Gov 2000 - 8. Regression with Two Independent Variables

Matthew Blackwell
Harvard University
mblackwell@gov.harvard.edu

Where are we? Where are we going?

- Last week: we learned about how to calculate a simple (bivariate) linear regression, what the properties of OLS was in this case, and how to do inference for regression parameters (slopes and intercepts).
- This week: we're going to think about how to model and estimate relationships between variables conditional on a third variable.
- Next week: generalize the entire regression model to the matrix framework and be very general.

Why do we want to add variables to the regression?

Berkeley gender bias

- In general, we want to add variables to a regression because relationships between variables in the entire sample might differ from those same relationships within subgroups of the sample.
- Graduate admissions data from Berkeley, 1973 is a famous example of this
- Acceptance rates:
  - Men: 8442 applicants, 44% admission rate

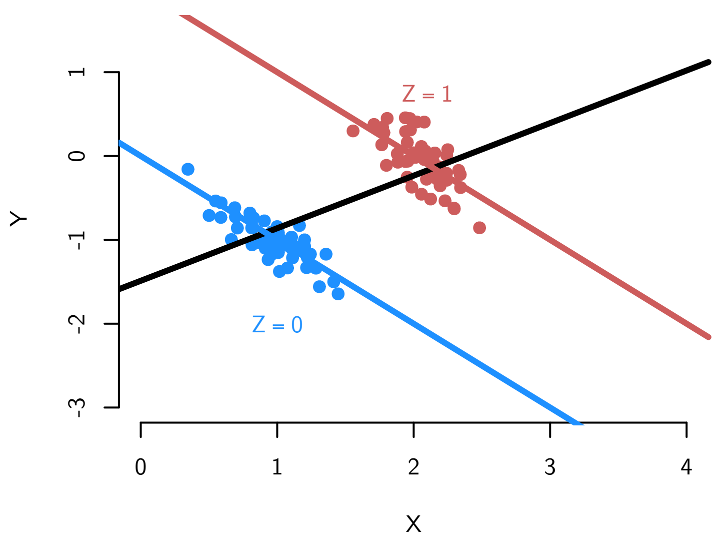
   - Women: 4321 applicants, 35% admission rate

- Evidence of discrimination toward women in admissions?
- What about 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%
D	373	6%	341	7%

- Within departments, women do somewhat better than men! Women apply to more challenging departments.
- Message: overall relationships (admissions and gender) might be different or the opposite of the same relationship conditional on a third variable (department)

Simpson's paradox



- Overall a positive relationship between Y_i and X_i here
- But within levels of Z_i, the opposite
- We call this **Simpson's paradox** or the **Yue-Simpson effect**

Basic idea

- Before our goal was to estimate the mean of Y (the dependent variable) as a function of some independent variable, X:

$$\mathbb{E}[Y_i|X_i]$$

- We learned how to do for this for binary and categorical X's with simple means.
- For continuous X's, we saw that our estimators were too noisy, so we modeled the CEF/regression function with a line:

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

- This week, we want to estimate the relationship of two variables, Y_i and X_i, conditional on a third variable, X_i:

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

- Once again, these β's are the population parameters we want to estimate. We don't get to observe them.

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.

Broad points to make

1. Slopes go from being predicted differences to predicted differences conditional on the other independent variable/covariate
2. OLS with two covariates is still just minimizing the sum of the squared residuals

3. OLS with two covariates is equivalent to two OLS regressions with 1 covariate each
4. Small adjustments to OLS assumptions and inference when adding a covariate
5. Adding or omitting variables in a regression can affect the bias and the variance of OLS

What we won't cover in lecture

1. The formula for the regression coefficients/slopes with more than 1 independent variable (we'll cover this with matrices in the coming weeks)
2. Proofs on the properties of OLS with 2 covariates (again, we'll tackle the fully general cases in future weeks)
3. The second covariate being a function of the first, such as:

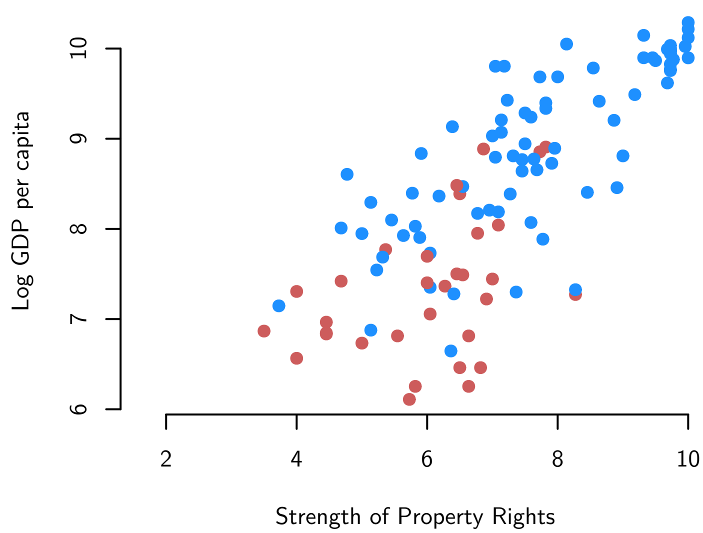
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u_i$$

   We'll get to this in future weeks too.
4. Goodness of fit for these regressions (we'll get to this as well)

ADDING A BINARY VARIABLE

Example

```
ajr <- foreign::read.dta("ajr.dta")
plot(ajr$avexpr, ajr$logpgp95, xlab = "Strength of Property Rights", ylab = "Log GDP per capita",
    pch = 19, bty = "n", col = ifelse(ajr$africa == 1, "indianred", "dodgerblue"))
```



Basics

- Let Z_i be Bernoulli/binary ($Z_i = 1$ or $Z_i = 0$)
- Here we'll use $Z_i = 1$ to indicate that i is an African country.
- Old model:
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$
- The concern might be that AJR are picking up an "African effect" if African countries have low incomes and weak property rights due to, say, a different type of colonialism.
- We include Z_i in the model to make sure that we are comparing differences in property rights within African countries and within non-African countries, not between these two groups.
- New model:
$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i$$

Two lines in one regression

- How can we interpret this model?

- One quick way is to notice that this equation with two covariates is actually just two different lines: one for when $Z_i = 1$ and one for when $Z_i = 0$

- When $Z_i = 0$:
$$\begin{aligned}\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 Z_i \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 \times 0 \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i}\end{aligned}$$

- When $Z_i = 1$:
$$\begin{aligned}\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 Z_i \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 \times 1 \\ &= (\widehat{\beta}_0 + \widehat{\beta}_2) + \widehat{\beta}_1 X_{1i}\end{aligned}$$

- This will make the interpretation of these estimates easier.

AJR model

- Let's see an example with the AJR data:

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

```
##
## Call:
## lm(formula = logpgp95 ~ avexpr + africa, data = ajr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83855 -0.28403  0.09149  0.37135  1.19757
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.65556    0.31344  18.043  < 2e-16 ***
## avexpr       0.42416    0.03971  10.681  < 2e-16 ***
## africa      -0.87844    0.14707  -5.973 3.03e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6253 on 108 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.7078, Adjusted R-squared:  0.7024
## F-statistic: 130.8 on 2 and 108 DF,  p-value: < 2.2e-16
```

Example interpretation of the coefficients

- Let's review what we've seen so far:

	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:

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

- We can read these as:
  - $\widehat{\beta}_0$: 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 change in property rights is associated with a 0.424 increase in average log incomes for two African countries
  - $\widehat{\beta}_1$: A one-unit change in property rights is associated with a 0.424 increase in average log incomes 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

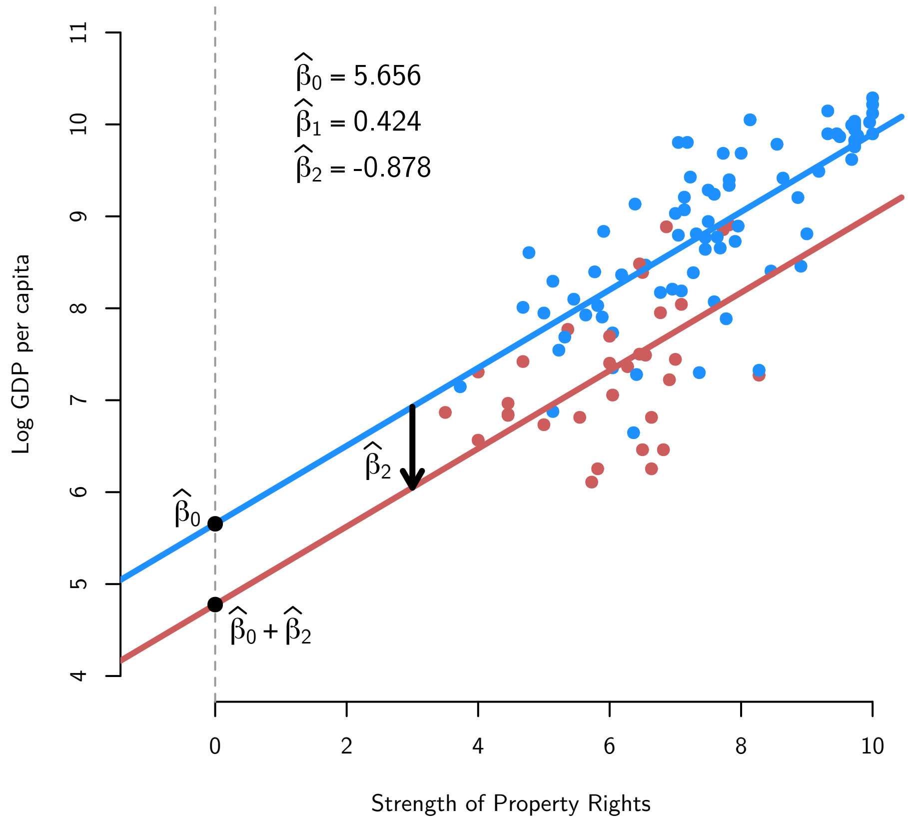
General interpretation of the coefficients

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

Adding a binary variable, visually

```
ajr.mod <- lm(logpgp95 ~ avexpr + africa, data = ajr)
plot(ajr$avexpr, ajr$logpgp95, xlab = "Strength of Property Rights", ylab = "Log GDP per capita",
    pch = 19, bty = "n", col = ifelse(ajr$africa == 1, "indianred", "dodgerblue"),
    xlim = c(-1, 10), ylim = c(4, 11))
abline(a = coef(ajr.mod)[1], b = coef(ajr.mod)[2], col = "dodgerblue", lwd = 3)
abline(a = coef(ajr.mod)[1] + coef(ajr.mod)[3], b = coef(ajr.mod)[2], col = "indianred",
    lwd = 3)
abline(v = 0, col = "grey60", lty = 2)
points(x = 0, y = coef(ajr.mod)[1], pch = 19, cex = 1.25)
text(x = 0, y = coef(ajr.mod)[1] + 0.1, expression(widehat(beta)[0]), pos = 2,
    cex = 1.25)
points(x = 0, y = coef(ajr.mod)[1] + coef(ajr.mod)[3], pch = 19, cex = 1.25)
text(x = 0, y = coef(ajr.mod)[1] + coef(ajr.mod)[3] - 0.3, expression(widehat(beta)[0] +
    widehat(beta)[2]), pos = 4, cex = 1.25)
arrows(x0 = 3, x1 = 3, y0 = coef(ajr.mod)[1] + 3 * coef(ajr.mod)[2], y1 = coef(ajr.mod)[1] +
    3 * coef(ajr.mod)[2] + coef(ajr.mod)[3], length = 0.1, lwd = 3)
text(x = 2.9, y = coef(ajr.mod)[1] + 3 * coef(ajr.mod)[2] + 0.75 * coef(ajr.mod)[3],
    expression(widehat(beta)[2]), cex = 1.25, pos = 2)
text(x = 1, y = 10.5, bquote(widehat(beta)[0] == .(round(coef(ajr.mod)[1], 3))),
    pos = 4, cex = 1.25)
text(x = 1, y = 10, bquote(widehat(beta)[1] == .(round(coef(ajr.mod)[2], 3))),
    pos = 4, cex = 1.25)
```

```r
text(x = 1, y = 9.5, bquote(widehat(beta)[2] == .(round(coef(ajr.mod)[3], 3))),
    pos = 4, cex = 1.25)
```



ADDING A CONTINUOUS VARIABLE

Basics

- Now suppose that Z_i is continuous, such as the mean temperature in that country.
- We might want to include this if geographic factors might influence the kinds of political institutions and average incomes (through health issues like malaria).
- Old model:

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

- New model:

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

AJR model, revisited

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

```
##
## Call:
## lm(formula = logpgp95 ~ avexpr + meantemp, data = ajr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7330 -0.4112  0.1191  0.4398  1.3044
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.80627    0.75184   9.053 1.27e-12 ***
## avexpr       0.40568    0.06397   6.342 3.94e-08 ***
## meantemp    -0.06025    0.01940  -3.105  0.00296 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6435 on 57 degrees of freedom
##   (103 observations deleted due to missingness)
## Multiple R-squared:  0.6155, Adjusted R-squared:  0.602
## F-statistic: 45.62 on 2 and 57 DF,  p-value: 1.481e-12
```

Interpretation

- With a continuous Z_i, we can have more than two values that it can take on:

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

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

- $\widehat{\beta}_0$: 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 change 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 a fixed value of Z_i.
- The coefficient $\widehat{\beta}_2$ measures how the predicted outcome varies in Z_i for a fixed value of X_i.

MECHANICS AND PARTIALING OUT REGRESSION

Fitted values and residuals

- Notice that we assumed that we have estimators for the various values here. But where did they come from?
- To answer this, we first need to redefine some terms from simple linear regression.
- Fitted values for $i = 1, \ldots, n$:

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

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

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

Least squares is still least squares

- How do we estimate $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$?
- Minimize the sum of the squared residuals, just like before:

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

- Not super-useful to derive these formulas, but you can do the calculus yourself if you're so inclined.
- We'll see the general version of this in the coming weeks

Estimating OLS using two steps

- We're not going to explicitly write out the OLS formulas for the two-covariate case, but there is a simple, intuitive way to do this using only simple/bivariate linear regression.
- Suppose we have the following model:

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

- We can write the OLS estimator for β_1 as:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n \widehat{r}_{xz,i} Y_i}{\sum_{i=1}^n \widehat{r}_{xz,i}^2}$$

- This is just the equation for a estimated slope in a bivariate regression where $\widehat{r}_{xz,i}$ is the only covariate
- Here, $\widehat{r}_{xz,i}$ are the residuals of a regression of X_i on Z_i:

$$X_i = \delta_0 + \delta_1 Z_i + r_{xz,i}$$
$$\widehat{r}_{xz,i} = X_i - \widehat{\delta}_0 + \widehat{\delta}_1 Z_i$$

- That is, we treat X_i as the dependent variable and Z_i as the independent variable and calculate the residuals from that regression.
- Then if we stick those residuals into a regression with Y_i as the outcome:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \widehat{r}_{xz,i}$$

- This will give us identical estimates for $\widehat{\beta}_1$ to when we run the full regression:

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

Regression property rights on mean temperature

- Let's show this with the AJR data. First we are going to regress the property rights variable, X_i, on the mean temperature variable, Z_i.
- Here we have to add an argument to the `lm()` function that tells R to exclude the missing values from the regression, but keep them in the residuals and fitted values. This is useful because we are going to create a new variable for the residuals and if R were to drop the missing values from the residuals, the columns wouldn't align properly.

```r
## when missing data exists, need the na.action in 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)
```

```
##
## Call:
## lm(formula = avexpr ~ meantemp, data = ajr, na.action = na.exclude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9770 -0.8888 -0.0350  0.8887  3.3993
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.95678    0.82015  12.140  < 2e-16 ***
## meantemp    -0.14900    0.03469  -4.295 6.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.321 on 58 degrees of freedom
##   (103 observations deleted due to missingness)
## Multiple R-squared:  0.2413, Adjusted R-squared:  0.2282
## F-statistic: 18.45 on 1 and 58 DF,  p-value: 6.733e-05
```

- Next, we store the residuals from this regression using the `residuals()` function in R. Again, the `na.exclude` option in the `lm()` call allows us to do this without errors.

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

Regression of log income on the residuals

- Now we compare the estimated slope for property rights from the regression on the residuals to the regression on the original variables:

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

```
## (Intercept)  avexpr.res
##   8.0542783   0.4056757
```

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

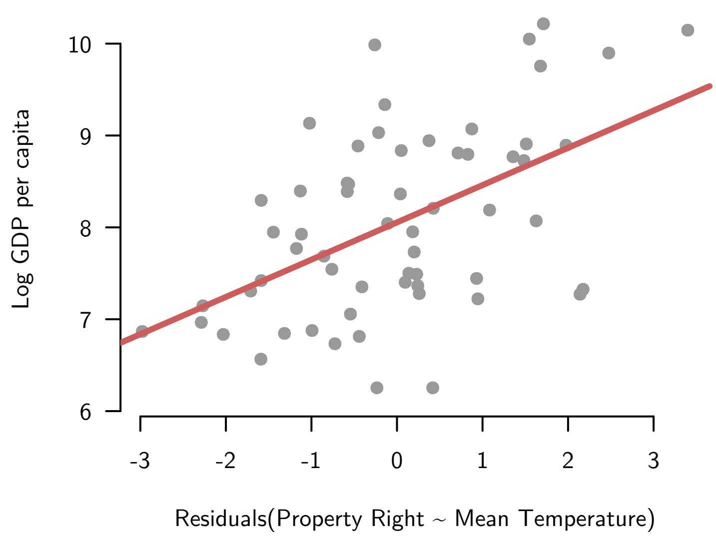
```
## (Intercept)      avexpr    meantemp
##  6.80627375  0.40567575 -0.06024937
```

- Notice how the estimated coefficient for property rights is the same in both.
- But also notice how the intercept is off. This won't be the main way we calculate OLS coefficients, but it's sometimes useful for intuition.
- It's especially useful for producing scatterplots, since this is more difficult when you have more than one explanatory variable.

Residual/partial regression plot

- We can plot the relationship between property rights and income conditional on temperature by plotting income against the same residuals.

```r
plot(x = ajr$avexpr.res, y = ajr$logpgp95, pch = 19, col = "grey60", bty = "n",
    xlab = "Residuals(Property Right ~ Mean Temperature)", ylab = "Log GDP per capita",
    las = 1)
abline(lm(logpgp95 ~ avexpr.res, data = ajr), col = "indianred", lwd = 3)
```



OLS ASSUMPTIONS & INFERENCE WITH 2 VARIABLES

OLS assumptions for unbiasedness

- When we have more than one independent variable, we need the following assumptions in order for OLS to be unbiased:

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
$$\mathbb{E}[u_i | X_i, Z_i] = 0$$

No perfect collinearity

- The "no perfect collinearity" is only truly new-sounding assumption. Notice that it replaces "variation in X_i."

Assumption 3 - (a) No independent variable is constant in the sample and (b) there are no exactly linear relationships among the independent variables.

- The first part here, (a), is just the same as in the bivariate regression. Both X_i and Z_i have to vary.
- The second part is new. It says that Z_i cannot be a deterministic, linear function of X_i. This rules out any function like this:
$$Z_i = a + bX_i$$
- Notice how this is linear (equation of a line) and there is no error, so it is deterministic. What's the correlation between Z_i and X_i? 1!

Perfect collinearity example (I)

- 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 not a linear 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))
```

```
##
## Call:
## lm(formula = logpgp95 ~ africa + nonafrica, data = ajr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.06999 -0.64783 -0.04867  0.72114  1.69849
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.71638    0.08991  96.941  < 2e-16 ***
## africa      -1.36119    0.16306  -8.348 4.87e-14 ***
## nonafrica         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9125 on 146 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.3231, Adjusted R-squared:  0.3184
## F-statistic: 69.68 on 1 and 146 DF,  p-value: 4.87e-14
```

Perfect collinearity example (II)

- Simple example:

- X_i = mean temperature in Celsius
- $Z_i = 1.8X_i + 32$ (mean temperature in Fahrenheit)

```
ajr$meantemp.f <- 1.8 * ajr$meantemp + 32
coef(lm(logpgp95 ~ meantemp + meantemp.f, data = ajr))
```

```
## (Intercept)    meantemp  meantemp.f
##  10.8454999  -0.1206948          NA
```

OLS assumptions for large-sample inference

- For large-sample inference and calculating SEs with more than one independent variable, we just need the two-variable version of the Gauss-Markov assumptions:

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

$$\mathbb{E}[u_i|X_i, Z_i] = 0$$

5. Homoskedasticity

$$\text{var}[u_i|X_i, Z_i] = \sigma_u^2$$

Inference with two independent variables in large samples

- Let's say that you have your OLS estimate $\widehat{\beta}_1$
- Furthermore, you have an estimate of the standard error for that coefficient, $\widehat{SE}[\widehat{\beta}_1]$. We haven't said how we're going to calculate those yet, but R gives them to you and we'll get to that shortly.
- Under assumption 1-5, in large samples, we'll have the following:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim N(0, 1)$$

- The same holds for the other coefficient:

$$\frac{\widehat{\beta}_2 - \beta_2}{\widehat{SE}[\widehat{\beta}_2]} \sim N(0, 1)$$

- In large samples, nothing changes about inference! Hypothesis test and confidence intervals are exactly the same as in the bivariate case.
- Note that this assumes that the number of independent variables stays fixed and n grows.

OLS assumptions for small-sample inference

- For small-sample inference, we need the Gauss-Markov plus Normal errors:

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
$$\mathbb{E}[u_i | X_i, Z_i] = 0$$
5. Homoskedasticity
$$\text{var}[u_i | X_i, Z_i] = \sigma_u^2$$
6. Normal conditional errors
$$u_i \sim N(0, \sigma_u^2)$$

Inference with two independent variables in small samples

- Under assumptions 1-6, we have the following small change to our small-n sampling distribution:
$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim t_{n-3}$$
- The same is true for the other coefficient:
$$\frac{\widehat{\beta}_2 - \beta_2}{\widehat{SE}[\widehat{\beta}_2]} \sim t_{n-3}$$
- Why $n - 3$ degrees of freedom now instead of the $n - 2$ in the simple linear regression case? Well, we've estimated another parameter, so we need to take off another degree of freedom.
- Thus, we need to make small adjustments to the critical values and the t-values for our hypothesis tests and confidence intervals.
- **Question** What happens to the size of the rejection region for an α-level test of $H_0 : \beta_1 = 0$ when we add another independent variable to the model? Does is get larger, smaller, or stay the same?

OMITTED VARIABLE BIAS

Unbiasedness revisited

- Remember that under assumptions 1-4, we get unbiased estimates of the coefficients.
- One question you might ask yourself is the following: what happens if we ignore the second independent variable and just run the simple linear regression with just X_i?
- Which of the four assumptions might we violate? Zero conditional mean error! Last week we said that for the simple linear regression we assume that:

$$\mathbb{E}[u_i|X_i] = 0$$

Omitted variable bias

- True model:
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$$
- Let's make Assumptions 1-4 about this model. Specifically, we'll say that $\mathbb{E}[u_i|X_i, Z_i] = 0$. Note that this implies that $E[u_i|X_i] = 0$ (the reverse is not true).
- Misspecified model:
$$Y_i = \beta_0 + \beta_1 X_i + u_i^*$$
- Notice here that $u_i^* = \beta_2 Z_i + u_i$, and while we know that $E[u_i|X_i] = 0$, we have made no assumptions about $E[Z_i|X_i]$, so $E[u_i^*|X_i] \neq 0$.
- Intuitively, this is saying that there is correlation between X_i and the misspecified error u_i^* due to the correlation between X_i and Z_i.
- OLS estimates from the misspecified model:

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

- Question: will $\mathbb{E}[\tilde{\beta}_1] = \beta_1$? If not, what will be the bias?

Omitted variable bias, derivation

- Simple linear regression parameter:

$$\tilde{\beta}_1 = \beta_1 + \beta_2 \widehat{\delta}_1$$

- Where the $\widehat{\delta}_1$ is the coefficient on X_i from a regression of Z_i on X_i:

$$Z_i = \delta_0 + \delta_1 X_i + v_i$$

- Remember that by OLS, this is just:

$$\widehat{\delta}_1 = \frac{\widehat{\text{cov}}(Z_i, X_i)}{\widehat{\text{var}}(X_i)}$$

- Will be positive when $\text{cov}(X_i, Z_i) > 0$ and negative when $\text{cov}(X_i, Z_i) < 0$. Will be 0 when X_i and Z_i are independent.

- Let's take expectations:

$$\begin{aligned} \mathbb{E}[\tilde{\beta}_1] &= \mathbb{E}[\beta_1 + \beta_2 \widehat{\delta}_1] \\ &= \beta_1 + \beta_2 \mathbb{E}[\widehat{\delta}_1] \\ &= \beta_1 + \beta_2 \delta_1 \end{aligned}$$

- Thus, we can calculate the bias here:

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

- In other words:

$$\text{omitted variable bias } = (\text{effect of } Z_i \text{ on } Y_i) \times (\text{effect of } X_i \text{ on } Z_i)$$

Omitted variable bias, summary

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

Including irrelevant variables

- What if we do the opposite? Include an irrelevant variable? Do we have bias in this case?

- What would it mean for Z_i to be an irrelevant variable? Basically, that we have

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

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

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

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

MULTICOLLINEARITY

Sampling variance for simple linear regression

- Under simple linear regression, we found that the distribution of the slope was the following:
$$\text{var}(\widehat{\beta}_1) = \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
- Factors affecting the standard errors (the square root of these sampling variances):
  - The error variance (higher conditional variance of Y_i leads to bigger SEs)
  - The variance of X_i (lower variation in X_i leads to bigger SEs)

Sampling variation for linear regression with two covariates

- Regression with an additional independent variable:
$$\text{var}(\widehat{\beta}_1) = \frac{\sigma_u^2}{(1 - R_1^2)\sum_{i=1}^n (X_i - \overline{X})^2}$$
- 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 (higher conditional variance of Y_i leads to bigger SEs)
  - The variance of X_i (lower variation in X_i leads to bigger SEs)
  - The strength of the relationship betwee X_i and Z_i (stronger relationships mean higher R_1^2 and thus bigger SEs)
- What happens with perfect collinearity? $R_1^2 = 1$ and the variances are infinite.

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:
$$\text{var}(\widehat{\beta}_1) = \frac{\sigma_u^2}{(1 - R_1^2)\sum_{i=1}^n (X_i - \overline{X})^2}$$
- Given the symmetry, it will also increase $\text{var}(\widehat{\beta}_2)$ as well.

Intuition for multicollinearity

- Remember that we can calculate the regression coefficient for X_i by first running a regression of X_i on Z_i and using the residuals from that regression as the independent variable:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \widehat{r}_{xz,i}$$

- But when Z_i and X_i have a strong relationship, then the residuals will be very small—we explain away a lot of the variation in X_i through Z_i.
- And we know that when the independent variable (here the residuals, $\widehat{r}_{xz,i}$) has low variance, then the standard errors of the estimator will increase.
- Basically, there is less residual variation left in X_i after "partialling out" the effect of Z_i

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.

APPENDIX

Deriving the formula for the misspecified coefficient

- Here we'll use $\widehat{\text{cov}}$ to mean the sample covariance, and $\widehat{\text{var}}$ to be the sample variance.

$$
\begin{aligned}
\tilde{\beta}_1 &= \frac{\widehat{\text{cov}}(Y_i, X_i)}{\widehat{\text{var}}(X_i)} && \text{(OLS formulas)} \\
&= \frac{\widehat{\text{cov}}(\beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i, X_i)}{\widehat{\text{var}}(X_i)} && \text{(Linearity in correct model)} \\
&= \frac{\widehat{\text{cov}}(\beta_0, X_i)}{\widehat{\text{var}}(X_i)} + \frac{\widehat{\text{cov}}(\beta_1 X_i, X_i)}{\widehat{\text{var}}(X_i)} + \frac{\widehat{\text{cov}}(\beta_2 Z_i, X_i)}{\widehat{\text{var}}(X_i)} + \frac{\widehat{\text{cov}}(u_i, X_i)}{\widehat{\text{var}}(X_i)} && \text{(covariance properties)} \\
&= 0 + \frac{\widehat{\text{cov}}(\beta_1 X_i, X_i)}{\widehat{\text{var}}(X_i)} + \frac{\widehat{\text{cov}}(\beta_2 Z_i, X_i)}{\widehat{\text{var}}(X_i)} + 0 && \text{(zero mean error)} \\
&= \beta_1 \frac{\widehat{\text{var}}(X_i)}{\widehat{\text{var}}(X_i)} + \beta_2 \frac{\widehat{\text{cov}}(Z_i, X_i)}{\widehat{\text{var}}(X_i)} && \text{(properties of covariance)} \\
&= \beta_1 + \beta_2 \frac{\widehat{\text{cov}}(Z_i, X_i)}{\widehat{\text{var}}(X_i)} \\
&= \beta_1 + \beta_2 \widehat{\delta}_1 && \text{(OLS formulas)}
\end{aligned}
$$