Gov 2000: 9. 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 phenomenon. In this year, there were 8442 male applicants with a 44% admission rate and 4321 female applicants with 35% admission rate. The key substantive question is whether or not this is evidence of discrimination toward women in admissions.

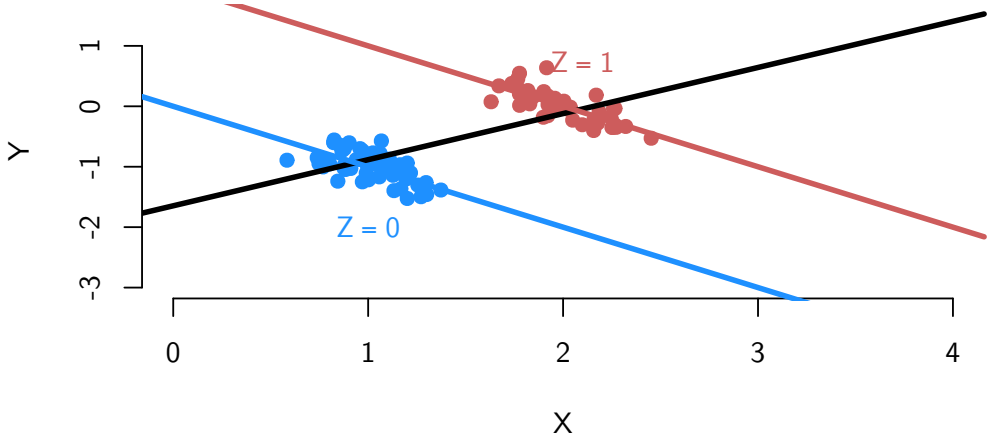
Of course, admission rates vary by department—some departments are "easier" to get into than others. In this case, it makes sense to look at the difference in admissions rates *within* departments. This data is in the following table:

	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! It seems that reason that women are admitted at lower rates overall is because women apply to departments with lower acceptance rates. (Of course, this doesn't imply no discrimination—it could be the case that the university restricts the size of departments with high interest from women.) The lesson here is that overall/marginal relationships (admissions and gender) might be different or the opposite of the same relationship conditional on a third variable (department).

This admissions data is an example of what we call **Simpson's paradox** of the **Yule-Simpson effect**. Another example is given in the figure below. In this simulated data, there is a positive relationship between Y_i and X_i among all observations. But if we look within levels of Z_i, there is a negative relationship.

```
## plot example of simpson's paradox
z <- rbinom(100, 1, 0.5)
x <- z + rnorm(100, 1, 0.2)
y <- 2 * z - x + rnorm(100, 0, 0.2)
plot(x,y, bty = 'n', xlim = c(0, 4), ylim = c(-3, 1.5), xlab = expression(X), ylab = expression(Y), col = ifelse
abline(lm(y~x), lwd = 3)
text(x = 1, y = -2, expression(Z == 0), col = 'dodgerblue')
text(x = 2.1, y = 0.75, expression(Z == 1), col = 'indianred')
abline(a=0, b=-1, col = 'dodgerblue', lwd = 3)
abline(a=2, b=-1, col = 'indianred', lwd = 3)
```



Why might we want to look at the relationship between X_i and Y_i within levels of Z_i? There are three primary reasons:

- **Descriptive**: Allows us to understand the relationships in the data. For example, 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.

With these goals in mind, we can start to think about how write the CEF as a function of two r.v.s as opposed to just one. 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, with the CEF, $\mathbb{E}[Y_i|X_i, Z_i]$. In general, we often will a linear and additive relationship:

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

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

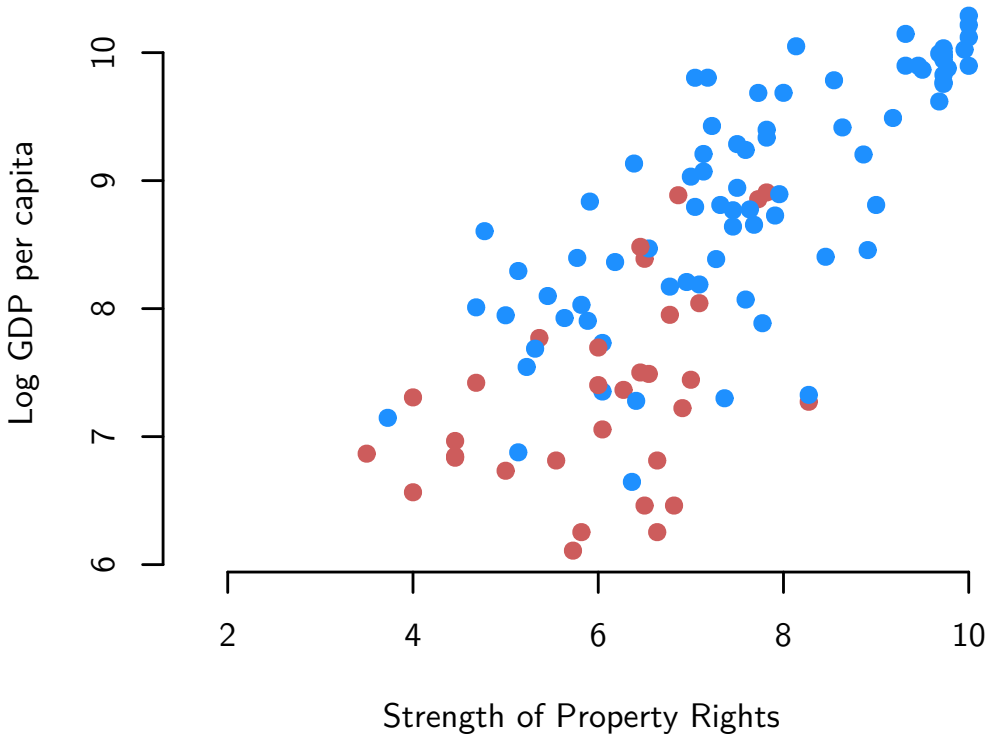
The major points of moving from a single independent variable to two independent variables are:

1. Estimated 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.

ADDING A BINARY VARIABLE

```
ajr <- foreign::read.dta('../data/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'))
```



Let Z_i be Bernoulli/binary ($Z_i = 1$ or $Z_i = 0$). Here we'll use $Z_i = 1$ to indicate that country i is an African country. Suppose we were to run a simple linear regression of log GDP per captia (Y_i) on just expropriation risk (X_i). What might be wrong with this analysis? One concern might be that this model i picking up an "African effect" if African countries have low incomes and weak property rights due to, say, a different type of colonialism. To avoid this problem, we might 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. This new regression will be:

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

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 $Z_i = 0$:

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

When $Z_i = 1$:

$$\begin{aligned}\widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 \times 1 \\ &= (\widehat{\beta}_0 + \widehat{\beta}_2) + \widehat{\beta}_1 X_1\end{aligned}$$

This will make the interpretation of these estimates easier.

Let's see an example with the AJR data:

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

```
##
## 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
```

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:

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

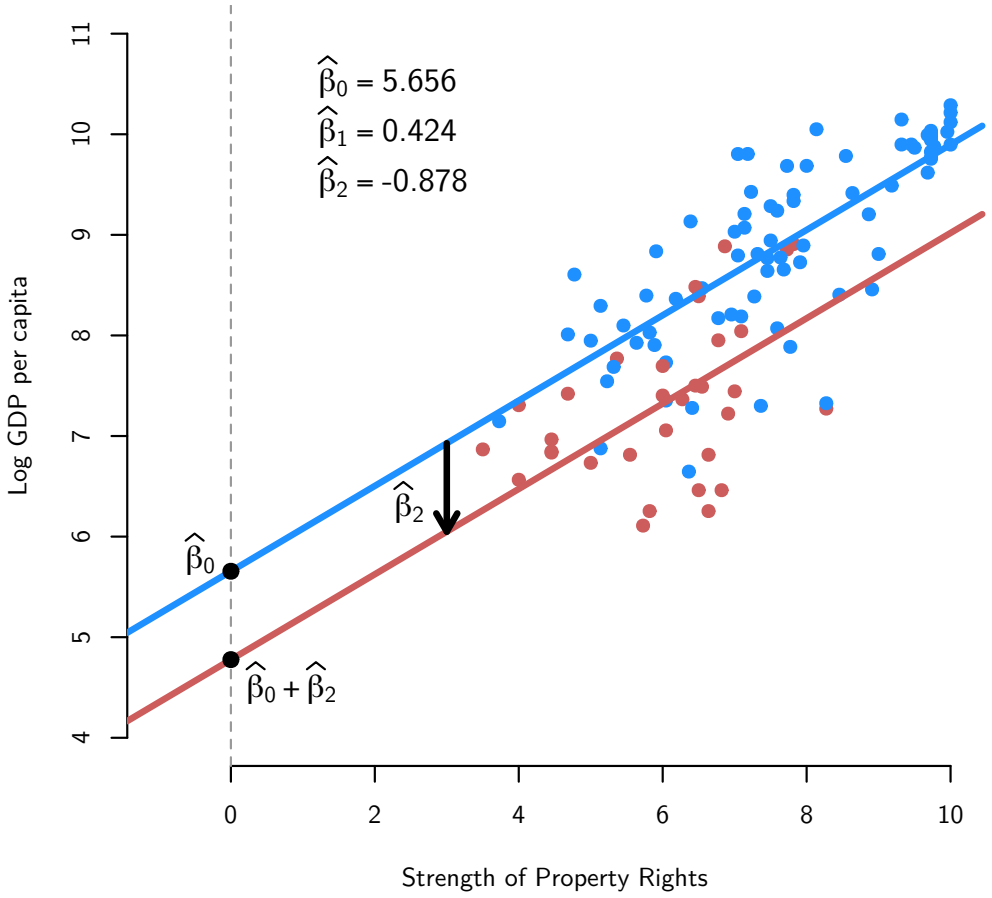
We can read these as:

- $\widehat{\beta}_0$: 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 (or between 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

More generally, we can interpret the coefficients with a binary 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 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

We can see how this works visually:



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). If we write the regression adding this Z_i it looks the same as the binary Z_i:

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

Let's see the output from this regression:

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

```
##
## 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
```

How do we interpret the coefficients from this regression? 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
$Z_i = 0\,^\circ\text{C}$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
$Z_i = 21\,^\circ\text{C}$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 21$	$\widehat{\beta}_1$
$Z_i = 24\,^\circ\text{C}$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 24$	$\widehat{\beta}_1$
$Z_i = 26\,^\circ\text{C}$	$\widehat{\beta}_0 + \widehat{\beta}_2 \times 26$	$\widehat{\beta}_1$

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

Specifically, we can interpret the coefficients from this regression:

- $\widehat{\beta}_0$: 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

More generally, with a regression of Y_i on a continuous X_i and Z_i, we can interpret the coefficients as:

- 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

Up to this point, we have just run our regressions without discussing how to calculate the estimators. Where do they come from? To answer this, we first need to redefine some terms from simple linear regression. We'll define the **fitted values** for $i = 1, \ldots, n$:

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

In a similar way to simple regression, we'll also define the **residuals** for $i = 1, \ldots, n$:

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

How do we estimate $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$? We'll simply generalize the method from simple regression. That is, we'll minimize the sum of the squared residuals:

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

It's possible to derive explicit formulas for these estimators, but we'll hold off on these until we can derive OLS for any number of independent variables.

Even thought 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:

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

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.

```
## 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)
```

```
##
## 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.

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

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

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

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

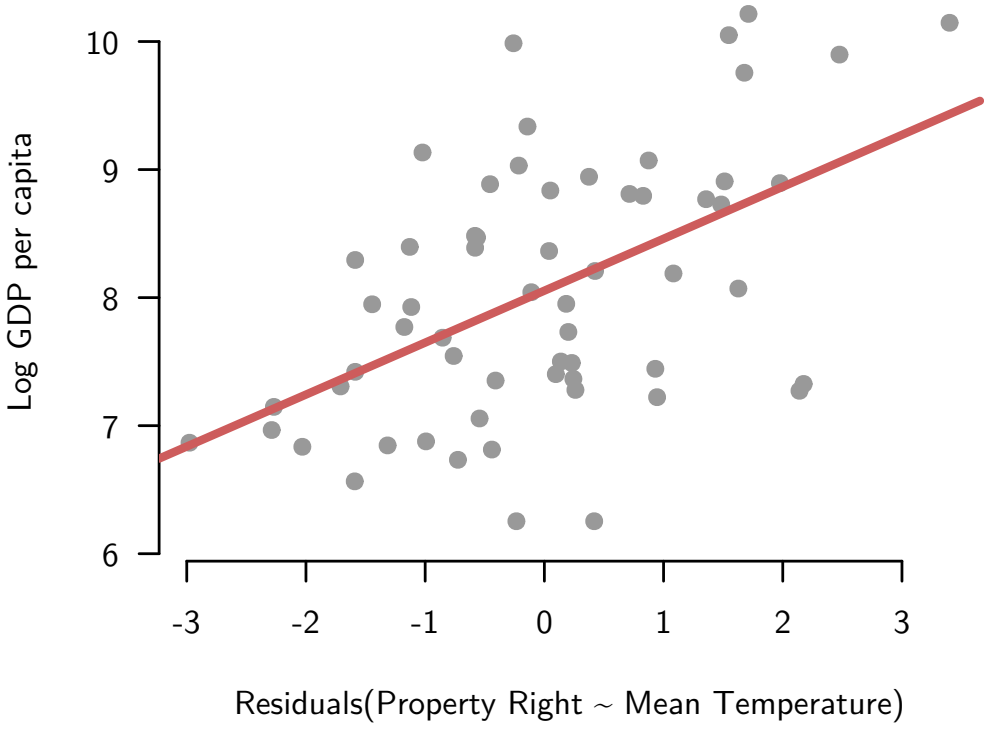
```
coef(lm(logpgp95 ~ avexpr + meantemp, data = ajr))
```

```
## (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.

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

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

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!

A simple example, if trivial example of a perfect collinearity is if we have the following two variables: $X_i = 1$ if a country is **not** in Africa and 0 otherwise, and $Z_i = 1$ 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.

A situation that may appear to be perfect collinearity is when X_i is property rights and $Z_i = X_i^2$. 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.

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

```
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)  8.71638    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
```

Another example of collinearity is when we rescale and recenter a variable. For example, let X_i be mean temperature in Celsius and let $Z_i = 1.8X_i + 32$ be the mean temperature in Fahrenheit. Obviously, this is a deterministic function, so R will act accordingly:

```
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

To arrive at a simple formula for the variance and standard error of the OLS coefficients, it is common to assume homoskedasticity. Again, this isn't strictly necessary to derive the sampling variance, but it does make the expression more simple and it is what almost all statistical software packages assume when they present SEs.

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: $\mathbb{V}[u_i|X_i, Z_i] = \sigma_u^2$

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, $\widehat{\mathsf{se}}[\widehat{\beta}_1]$. 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:

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

The same holds for the other coefficient:

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

In large samples, nothing changes about inference! Hypothesis test and confidence intervals are exactly the same as in the bivariate case.

For small-sample exact 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: $\mathbb{V}[u_i|X_i, Z_i] = \sigma_u^2$
6. Normal conditional errors: $u_i \sim N(0, \sigma_u^2)$

Under assumptions 1-6, we have the following small change to our small-n sampling distribution:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\mathsf{se}}[\widehat{\beta}_1]} \sim t_{n-3}$$

The same is true for the other coefficient:

$$\frac{\widehat{\beta}_2 - \beta_2}{\widehat{\mathsf{se}}[\widehat{\beta}_2]} \sim t_{n-3}$$

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.

OMITTED VARIABLE BIAS

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:

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

In this scenario, the true model would be:

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

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). Then, let's think about running a misspecified model that omits Z_i:

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

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.

Let's write the OLS estimates from the misspecified model as:

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

Our main question is how to relate $\tilde{\beta}_1$ to $\widehat{\beta}_1$ from the true model. In particular, we want to know if the OLS estimator on the misspecified model will be unbiased, so that $\mathbb{E}[\tilde{\beta}_1] = \beta_1$? If not, what will be the bias?

In short, we can write the OLS estimator from the misspecified simple linear regression as:

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

Here 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:

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

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

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

With this in hand, we can sign the possible bias if we know the sign of the X_i and Z_i relationship and the Y_i and Z_i relationship.

	$\text{cov}(X_i, Z_i) > 0$	$\text{cov}(X_i, Z_i) < 0$	$\text{cov}(X_i, Z_i) = 0$
$\beta_2 > 0$	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:

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

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

MULTICOLLINEARITY

Sampling variance for simple linear regression

Under simple linear regression, we found that the distribution of the slope was the following:

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

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)

Regression with an additional independent variable:

$$\mathbb{V}(\widehat{\beta}_1) = \frac{\sigma_u^2}{(1 - R_1^2)\sum_{i=1}^n (X_i - \overline{X})^2}$$

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.

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:

$$\mathbb{V}(\widehat{\beta}_1) = \frac{\sigma_u^2}{(1 - R_1^2) \sum_{i=1}^{n} (X_i - \overline{X})^2}$$

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

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

What is the effect of multicollinearity? Importantly, there is no effect on the bias of OLS. It only increases the standard errors. In some sense, it is 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 $\widehat{\mathbb{V}}$ to be the sample variance.

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