Gov 2000: 9. Regression with Two Independent Variables

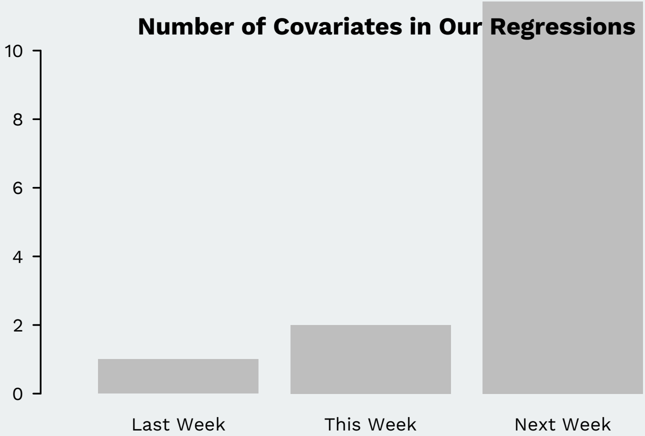
Matthew Blackwell

Fall 2016

Where are we? Where are we going?

Where are we? Where are we going?

Where are we? Where are we going?



1/ Why Add Variables to a Regression?

Berkeley gender bias

- Graduate admissions data from Berkeley, 1973
- Acceptance rates:
  - Men: 8442 applicants, 44% admission rate
  - Women: 4321 applicants, 35% admission rate
- Evidence of discrimination toward women in admissions?
- This is a marginal relationship.
- What about the conditional relationship within departments?

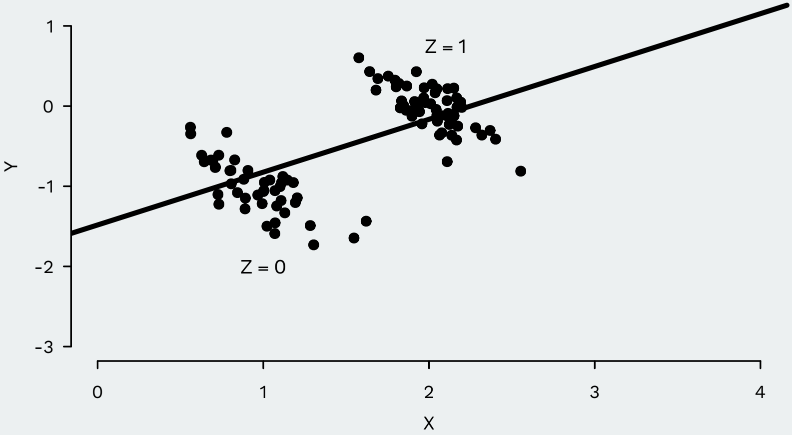
Berkeley gender bias, II

- Within departments:

	Men		Women	
Dept	Applied	Admitted	Applied	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

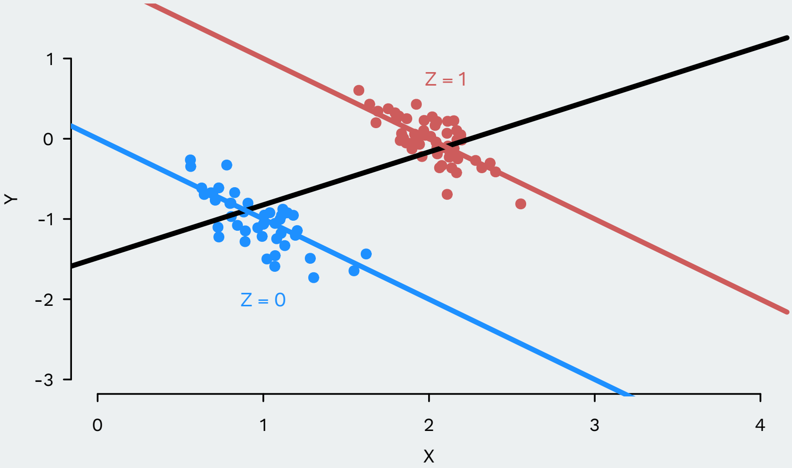
- Within departments, women do somewhat better than men!
- Women apply to more challenging departments.
- Marginal relationships (admissions and gender) $\neq$ conditional relationship given third variable (department).

Simpson's paradox



- Overall a positive relationship between Y_i and X_i.

Simpson's paradox



- Overall a positive relationship between Y_i and X_i.
- But within levels of Z_i, the opposite.

Basic idea

- Old goal: estimate the mean of Y as a function of some independent variable, X: $\mathbb{E}[Y_i|X_i]$.
- For continuous X's, we modeled the CEF/regression function with a line:

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

- New goal: estimate the relationship of two variables, Y_i and X_i, conditional on a third variable, Z_i:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

- β's are the population parameters we want to estimate.

Why control for another variable

- Descriptive
  - Get a sense for the relationships in the data.
  - Conditional on the number of steps I've taken, does higher activity levels correlate with less weight?
- Predictive
  - We can usually make better predictions about the dependent variable with more information on independent variables.
- Causal
  - Block potential confounding, which is when X doesn't cause Y, but only appears to because a third variable Z causally affects both of them.

Plan of attack

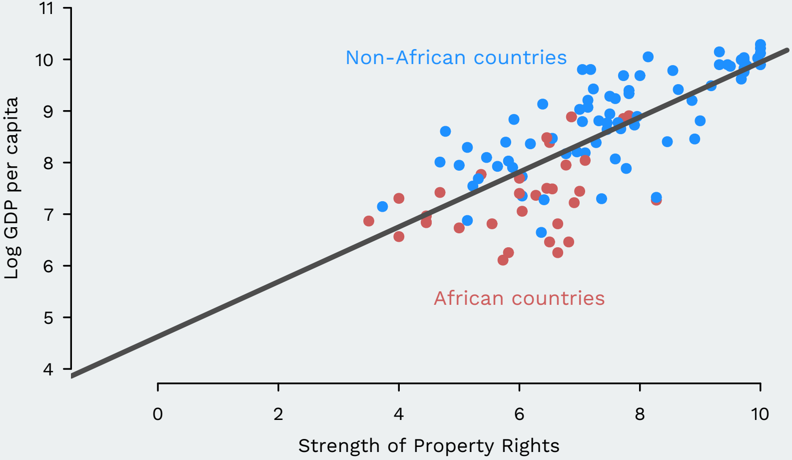
1. Interpretation with a binary Z_i
2. Interpretation with a continuous Z_i
3. Mechanics of OLS with 2 covariates
4. OLS assumptions with 2 covariates:
    - Omitted variable bias
    - Multicollinearity

What we won't cover in lecture

1. The OLS formulas for 2 covariates
2. Proofs
3. The second covariate being a function of the first: $Z_i = X_i^2$
4. Hypothesis test/confidence intervals (almost exactly the same)

2/ Adding a Binary Covariate

Example



Non-African countries
African countries
Log GDP per capita
11
10
9
8
7
6
5
4
0
2
4
6
8
10
Strength of Property Rights

Basics

- Ye olde model:

$$\mathbb{E}[Y_i|X_i] = \alpha_0 + \alpha_1 X_i$$

  - (α_0, α_1) are the bivariate intercept/slope, e_i is the bivariate error.
- Concern: AJR might be picking up an "African effect":
  - African countries might have low incomes and weak property rights.
- Condition on country being in Africa or not to remove this:

$$\mathbb{E}[Y_i|X_i, Z_i] = \beta_0 + \beta_1 X_i + \beta_2 Z_i$$

  - $Z_i = 1$ to indicate that i is an African country
  - $Z_i = 0$ to indicate that i is an non-African country
  - Effects are now within Africa or within non-Africa, not between

AJR model

```r
ajr.mod <- lm(logpgp95 ~ avexpr + africa, data = ajr)
summary(ajr.mod)
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.6556     0.3134   18.04   <2e-16 ***
## avexpr        0.4242     0.0397   10.68   <2e-16 ***
## africa       -0.8784     0.1471   -5.97    3e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.625 on 108 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.708,  Adjusted R-squared:  0.702
## F-statistic:  131 on 2 and 108 DF,  p-value: <2e-16
```

Two lines, one regression

- How can we interpret this model?
- Plug in two possible values for Z_i and rearrange
- When $Z_i = 0$:

$$\begin{aligned}\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 \times 0 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i\end{aligned}$$

- When $Z_i = 1$:

$$\begin{aligned}\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 \times 1 \\ &= (\widehat{\beta}_0 + \widehat{\beta}_2) + \widehat{\beta}_1 X_i\end{aligned}$$

- Two different intercepts, same slope

Interpretation of the coefficients

	Intercept for X_i	Slope for X_i
Non-African country ($Z_i = 0$)	$\widehat{\beta}_0$	$\widehat{\beta}_1$
African country ($Z_i = 1$)	$\widehat{\beta}_0 + \widehat{\beta}_2$	$\widehat{\beta}_1$

- In this example, we have:

$$\widehat{Y}_i = 5.656 + 0.424 \times X_i - 0.878 \times Z_i$$

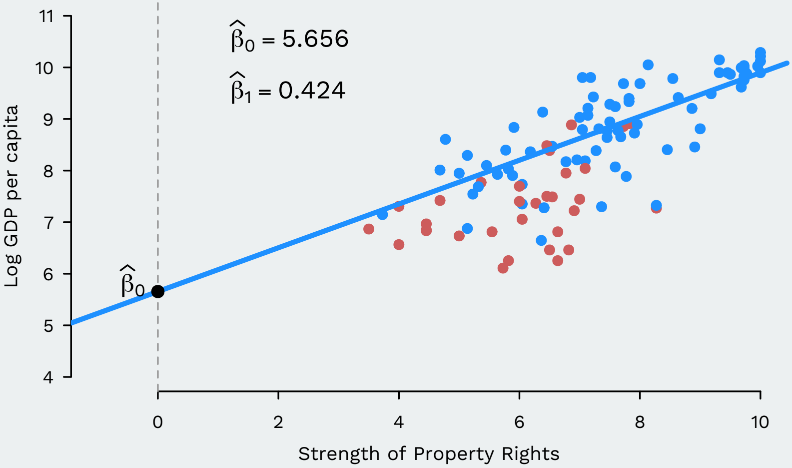
- $\widehat{\beta}_0$: average log income for non-African country ($Z_i = 0$) with property rights measured at 0 is 5.656
- $\widehat{\beta}_1$: A one-unit increase in property rights is associated with a 0.424 increase in average log incomes for two African countries (or for two non-African countries)
- $\widehat{\beta}_2$: there is a -0.878 average difference in log income per capita between African and non-African counties conditional on property rights

General interpretation of the coefficients

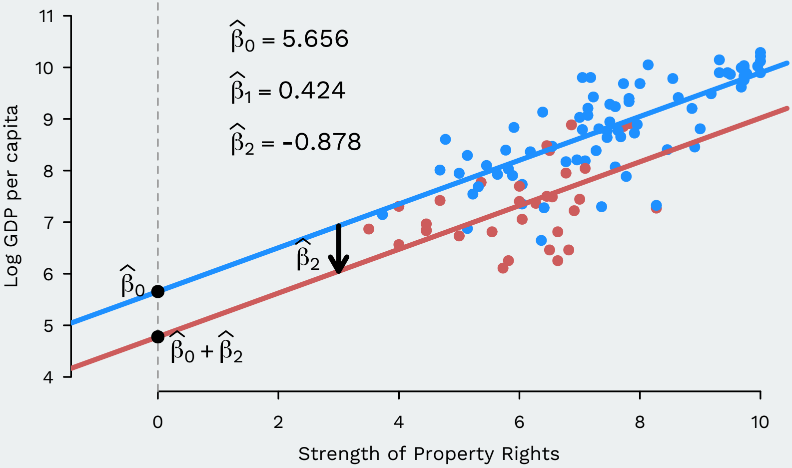
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i$$

- $\widehat{\beta}_0$: average value of Y_i when both X_i and Z_i are equal to 0
- $\widehat{\beta}_1$: A 1-unit increase in X_i is associated with a $\widehat{\beta}_1$-unit change in Y_i for units with the same value of Z_i
- $\widehat{\beta}_2$: average difference in Y_i between $Z_i = 1$ group and $Z_i = 0$ group for units with the same value of X_i

Adding a binary variable, visually



Adding a binary variable, visually



Marginal vs conditional

3/ Adding a Continuous Covariate

Adding a continuous variable

- Ye olde model:

$$\mathbb{E}[Y_i|X_i] = \alpha_0 + \alpha_1 X_i$$

- New concern: geography is confounding the effect
  - geography might affect political institutions
  - geography might affect average incomes (through diseases like malaria)
- Condition on Z_i: mean temperature in country i (continuous)

$$\mathbb{E}[Y_i|X_i, Z_i] = \beta_0 + \beta_1 X_i + \beta_2 Z_i$$

AJR model, revisited

```
ajr.mod2 <- lm(logpgp95 ~ avexpr + meantemp, data = ajr)
summary(ajr.mod2)
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.8063     0.7518    9.05  1.3e-12 ***
## avexpr        0.4057     0.0640    6.34  3.9e-08 ***
## meantemp     -0.0602     0.0194   -3.11    0.003 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.643 on 57 degrees of freedom
##   (103 observations deleted due to missingness)
## Multiple R-squared:  0.615,  Adjusted R-squared:  0.602
## F-statistic: 45.6 on 2 and 57 DF,  p-value: 1.48e-12
```

Interpretation with a continuous Z

	Intercept for X_i	Slope for X_i
$Z_i = 0\,^\circ\mathrm{C}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
$Z_i = 21\,^\circ\mathrm{C}$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 21$	$\widehat{\beta}_1$
$Z_i = 24\,^\circ\mathrm{C}$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 24$	$\widehat{\beta}_1$
$Z_i = 26\,^\circ\mathrm{C}$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 26$	$\widehat{\beta}_1$

- In this example we have:

$$\widehat{Y}_i = 6.806 + 0.406 \times X_i - 0.06 \times Z_i$$

- $\widehat{\beta}_0$: average log income for a country with property rights measured at 0 and a mean temperature of 0 is 6.806
- $\widehat{\beta}_1$: A one-unit increase in property rights is associated with a 0.406 change in average log incomes conditional on a country's mean temperature
- $\widehat{\beta}_2$: A one-degree increase in mean temperature is associated with a -0.06 change in average log incomes conditional on strength of property rights

General interpretation

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i$$

- The coefficient $\widehat{\beta}_1$ measures how the predicted outcome varies in X_i for units with the same value of Z_i.
- The coefficient $\widehat{\beta}_2$ measures how the predicted outcome varies in Z_i for units with the same value of X_i.

4/ OLS Mechanics with Two Covariates

Fitted values and residuals

- Where do we get our hats? $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2$
- Fitted values for $i = 1, \ldots, n$:

$$\widehat{Y}_i = \widehat{\mathbb{E}}[Y_i | X_i, Z_i] = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i$$

- Residuals for $i = 1, \ldots, n$:

$$\hat{u}_i = Y_i - \widehat{Y}_i$$

- Minimize the sum of the squared residuals, just like before:

$$(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2) = \underset{b_0, b_1, b_2}{\arg\min} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i - b_2 Z_i)^2$$

- We'll derive closed-form estimators with arbitrary covariates next week.

OLS estimator recipe using two steps

- No explicit OLS formulas this week, but a recipe instead
- "Partialling out" OLS recipe:
    1. Run regression of X_i on Z_i:
    $$\widehat{X}_i = \widehat{\mathbb{E}}[X_i|Z_i] = \widehat{\delta}_0 + \widehat{\delta}_1 Z_i$$
    2. Calculate residuals from this regression:
    $$\hat{r}_{xz,i} = X_i - \widehat{X}_i$$
    3. Run a simple regression of Y_i on residuals, $\hat{r}_{xz,i}$:
    $$\widehat{Y}_i = \widehat{\alpha}_0 + \widehat{\alpha}_1 \hat{r}_{xz,i}$$
- Estimate of $\widehat{\alpha}_1$ will be equivalent to $\widehat{\beta}_1$ from the "big" regression:
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i$$

First regression

- Regress X_i on Z_i:

```r
## when missing data exists, need the na.actionin order
## to place residuals or fitted values back into the data
ajr.first <- lm(avexpr ~ meantemp, data = ajr,
                na.action = na.exclude)
summary(ajr.first)
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.9568     0.8202    12.1  < 2e-16 ***
## meantemp     -0.1490     0.0347    -4.3 0.000067 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.32 on 58 degrees of freedom
##   (103 observations deleted due to missingness)
## Multiple R-squared:  0.241,  Adjusted R-squared:  0.228
## F-statistic: 18.4 on 1 and 58 DF,  p-value: 0.0000673
```

Regression of log income on the residuals

- Save residuals:

```
## store the residuals
ajr$avexpr.res <- residuals(ajr.first)
```

- Now we compare the estimated slopes:

```
coef(lm(logpgp95 ~ avexpr.res, data = ajr))
```

```
## (Intercept)  avexpr.res
##      8.0543      0.4057
```

```
coef(lm(logpgp95 ~ avexpr + meantemp, data = ajr))
```

```
## (Intercept)      avexpr    meantemp
##     6.80627     0.40568    -0.06025
```

Residual/partial regression plot

- Can plot the conditional relationship between property rights and income given temperature:

5/ OLS Assumptions with Two Covariates

OLS assumptions for unbiasedness

- Simple regression assumptions unbiasedness/consistency of OLS:
    1. Linearity
    2. Random/iid sample
    3. Variation in X_i
    4. Zero conditional mean error: $\mathbb{E}[u_i|X_i] = 0$
- Small modification to these with 2 covariates:
    1. Linearity
    $$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$
    2. Random/iid sample
    3. No perfect collinearity
    4. Zero conditional mean error (both X_i and Z_i unrelated to u_i)
    $$\mathbb{E}[u_i|X_i, Z_i] = 0$$

New assumption

Assumption 3: No perfect collinearity

(1) No independent variable is constant in the sample and (2) there are no exactly linear relationships among the independent variables.

- Two components
  1. Both X_i and Z_i have to vary.
  2. Z_i cannot be a deterministic, linear function of X_i.
- Part 2 rules out anything of the form:

$$Z_i = a + bX_i$$

- What's the correlation between Z_i and X_i? 1!

Perfect collinearity example

- Simple example:
  - $X_i = 1$ if a country is **not** in Africa and 0 otherwise.
  - $Z_i = 1$ if a country **is** in Africa and 0 otherwise.
- But, clearly we have the following:

$$Z_i = 1 - X_i$$

- These two variables are perfectly collinear.
- What about the following:
  - X_i = property rights
  - $Z_i = X_i^2$
- Do we have to worry about collinearity here?
- No! Because while Z_i is a deterministic function of X_i, it is a nonlinear function of X_i.

R and perfect collinearity

- R, Stata, et al will drop one of the variables if there is perfect collinearity:

```r
ajr$nonafrica <- 1 - ajr$africa
summary(lm(logpgp95 ~ africa + nonafrica, data = ajr))
```

```
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.7164     0.0899   96.94  < 2e-16 ***
## africa       -1.3612     0.1631   -8.35  4.9e-14 ***
## nonafrica         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.913 on 146 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.323,  Adjusted R-squared:  0.318
## F-statistic: 69.7 on 1 and 146 DF,  p-value: 4.87e-14
```

6/ Omitted Variable Bias

Unbiasedness revisited

- Long regression:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$

- Assumptions 1-4 ⇒ OLS is unbiased for $\beta_0, \beta_1, \beta_2$
- What happens if we ignore the Z_i and just run the simple linear regression with just X_i?
- Short regression:

$$Y_i = \alpha_0 + \alpha_1 X_i + u_i^*$$

- OLS estimates from the short regression: $(\widehat{\alpha}_0, \widehat{\alpha}_1)$
- Question: will $\mathbb{E}[\widehat{\alpha}_1] = \beta_1$? If not, what will be the difference?

Deriving the short regression

- How can we relate α_1 to β_1?
  - Short regression will be unbiased for CEF of Y_i just given X_i.
- Write "short CEF" in terms of the "long" regression model:

$$\begin{aligned} \mathbb{E}[Y_i|X_i] &= \mathbb{E}[\beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i|X_i] \\ &= \beta_0 + \beta_1 X_i + \beta_2 \mathbb{E}[Z_i|X_i] + \mathbb{E}[u_i|X_i] \end{aligned}$$

- By assumption 4, X_i is unrelated to the long-regression error, so $\mathbb{E}[u_i|X_i] = 0$.

$$\mathbb{E}[Y_i|X_i] = \beta_0 + \beta_1 X_i + \beta_2 \mathbb{E}[Z_i|X_i]$$

Deriving the short regression

$$\mathbb{E}[Y_i|X_i] = \beta_0 + \beta_1 X_i + \beta_2 \mathbb{E}[Z_i|X_i]$$

- Let $\mathbb{E}[Z_i|X_i] = \gamma_0 + \gamma_1 X_i$ be the (population) CEF from a regression of Z_i on X_i.
- Then, we can write the short CEF as:

$$\begin{aligned}\mathbb{E}[Y_i|X_i] &= \beta_0 + \beta_1 X_i + \beta_2(\gamma_0 + \gamma_1 X_i) \\ &= (\beta_0 + \gamma_0) + (\beta_1 + \beta_2\gamma_1)X_i \\ &= \alpha_0 + \alpha_1 X_i\end{aligned}$$

- Under these assumptions, short regression OLS unbiased for α_1:

$$\mathbb{E}[\widehat{\alpha}_1] = \alpha_1 = \beta_1 + \beta_2\gamma_1$$

Omitted variable bias

- Omitted variable bias: bias for long regression coefficient from omitting Z_i:

$$\text{Bias}(\widehat{\alpha}_1) = \mathbb{E}[\widehat{\alpha}_1] - \beta_1 = \beta_2\delta_1$$

- In other words omitted variable bias is:

$$(\text{“effect” of } Z_i \text{ on } Y_i) \times (\text{“effect” of } X_i \text{ on } Z_i)$$
$$(\text{omitted} \rightarrow \text{outcome}) \times (\text{included} \rightarrow \text{omitted})$$

Omitted variable bias, summary

- Remember that by OLS, the effect of X_i on Z_i is:

$$\delta_1 = \frac{\text{cov}(Z_i, X_i)}{\text{var}(X_i)}$$

- We can summarize the direction of bias like so:

	$\text{cov}(X_i, Z_i) > 0$	$\text{cov}(X_i, Z_i) < 0$	$\text{cov}(X_i, Z_i) = 0$
$\beta_2 > 0$	Positive bias	Negative Bias	No bias
$\beta_2 < 0$	Negative bias	Positive Bias	No bias
$\beta_2 = 0$	No bias	No bias	No bias

- Very relevant if Z_i is unobserved for some reason!

Including irrelevant variables

- What if we do the opposite and include an irrelevant variable?
- What would it mean for Z_i to be an irrelevant variable?

$$Y_i = \beta_0 + \beta_1 X_i + 0 \times Z_i + u_i$$

- So in this case, the true value of $\beta_2 = 0$. But under Assumptions 1-4, OLS is unbiased for all the parameters:

$$\mathbb{E}[\widehat{\beta}_0] = \beta_0$$
$$\mathbb{E}[\widehat{\beta}_1] = \beta_1$$
$$\mathbb{E}[\widehat{\beta}_2] = 0$$

- Including an irrelevant variable will increase the standard errors for $\widehat{\beta}_1$.

7/ Goodness of Fit & Multicollinearity

Prediction error

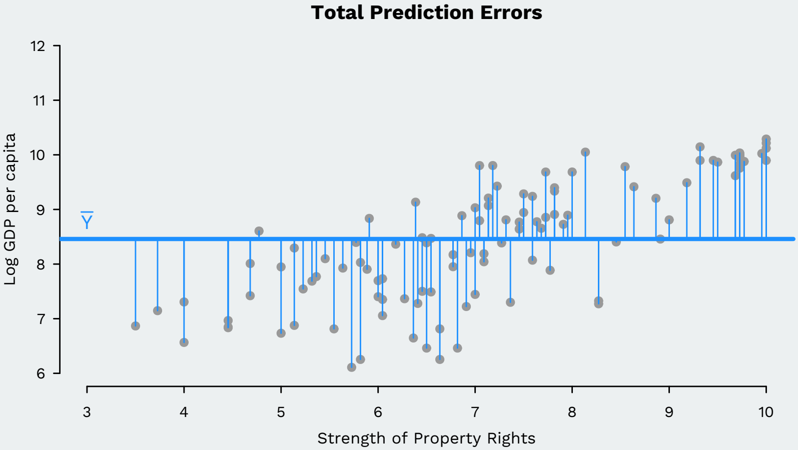
- How do we judge how well a regression fits the data?
- How much does X_i help us predict Y_i?
- Prediction errors without X_i:
  - Best prediction is the mean, $\overline{Y}$
  - Prediction error is called the total sum of squares (SS_{tot}) would be:

$$SS_{tot} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

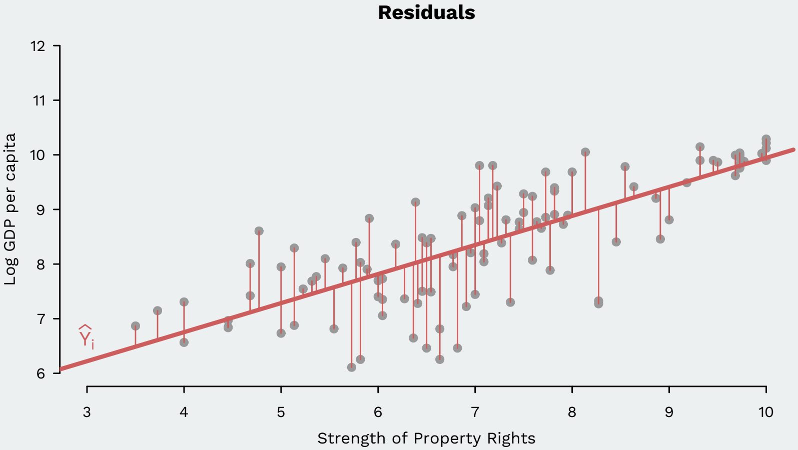
- Prediction errors with X_i:
  - Best predictions are the fitted values, $\widehat{Y}_i$.
  - Prediction error is the the sum of the squared residuals or SS_{res}:

$$SS_{res} = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$$

Total SS vs SSR



Total SS vs SSR



R-square

- Regression will always improve in-sample fit: $SS_{tot} > SS_{res}$
- How much better does using X_i do? Coefficient of determination or R^2:

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- R^2 = fraction of the total prediction error eliminated by conditioning on X_i.
- Common interpretation: R^2 is the fraction of the variation in Y_i is "explained by" X_i.
  - $R^2 = 0$ means no relationship
  - $R^2 = 1$ implies perfect linear fit

Sampling variance for bivariate regression

- Under simple linear regression and homoskadasticity, the sampling variance of the slope was:

$$\mathbb{V}[\widehat{\beta}_1|X] = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sigma_u^2}{(n-1)S_X^2}$$

- Factors affecting the standard errors:
  - The error variance σ_u^2 (higher conditional variance of Y_i leads to bigger SEs)
  - The sample variance of X_i: S_X^2 (lower variation in X_i leads to bigger SEs)
  - The sample size n (higher sample size leads to lower SEs)

Sampling variation with 2 covariates

- Regression with an additional independent variable:

$$\mathbb{V}[\widehat{\beta}_1 | X_i, Z_i] = \frac{\sigma_u^2}{(1 - R_1^2)(n-1)S_X^2}$$

- Here, R_1^2 is the R^2 from the regression of X_i on Z_i:

$$\widehat{X}_i = \widehat{\delta}_0 + \widehat{\delta}_1 Z_i$$

- Factors now affecting the standard errors:
  - The error variance: σ_u^2
  - The sample variance of X_i: S_X^2
  - The sample size n
  - The strength of the (linear) relationship betwee X_i and Z_i (stronger relationships mean higher R_1^2 and thus bigger SEs)

Multicollinearity

Definition

Multicollinearity is defined to be high, but not perfect, correlation between two independent variables in a regression.

- With multicollinearity, we'll have $R_1^2 \approx 1$, but not exactly.
- The stronger the relationship between X_i and Z_i, the closer the R_1^2 will be to 1, and the higher the SEs will be:

$$\mathbb{V}[\widehat{\beta}_1 | X_i, Z_i] = \frac{\sigma_u^2}{(1 - R_1^2)(n-1)S_X^2}$$

- Given the symmetry, it will also increase $\text{var}(\widehat{\beta}_2)$ as well.

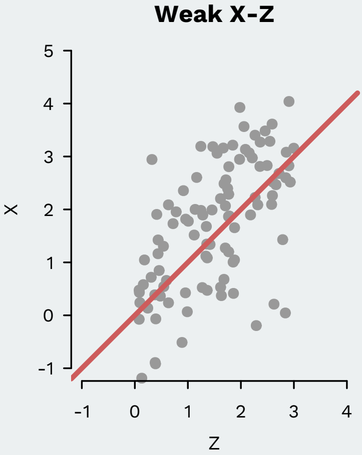
Intuition for multicollinearity

- Remember the OLS recipe:
  - $\hat{r}_{xz,i}$ are the residuals from the regression of X_i on Z_i
  - $\widehat{\beta}_1$ from regression of Y_i on $\hat{r}_{xz,i}$
- Estimated coefficient:

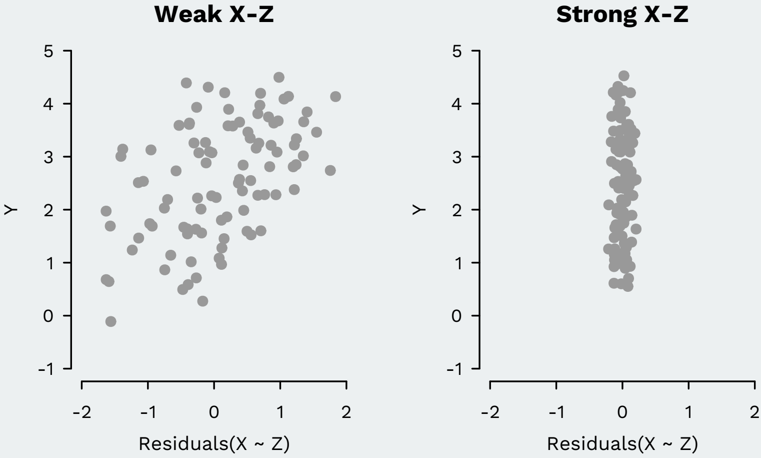
$$\widehat{\beta}_1 = \frac{\widehat{\text{cov}}[\hat{r}_{xz,i} Y_i]}{\widehat{\mathbb{V}}[\hat{r}^2_{xz,i}]}$$

- When Z_i and X_i have a strong relationship, then the residuals will have low variation
- We explain away a lot of the variation in X_i through Z_i.

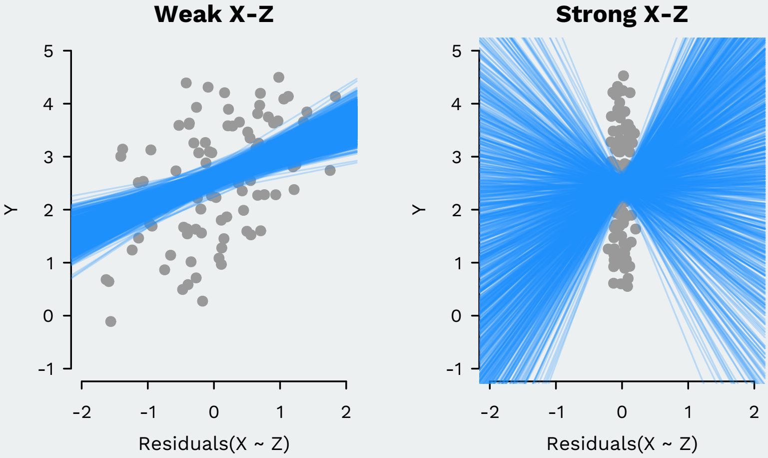
Multicollinearity, visualized



Multicollinearity, visualized



Multicollinearity, visualized



Effects of multicollinearity

- No effect on the bias of OLS.
- Only increases the standard errors.
- Really just a sample size problem:
  - If X_i and Z_i are extremely highly correlated, you're going to need a much bigger sample to accurately differentiate between their effects.

Conclusion

- In this brave new world with 2 independent variables:
  1. β's have slightly different interpretations
  2. OLS still minimizing the sum of the squared residuals
  3. Adding or omitting variables in a regression can affect the bias and the variance of OLS
- Remainder of class:
  1. Regression in most general glory (matrices)
  2. How to diagnose and fix violations of the OLS assumptions