Multiple Linear Regression: Estimation

EC 320: Introduction to Econometrics

Winter 2022

Prologue

Other Things Being Equal

Goal: Isolate the effect of one variable on another.

- All else equal, how does increasing X affect Y.

Challenge: Changes in X often coincide with changes in other variables.

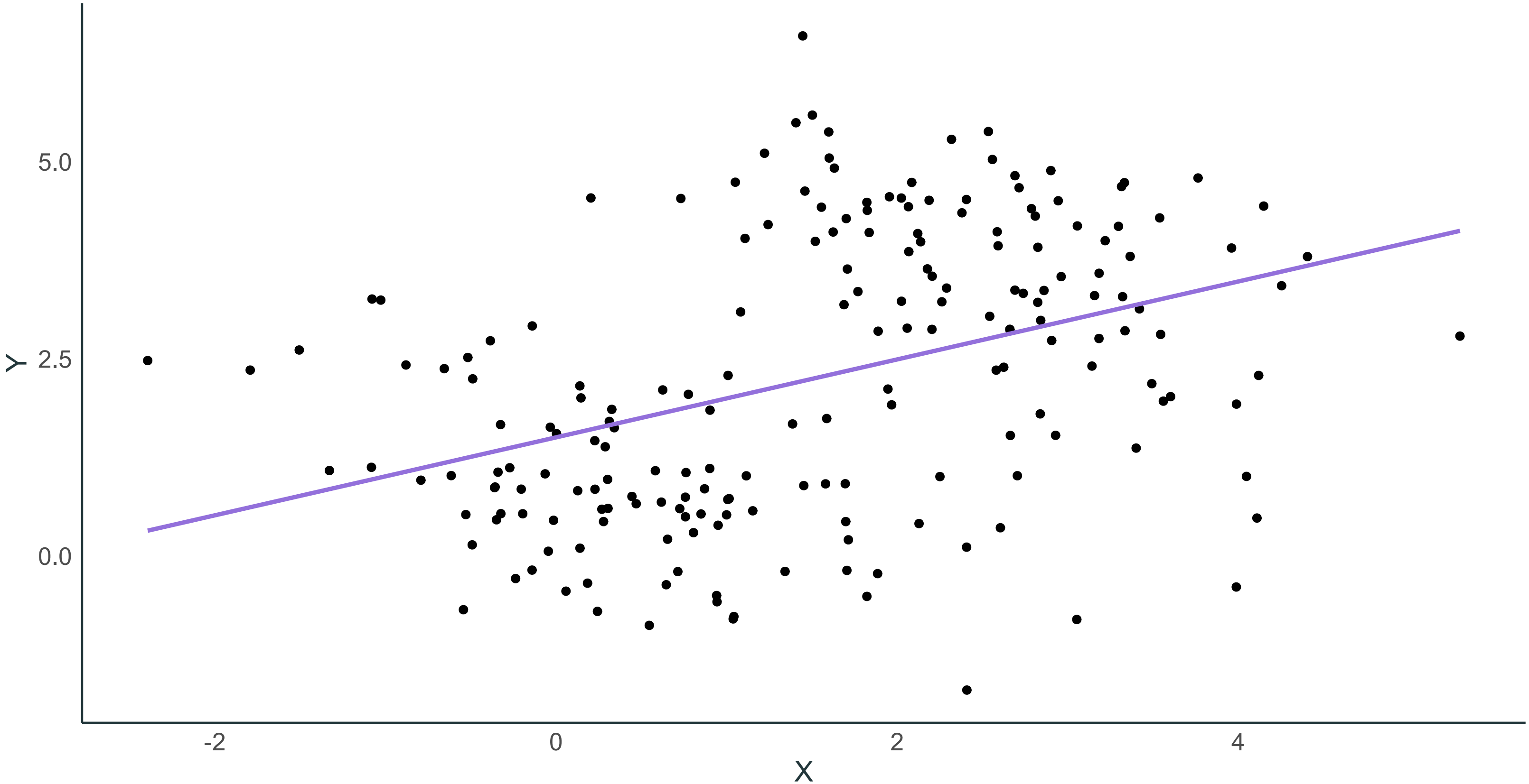
- Failure to account for other changes can *bias* OLS estimates of the effect of X on Y.

A potential solution: Account for other variables using **multiple linear regression**.

- Easier to defend the exogeneity assumption.

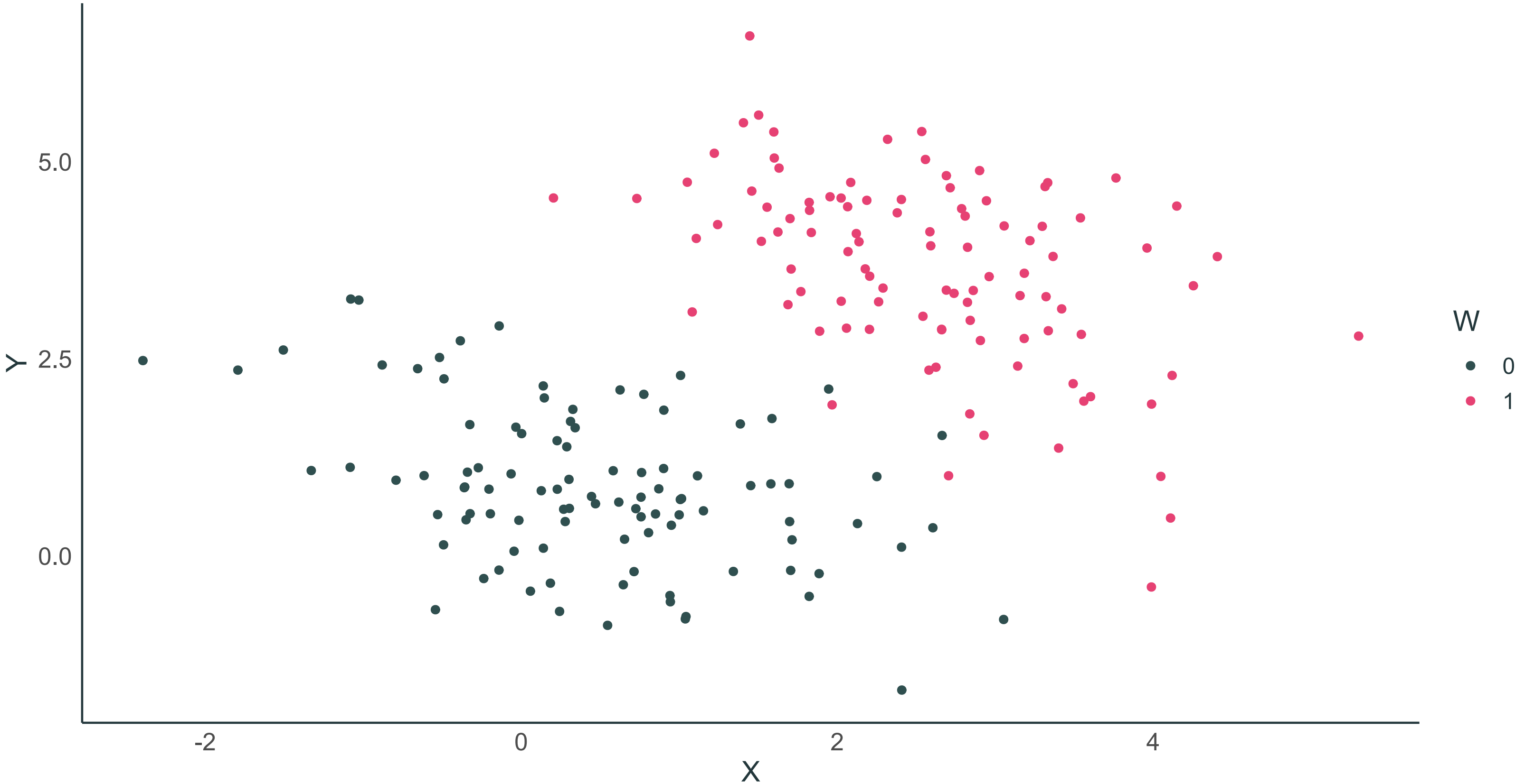
Other Things Equal?

OLS picks $\hat{\beta}_0$ and $\hat{\beta}_1$ that trace out the line of best fit. Ideally, we wound like to interpret the slope of the line as the causal effect of X on Y.



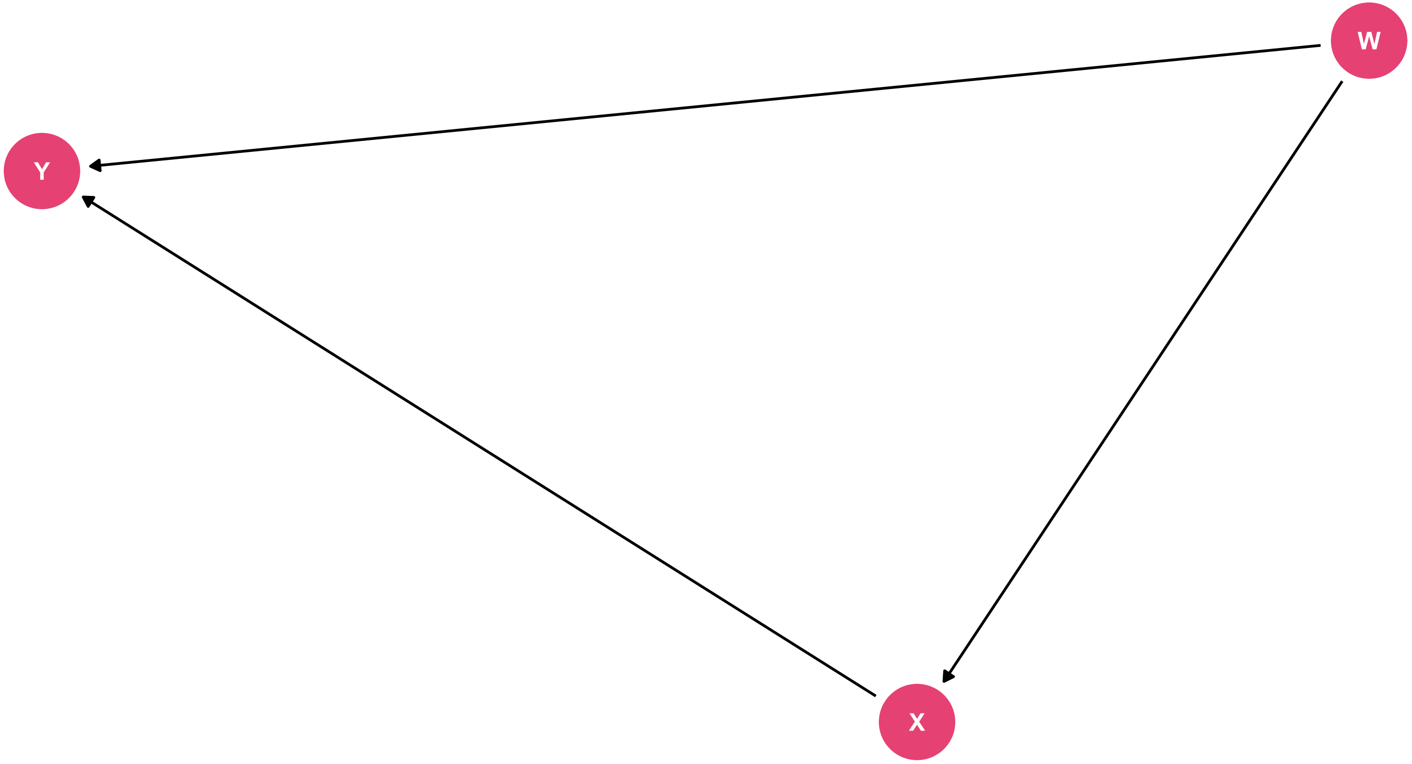
Confounders

However, the data are grouped by a third variable W. How would omitting W from the regression model affect the slope estimator?

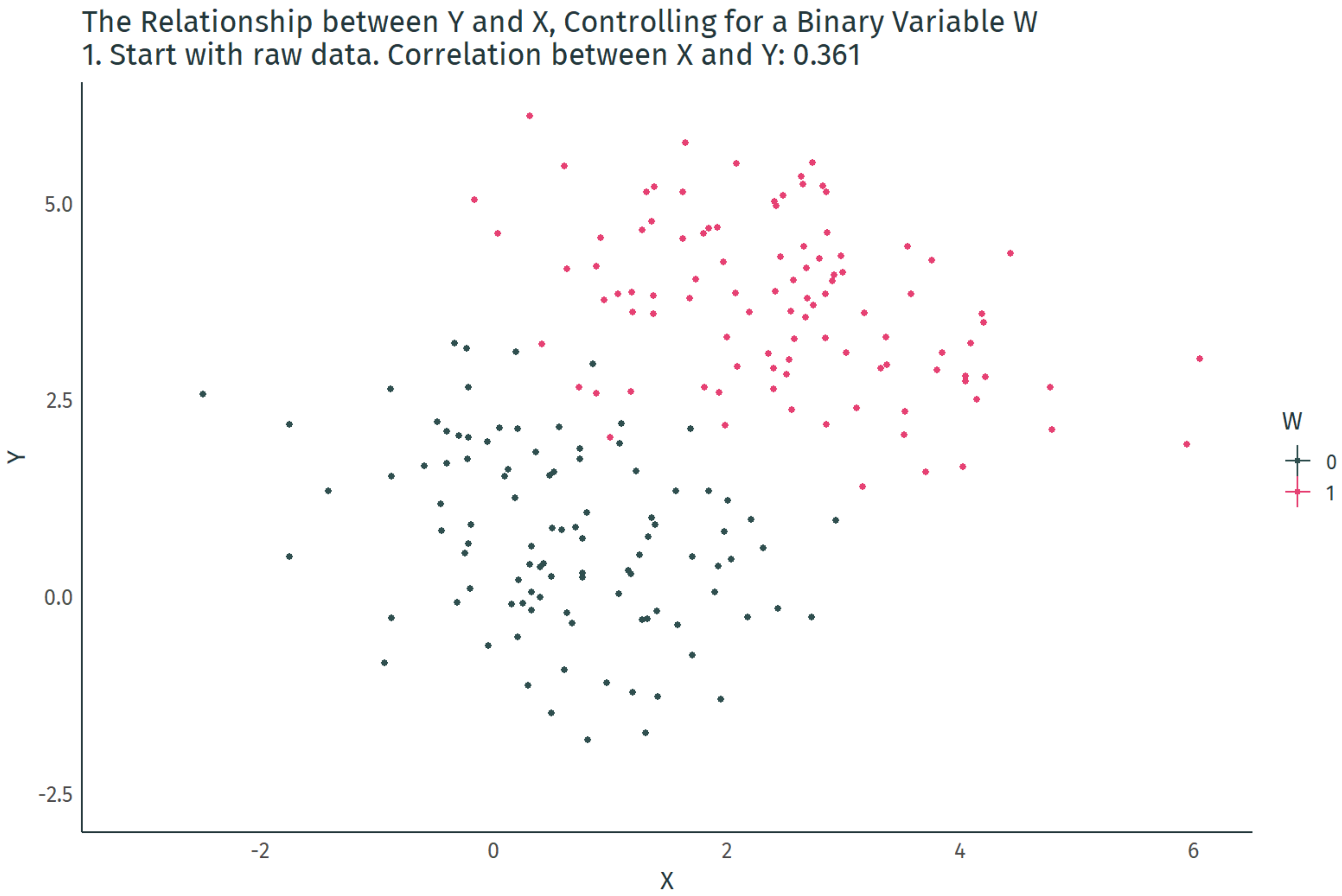


Confounders

The problem with W is that it affects both Y and X. Without adjusting for W, we cannot isolate the causal effect of X on Y.



Controlling for Confounders



Controlling for Confounders

```
lm(Y ~ X, data = df) %>% tidy()
```

```
#> # A tibble: 2 × 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)    1.51     0.169       8.91 3.36e-16
#> 2 X              0.494    0.0811      6.10 5.53e- 9
```

```
lm(Y ~ X + W, data = df) %>% tidy()
```

```
#> # A tibble: 3 × 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)    1.11     0.104      10.6  3.57e-21
#> 2 X             -0.518    0.0731     -7.09 2.32e-11
#> 3 W              3.88     0.208      18.6  2.32e-45
```

Multiple Linear Regression

Multiple Linear Regression

More explanatory variables

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

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

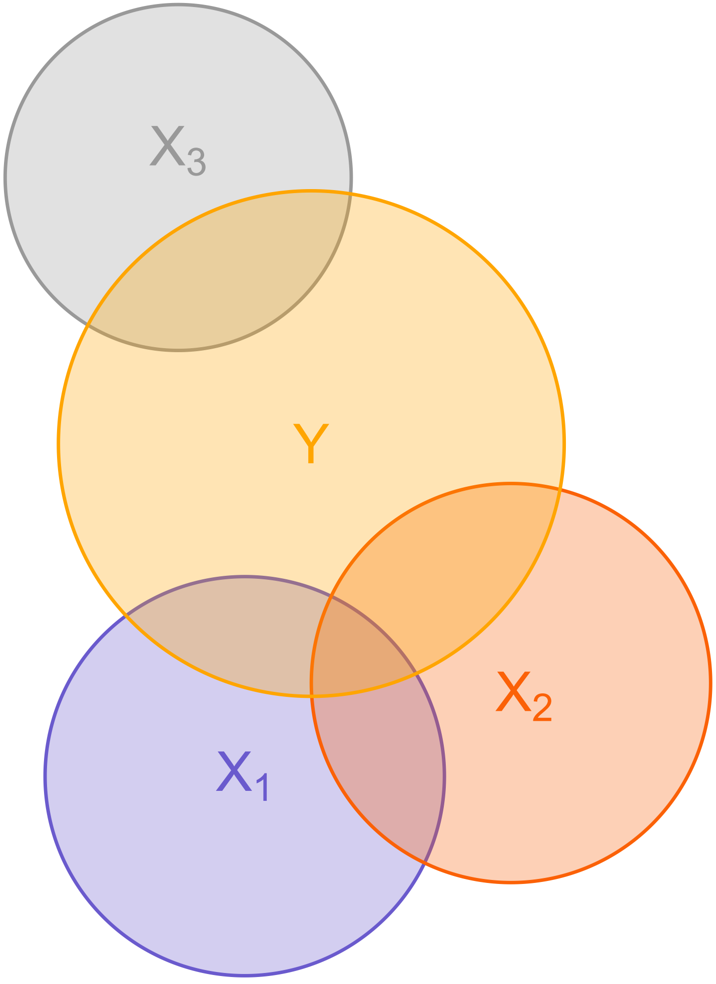
Multiple linear regression features one outcome variable and multiple explanatory variables:

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

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

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

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

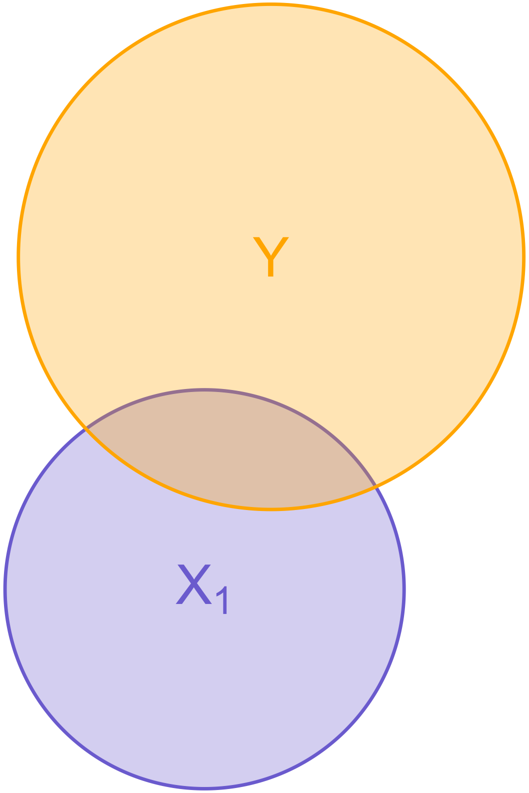
True or false?

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

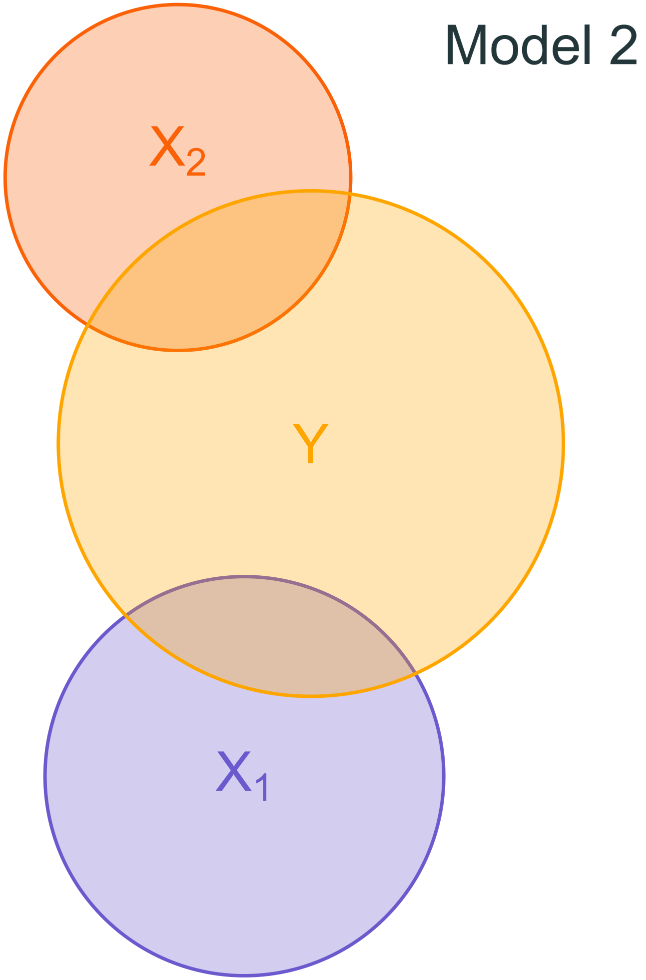
- Hint: Think of R^2 as $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$.

Goodness of Fit

Model 1



Goodness of Fit



Goodness of Fit

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

To see this problem, we can simulate a dataset of 10,000 observations on y and 1,000 random x_k variables. **No relations between y and the x_k!**

Pseudo-code outline of the simulation:

- Generate 10,000 observations on y
- Generate 10,000 observations on variables x_1 through x_{1000}
- Regressions
  - LM_1: Regress y on x_1; record R^2
  - LM_2: Regress y on x_1 and x_2; record R^2
  - ...
  - LM_{1000}: Regress y on x_1, x_2, ..., x_{1000}; record R^2

Goodness of Fit

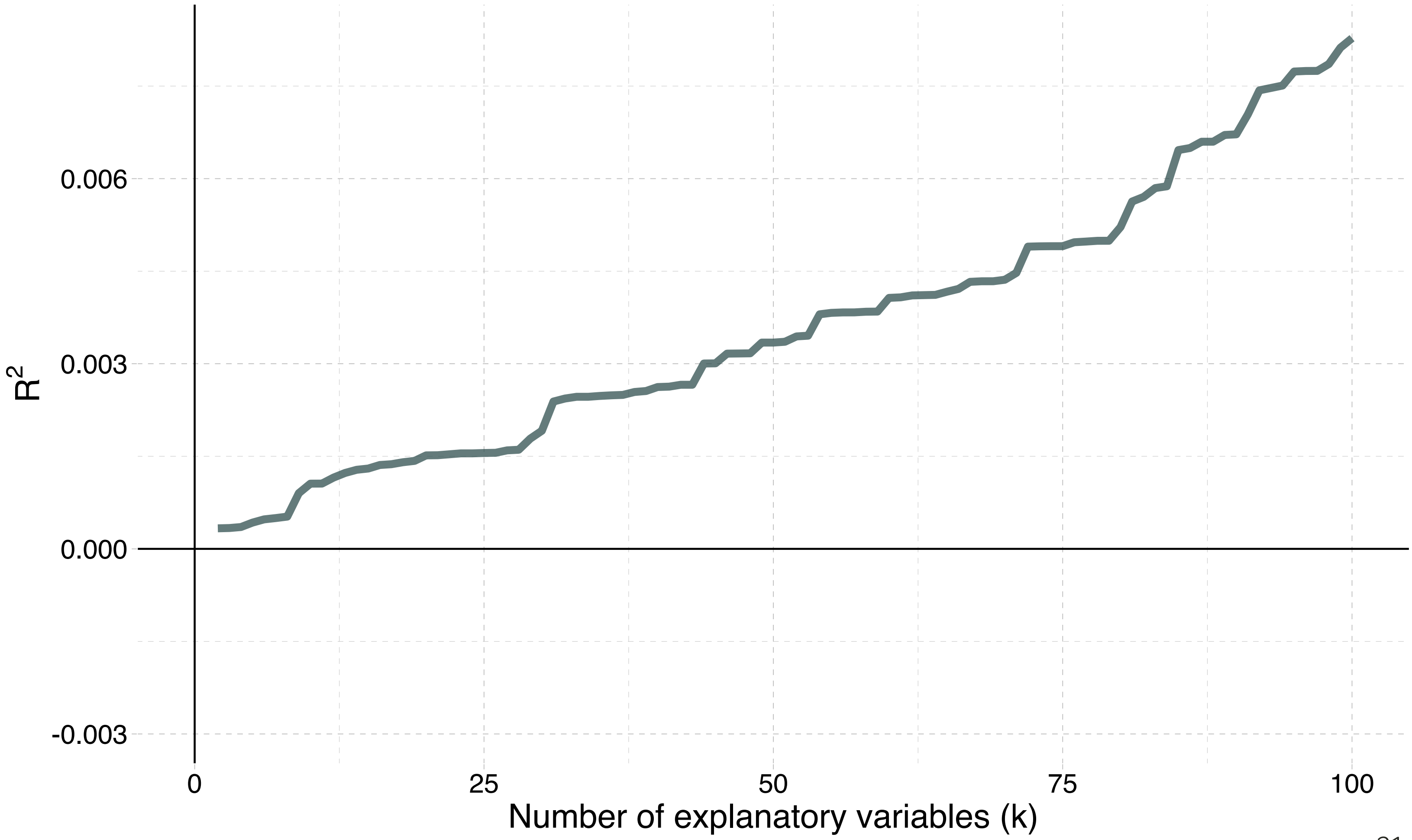
Problem: As we add variables to our model, R^2 *mechanically* increases.

R code for the simulation:

```
set.seed(1234)
#plan(multiprocess)
y <- rnorm(1e4) # 10000 obs
x <- matrix(data = rnorm(1e6), nrow = 1e4) # 10000 by 100 matrix
x %<>% cbind(matrix(data = 1, nrow = 1e4, ncol = 1) # 10000 by 1 vector
             , x)
r_fun <- function(i) {
  tmp_reg <- lm(y ~ x[,1:(i + 1)]) %>% summary()
  data.frame(
  k = i + 1,
  r2 = tmp_reg$r.squared,
  r2_adj = tmp_reg$adj.r.squared)
}
r_df <- future_map(1:(1e2-1), r_fun) %>% bind_rows()
r_df
```

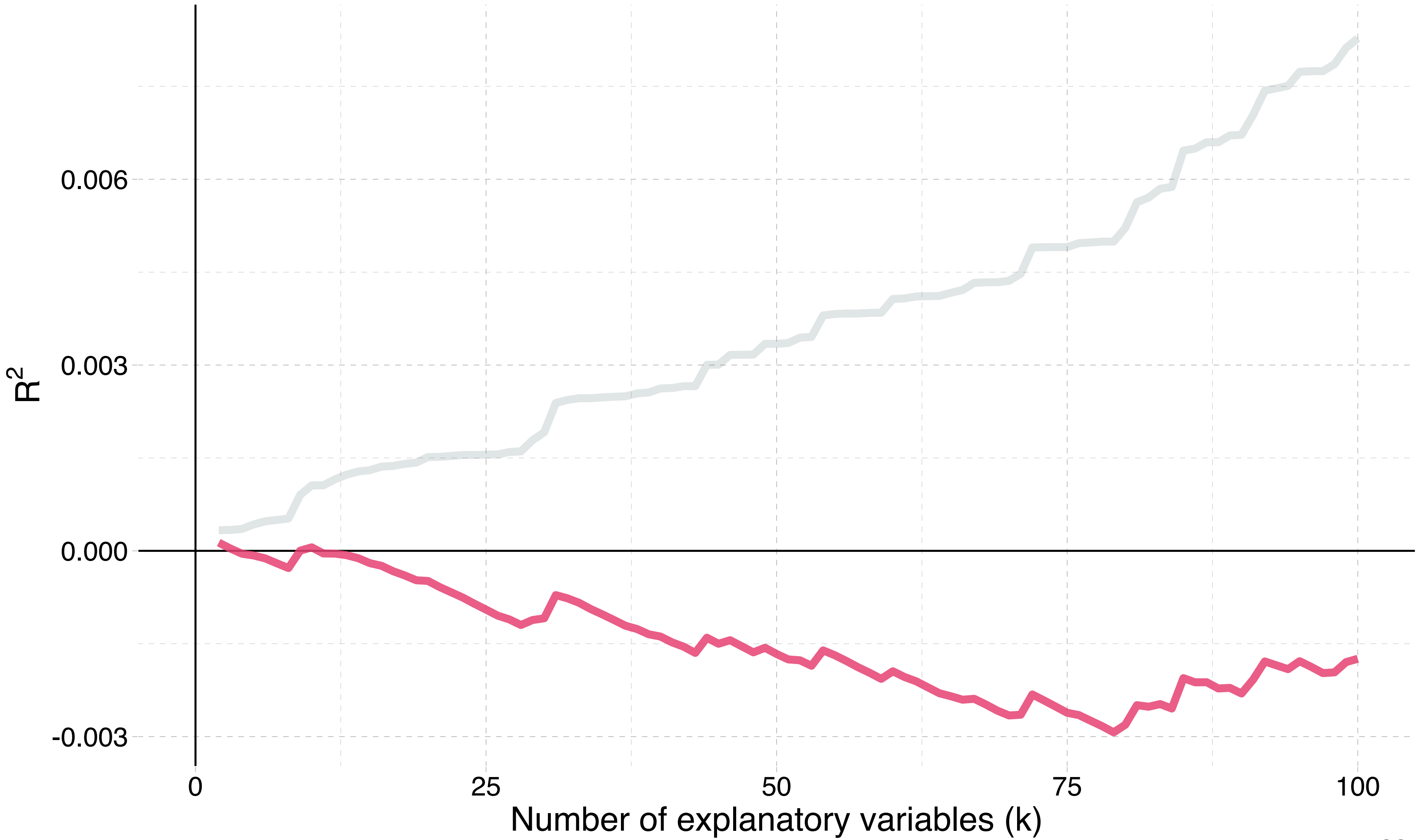
Goodness of Fit

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



Goodness of Fit

One solution: Adjusted R^2



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. **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.*, $\mathrm{Var}(u|X) = \sigma^2$).
5. **Non-autocorrelation:** The values of error terms are independent from one another (*i.e.*, $E[u_i u_j] = 0, \forall i \text{ s.t. } i \neq j$)
6. **Normality:** The population error term is normally distributed with mean zero and variance σ^2 (*i.e.*, $u \sim N(0, \sigma^2)$)

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.

Tradeoffs

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

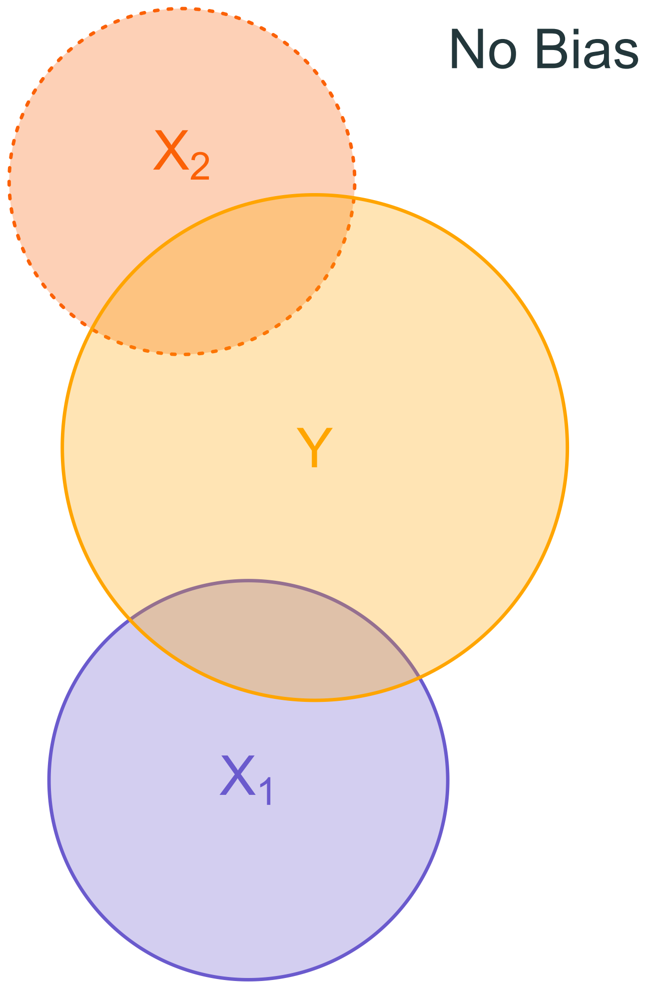
Fewer variables

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

More variables

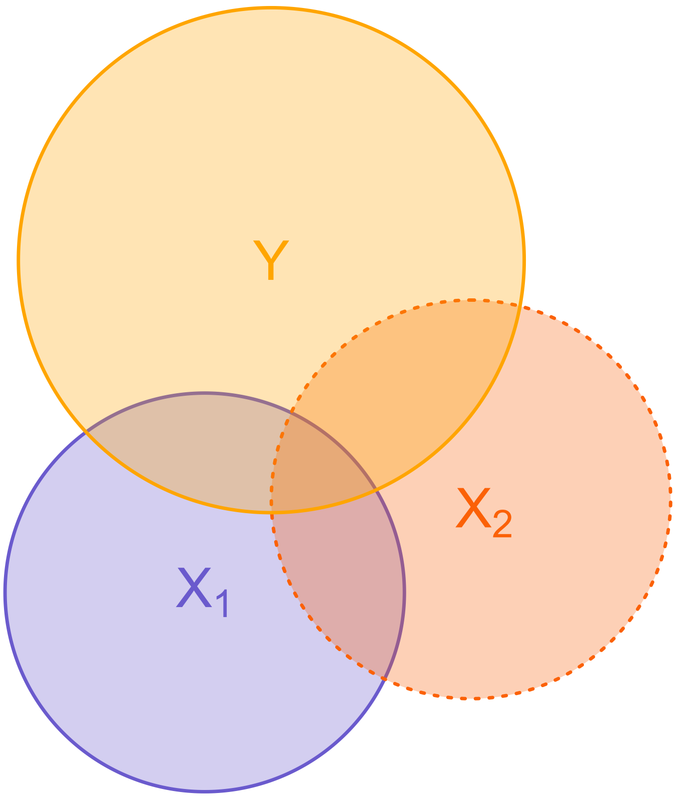
- More likely to find *spurious* relationships (statistically significant due to chance; do not reflect true, population-level relationships).
- More difficult to interpret the model.
- May still leave out important variables.

Omitted Variables



Omitted Variables

Bias



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.*, β_2.
2. The correlation between X_1 and X_2, *i.e.*, $\text{Cov}(X_{1i}, X_{2i})$.