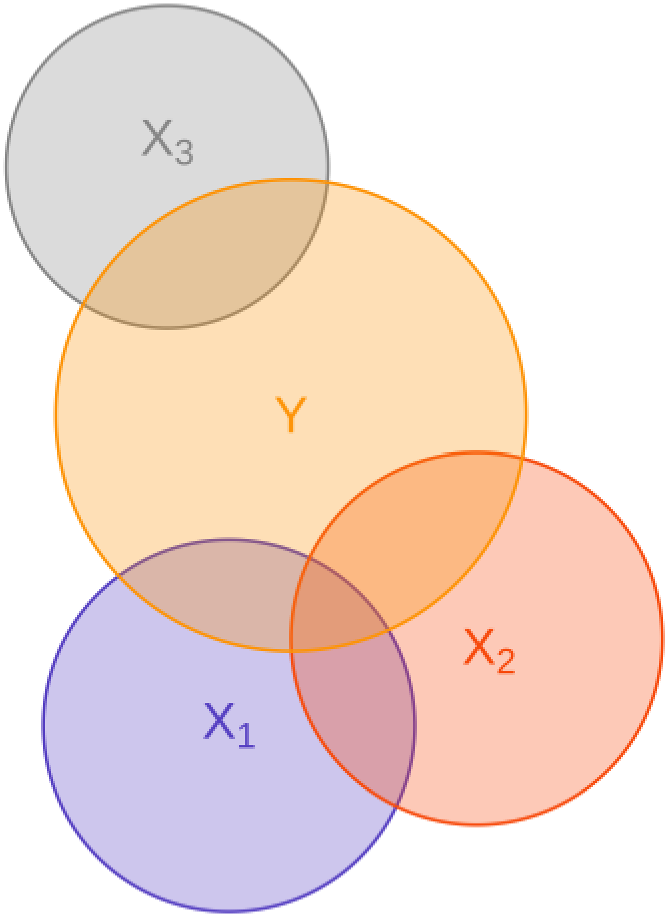
**Multiple linear regression** features one outcome variable and multiple explanatory variables:

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

**Why?**

- Better explain the variation in $Y$.
- Improve predictions.
- Avoid bias.

# Multiple Linear Regression



# OLS Estimation

As was the case with simple linear regressions, OLS minimizes the sum of squared residuals (RSS).

However, residuals are now defined as

$$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki}.$$

To obtain estimates, take partial derivatives of RSS with respect to each $\hat{\beta}$, set each derivative equal to zero, and solve the system of $k+1$ equations.

- Without matrices, the algebra is difficult. For the remainder of this course, we will let R do the work for us.

# Coefficient Interpretation

**Model**

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

**Interpretation**

- The intercept $\hat{\beta}_0$ is the average value of $Y_i$ when all of the explanatory variables are equal to zero.
- Slope parameters $\hat{\beta}_1, \ldots, \hat{\beta}_k$ give us the change in $Y_i$ from a one-unit change in $X_j$, holding the other $X$ variables constant.

# Algebraic Properties of OLS

The OLS first-order conditions yield the same properties as before.

1. Residuals sum to zero: $\sum_{i=1}^{n} \hat{u}_i = 0$.
2. The sample covariance between the independent variables and the residuals is zero.
3. The point $(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k, \bar{Y})$ is always on the fitted regression "line."

# Goodness of Fit

Fitted values are defined similarly:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}.$$

The formula for $R^2$ is the same as before:

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}.$$

# Goodness of Fit

**Model 1:** $Y_i = \beta_0 + \beta_1 X_{1i} + u_i$.

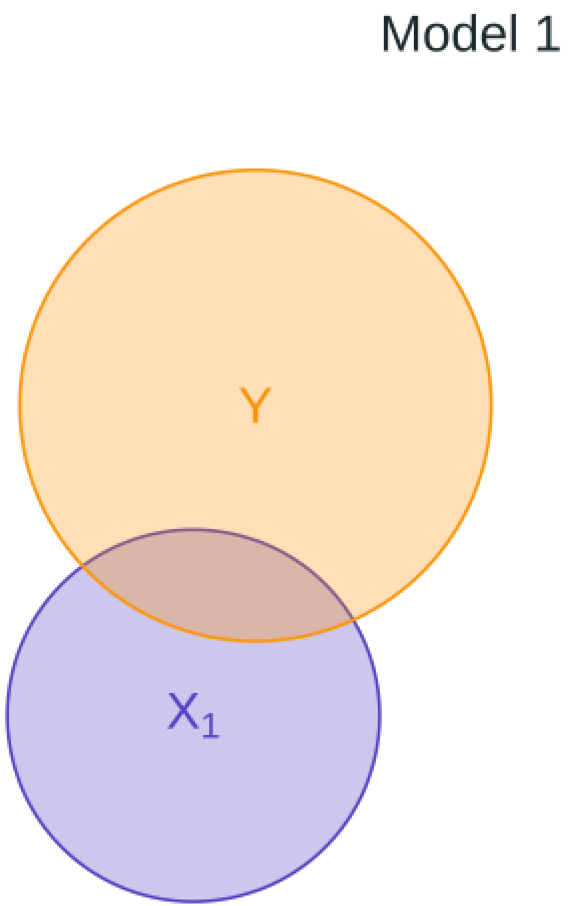
**Model 2:** $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + v_i$

True or false?

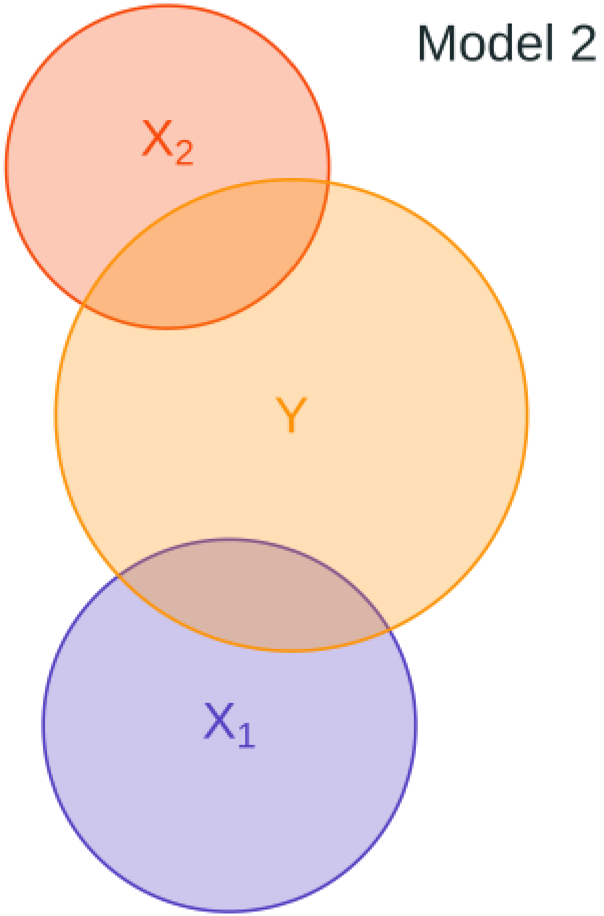
Model 2 will yield a lower $R^2$ than Model 1.

- Hint: Think of $R^2$ as $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$.

# Goodness of Fit



# Goodness of Fit



# Goodness of Fit

**Problem:** As we add variables to our model, $R^2$ *mechanically* increases.

**One solution:** Penalize for the number of variables, *e.g.*, adjusted $R^2$:

$$\bar{R}^2 = 1 - \frac{\sum_i \left(Y_i - \hat{Y}_i\right)^2 / (n - k - 1)}{\sum_i \left(Y_i - \bar{Y}\right)^2 / (n - 1)}$$

*Note:* Adjusted $R^2$ need not be between 0 and 1.

# Goodness of Fit

## Example: 2016 Election

```
lm(trump_margin ~ white, data = election) %>% glance()
```

```
#> # A tibble: 1 × 12
#>   r.squared adj.r.squared sigma statistic   p.value    df  logLik    AIC    BIC
#>       <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>  <dbl>  <dbl>
#> 1     0.320         0.320  25.4     1462. 1.51e-262     1 -14472. 28950. 28969.
#> # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
lm(trump_margin ~ white + poverty, data = election) %>% glance()
```

```
#> # A tibble: 1 × 12
#>   r.squared adj.r.squared sigma statistic   p.value    df  logLik    AIC    BIC
#>       <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>  <dbl>  <dbl>
#> 1     0.345         0.344  24.9      818. 4.20e-286     2 -14414. 28836. 28860.
#> # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

# OLS Assumptions

Same as before, except for assumption 2:

1. **Linearity:** The population relationship is linear in parameters with an additive error term.
2. No perfect collinearity: No $X$ variable is a perfect linear combination of the others.
3. **Random Sampling:** We have a random sample from the population of interest.
4. **Exogeneity:** The $X$ variable is exogenous (*i.e.*, $\mathbb{E}(u|X) = 0$).
5. **Homoskedasticity:** The error term has the same variance for each value of the independent variable (*i.e.*, $\text{Var}(u|X) = \sigma^2$).
6. **Normality:** The population error term is normally distributed with mean zero and variance $\sigma^2$ (*i.e.*, $u \sim N(0, \sigma^2)$)

# Perfect Collinearity

## Example: 2016 Election

OLS cannot estimate parameters for white and nonwhite simultaneously.

- white = 100 - nonwhite.

```
lm(trump_margin ~ white + nonwhite, data = election) %>% tidy()
```

```
#> # A tibble: 3 × 5
#>   term        estimate std.error statistic    p.value
#>   <chr>          <dbl>     <dbl>     <dbl>      <dbl>
#> 1 (Intercept)  -40.7      1.95       -20.9  6.82e- 91
#> 2 white          0.910    0.0238      38.2  1.51e-262
#> 3 nonwhite      NA       NA           NA   NA
```

R drops perfectly collinear variables for you.

# Multiple Linear Regression

## Tradeoffs

There are tradeoffs to remember as we add/remove variables:

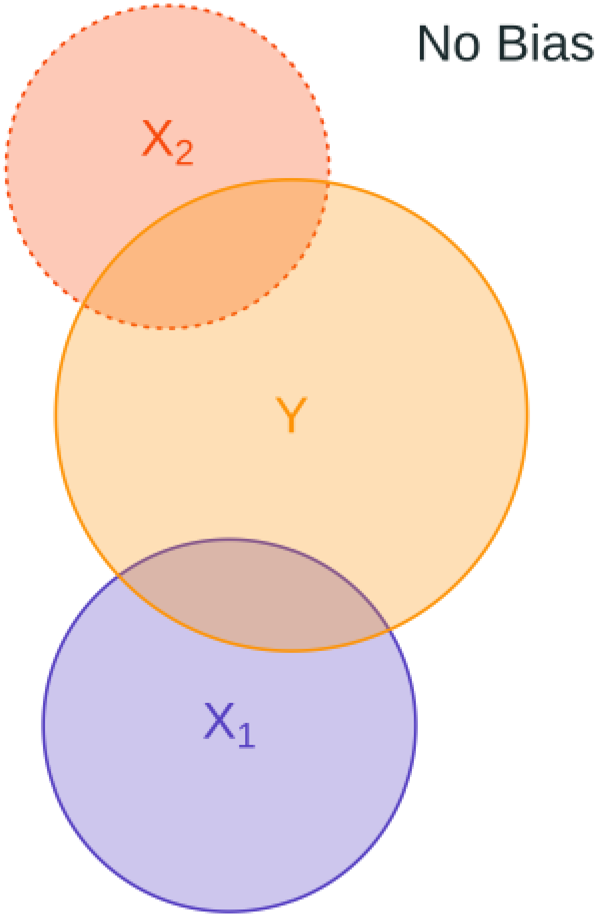
**Fewer variables**

- Generally explain less variation in $y$.
- Provide simple interpretations and visualizations (*parsimonious*).
- May need to worry about omitted-variable bias.

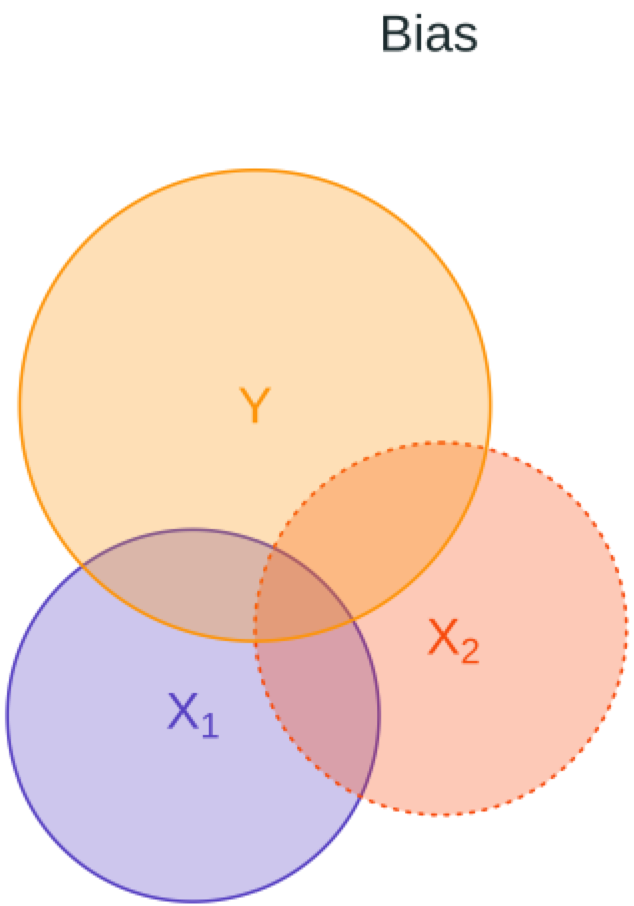
**More variables**

- More likely to find *spurious* relationships (statistically significant due to chance; do not reflect true, population-level relationships).
- More difficult to interpret the model.
- May still leave out important variables.

# Omitted Variables



# Omitted Variables



# Omitted Variables

| Math Score | | |
|---|---|---|
| **Explanatory variable** | **1** | **2** |
| *Intercept* | -84.84 | **-6.34** |
| | (18.57) | **(15.00)** |
| *log(Spend)* | -1.52 | **11.34** |
| | (2.18) | **(1.77)** |
| *Lunch* | | **-0.47** |
| | | **(0.01)** |

Data from 1823 elementary schools in Michigan

- *Math Score* is average fourth grade state math test scores.
- *log(Spend)* is the natural logarithm of spending per pupil.
- *Lunch* is the percentage of student eligible for free or reduced-price lunch.

# Omitted-Variable Bias

**Model 1:** $Y_i = \beta_0 + \beta_1 X_{1i} + u_i$.

**Model 2:** $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + v_i$

Estimating Model 1 (without $X_2$) yields omitted-variable bias:

$$\text{Bias} = \beta_2 \frac{\text{Cov}(X_{1i}, X_{2i})}{\text{Var}(X_{1i})}.$$

The sign of the bias depends on

1. The correlation between $X_2$ and $Y$, *i.e.*, $\beta_2$.
2. The correlation between $X_1$ and $X_2$, *i.e.*, $\text{Cov}(X_{1i}, X_{2i})$.