11. Multiple regression: introduction

LPO.8800: Statistical Methods in Education Research

Sean P. Corcoran

Last time

- Bivariate regression
- Prediction equation, predicted values, residuals (prediction errors)
- Ordinary least squares (OLS) "line of best fit"
- Interpreting regression intercept and slope
- Assessing goodness of fit (R^2)
- Conditional mean interpretation of regression
- Inference about the population slope: confidence intervals and hypothesis tests
- Regression diagnostics with residuals

Correlation vs. causation

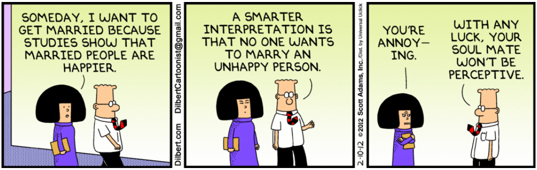
Correlation vs. causality

Generally speaking, regression slopes (and correlations) *cannot* be interpreted as *causal*. Examples:

- Russian cholera epidemic: peasants observed that in communities with lots of doctors, there were lots of cholera cases; doctors were murdered.
- SAT prep courses: in 1988 Harvard interviewed its freshmen and found that those who took SAT coaching courses scored 63 points lower than those who did not.
  - A dean concluded that the SAT courses were unhelpful and that "the coaching industry is playing on parental anxiety."

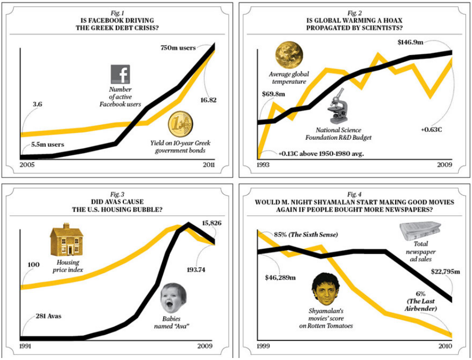
On the Russian cholera riots of 1830-31 see:
`http://www.unm.edu/~ybosin/documents/rus_chol.pdf`

Correlation vs. causality



Not an endorsement of the *Dilbert* cartoonist.

Correlation vs. causality



What is causality?

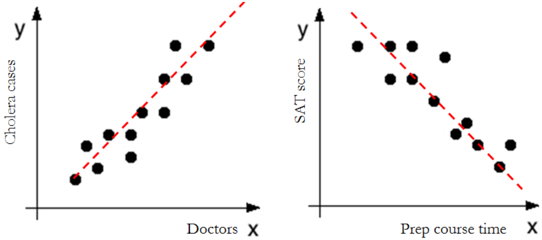
There is clearly an *association* between these variables, but can we say that changes in X are *causing* changes in Y ($\mathbf{X} \rightarrow \mathbf{Y}$)? What is a causal effect?

A causal effect involves a **counterfactual** comparison between two different states of the world: e.g., y whenever $x = 1$ vs. y whenever $x = 0$, assuming all else is held constant.

A regression can be considered causal when it provides a counterfactual comparison.

Correlation vs. causality

Imagine collecting data and conducting a simple regression analysis for the cholera and SAT score examples:



Do these regressions provide counterfactual comparisons? Most likely not.

Correlation vs. causality

Considering the above two examples:

- Russian cholera epidemic: it is unlikely doctors (X) caused the cholera cases (Y), since the presence of cholera preceded the arrival of the doctors. This is a case of **reverse causality** ($\mathbf{Y} \rightarrow \mathbf{X}$).

- SAT prep courses: it is *possible* that the prep course worsened SAT performance (the time ordering is appropriate). But it is more likely a **confounding factor** explains *both* enrollment in the prep course *and* low SAT scores (e.g., test anxiety, poor prior academic preparation). The association may be **spurious**.

Good research design involves ruling out alternative explanations for an observed association.

Correlation vs. causality


Rigby

Enough with the wind already

Received April 1

Ever since they installed all those big fans up on the hill it's become even windier. Whose bright idea was that?

I've noticed when they're off, we get a nice calm spell. Please turn them off, at least on weekends. (Word count: 40)

JEFF FORBES
Idaho Falls

Guest columns, solicited: 450 words max • Guest column

Experimental vs. observational data

Ruling out alternative explanations can be very difficult to do in social science and education research. The researcher is typically working with *observational* data, and has no control over assignment to "treatment" conditions of interest. Examples:

- Does smoking cause lung cancer?
- Would a smaller class size improve learning?
- Does education increase labor market productivity and earnings?
- Is parental divorce detrimental to childrens' outcomes?
- Do mask mandates reduce the transmission of infectious diseases?

Experimental vs. observational data

This is in contrast to the medical researcher who can randomly assign subjects to receive a new drug or a placebo. With a **randomized controlled trial**, she can confidently attribute any systematic differences in the subjects' outcomes to the drug (and not due to a confounder).

Experimental vs. observational data

In the absence of random assignment, attributing causality is difficult to do, and depends on sound research design, data availability, and a good theoretical understanding of factors that affect variation in the outcome Y.

Side note: outliers and examples of contradictory cases are **not** sufficient for ruling out causal relationships! Causal effects are a description of how X affects Y *on average*, not in a deterministic sense.

- A high-poverty school that is "beating the odds" does not demonstrate that poverty has no effect on academic achievement.
- A smoker that lives to 102 is not proof that smoking does not cause lung cancer.

Regression and causal inference

Does computer ownership improve math performance?

Does access to a computer at home improve math performance? Suppose you use a random sample of 8th grade students to estimate the following population regression:

$$E(y|x) = \beta_0 + \beta_1 x$$

where y is an 8th grade math test score and $x = 1$ if the student has a computer at home and $x = 0$ otherwise. u is the population error term:

$$y = \beta_0 + \beta_1 x + u$$

Does computer ownership improve math performance?

x is a dichotomous variable, so the intercept and slope provide two population means (no computer at home vs. computer at home):

$$E(y|x) = \beta_0 + \beta_1 x$$

$$E(y|x = 0) = \beta_0$$
$$E(y|x = 1) = \beta_0 + \beta_1$$

β_1 is the *difference* in the two population means.

Does computer ownership improve math performance?

Estimates of β_0 and β_1 using the NELS data:

```
. tabstat achmat08, by(computer) stat(mean n)

Summary for variables: achmat08
     by categories of: computer (computer owned by family in eighth grade?)

computer |      mean         N
---------+--------------------
      no |  54.94897       263
     yes |  58.41321       237
---------+--------------------
   Total |  56.59102       500
------------------------------

. reg achmat08 computer

      Source |       SS           df       MS      Number of obs   =       500
-------------+----------------------------------   F(1, 498)       =     17.73
       Model |  1496.05776         1  1496.05776   Prob > F        =    0.0000
    Residual |  42030.8486       498  84.3992944   R-squared       =    0.0344
-------------+----------------------------------   Adj R-squared   =    0.0324
       Total |  43526.9064       499  87.2282693   Root MSE        =    9.1869

------------------------------------------------------------------------------
    achmat08 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    computer |   3.464233   .8228153     4.21   0.000     1.847616    5.080851
       _cons |   54.94897   .5664891    97.00   0.000     53.83597    56.06198
------------------------------------------------------------------------------
```


Does computer ownership improve math performance?

Question: should we think of β_1 as the *causal* effect of computer ownership on math performance?

Answer: only if it describes a counterfactual comparison—the difference in (mean) y between having a computer and not, assuming all else is held constant.

It seems unlikely β_1 can be interpreted this way. There are likely confounding variables related to both computer ownership and math achievement.

Does computer ownership improve math performance?

When we use data to estimate the population regression:

$$E(y|x) = \beta_0 + \beta_1 x$$
$$y = \beta_0 + \beta_1 x + u$$

we are thinking of u as things that affect y that are unrelated to x. OLS uses that assumption when finding $\hat{\beta}_0$ and $\hat{\beta}_1$.

But what if the (causal) relationship we care about involves things that are related to both x and y that we would like to hold constant?

Does computer ownership improve math performance?

For instance, suppose kids with computers at home would do better in math *with or without a computer*—perhaps because they live in higher income households. We could represent this as:

$$E(u|x = 0) = 0$$
$$E(u|x = 1) = \gamma$$

We can think of γ as representing "baseline differences" between kids with computers and kids without computers.

Does computer ownership improve math performance?

The implication of this is:

$$E(y|x=0) = \beta_0$$
$$E(y|x=1) = \beta_0 + \beta_1 + \gamma$$

The difference in means is the causal effect of computer ownership (β_1) plus **selection bias** (γ). The kinds of kids who have computers at home have other resources that would help them in math in any case.

We can't disentangle these with a simple linear regression! This implies that our slope estimate $\hat{\beta}_1$ is a biased estimator of the causal effect β_1. We have an **omitted variables bias** problem.

Omitted variables bias

If we are interested in the causal effect of x on y and estimate:

$$y = \beta_0 + \beta_1 x + u$$

then the OLS estimator $\hat{\beta}_1$ suffers from **omitted variables bias** (OVB) if:

1. x is correlated with another variable that is not included in the analysis (i.e., it is part of u) **and**
2. that omitted variable is also a determinant of y

Omitted variables bias: class size example

Consider the class size and test scores example (from Lecture 10):

$$testscr = \beta_0 + \beta_1 str + u$$

Now consider three potential omitted variables:

- Percent of students in the school district who are English learners
- Time of day the test is administered
- Staff parking lot area per pupil

Which of these will result in omitted variables bias? Why?

Signing the direction of omitted variables bias

It is possible to sign $(+/-)$ the direction of OVB in a simple regression. It can be shown that in large samples, the OLS slope estimator will estimate:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \underbrace{\rho_{xu}\frac{\sigma_u}{\sigma_x}}_{OVB}$$

In other words, it estimates the true β_1 plus OVB. The two σ's are positive, so the direction of the bias is determined by ρ_{xu}, the correlation between x and the omitted variable u.

Note: a larger sample will <u>not</u> help with OVB!

Another note: with random assignment, there is no OVB! ($\rho_{xu} = 0$)

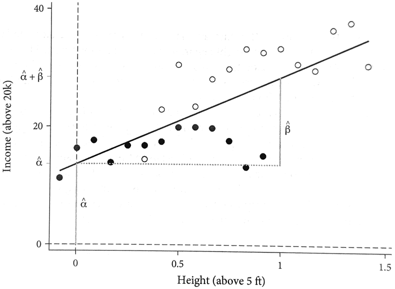
Signing the direction of omitted variables bias

Let's apply this:

- Math performance and computer ownership
- District test scores and class size (where %EL is the omitted variable)
- Annual income and height

Do taller people earn more?

Regressing annual income on height:



Note: women shown in solid dots, men shown in hollow dots.

Controlling for confounders

Does computer ownership improve math performance?

One way we might **control** for the effects of SES in the above example is to condition on SES. For sake of example (there are better ways to do this) divide SES into two groups, low or high:

```
. egen ses2=cut(ses), group(2)

. * the above command creates a new variable 'ses2' that splits 'ses' into two equal-sized groups (low and high)

. table computer ses2, contents(mean achmat08 n achmat08)
```

computer owned by family in eighth grade?	ses2 0	1
no	**53.5129** **169**	**57.53085** **94**
yes	**55.46333** **78**	**59.86031** **159**

The 3.46 point "effect" of computer ownership on math achievement is smaller after conditioning on SES. For high SES students, the "effect" is 2.32 points; for low SES students, 1.95 points.

Do AP courses improve high school math achievement?

Using the NELS data:

```
. reg achmat12 approg

      Source |       SS           df       MS      Number of obs   =       493
-------------+----------------------------------   F(1, 491)       =     88.17
       Model |  4688.50685         1  4688.50685   Prob > F        =    0.0000
    Residual |   26110.315       491   53.177831   R-squared       =    0.1522
-------------+----------------------------------   Adj R-squared   =    0.1505
       Total |  30798.8219       492  62.5992314   Root MSE        =    7.2923

------------------------------------------------------------------------------
    achmat12 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      approg |   6.175654   .6577047     9.39   0.000     4.883391    7.467917
       _cons |   53.69983   .4767134   112.65   0.000     52.76318    54.63648
------------------------------------------------------------------------------
```

We might be concerned that β_1 does not represent a causal effect. Factors related to math achievement (in u) are related to AP course taking.

Do AP courses improve high school math achievement?

For sake of example, divide students into quintiles (5 groups) of SES. How does course taking vary with SES quintile?

Quintile	took AP
1	0.409
2	0.385
3	0.465
4	0.610
5	0.718

Kids from higher SES households are more likely to take AP. It is likely that $\rho_{xu} > 0$.

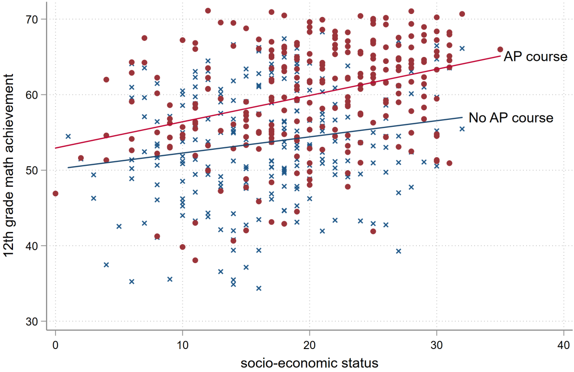
Do AP courses improve high school math achievement?

How does 12th grade math achievement differ by AP course status *conditional* on SES quintile?

		Mean math:		
Quintile	took AP	no AP	AP	**Diff**
1	0.409	52.2	55.7	**3.5**
2	0.385	52.3	57.1	**4.9**
3	0.465	55.4	59.4	**4.1**
4	0.610	54.3	60.4	**6.1**
5	0.718	55.3	62.9	**7.6**

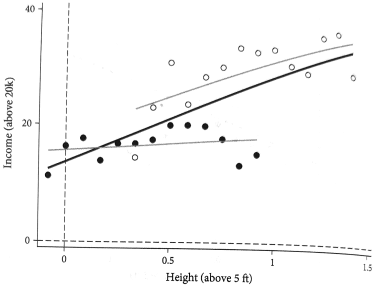
Note: you would get the same Diff values by estimating a separate regression for each SES quintile.

Do AP courses improve high school math achievement?



Do taller people earn more? Revisited

Regressing annual income on height separately for men and women:



Note: women shown in solid dots, men shown in hollow dots.

Do AP courses improve high school math achievement?

We might like a single estimate that represents the *average* difference in 12th grade math achievement that comes from AP course attendance *conditional on* SES. This implies some kind of weighted average across SES groups.

Multiple regression is one way of obtaining such an average.

Multiple regression

Multiple regression

Multiple regression extends the linear regression model to > 1 explanatory variable. It is a way of statistically controlling for other variables that are ignored in simple regression. With two explanatory variables:

$$E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

and with k explanatory variables:

$$E(y|x_1, ..., x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

There is now an intercept and k slope coefficients.

Multiple regression: interpretation

The multiple regression has a similar interpretation to the single variable regression. It gives us a mean value for y given values of x_1, x_2, etc.

$$E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Each slope coefficient is interpreted as the **partial** effect of that x on y *holding the other explanatory variable(s) constant*.

For example:

- β_1 is the change in the mean y for a one-unit change in x_1, holding x_2 constant.
- β_2 is the change in the mean y for a one-unit change in x_2, holding x_1 constant.

Multiple regression: least squares

We can again use ordinary least squares (OLS) to find the intercept $\hat{\beta}_0$ and slope coefficients $\hat{\beta}_1, ..., \hat{\beta}_k$ by minimizing the sum of the squared deviations between the actual data points and predicted values:

$$\overset{\min}{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

$$\overset{\min}{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - ... - \hat{\beta}_k x_k\right)^2$$

Multiple regression: predicted values and residuals

The definition of predicted values and residuals are the same for multiple regression. For a given $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$:

The predicted value of y is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

The residual is the difference between the actual and predicted y:

$$\hat{u} = y - \hat{y}$$

Multiple regression in Stata

To implement multiple regression in Stata, continue to use `regress` and include the additional explanatory variables in your variable list:

```
. reg achmat08 computer i.ses2

      Source |       SS           df       MS      Number of obs   =       500
-------------+----------------------------------   F(2, 497)       =     21.59
       Model |  3478.87468         2  1739.43734   Prob > F        =    0.0000
    Residual |  40048.0317       497  80.5795407   R-squared       =    0.0799
-------------+----------------------------------   Adj R-squared   =    0.0762
       Total |  43526.9064       499  87.2282693   Root MSE        =    8.9766

------------------------------------------------------------------------------
    achmat08 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    computer |   2.149567   .8465355     2.54   0.011     .4863372    3.812796
      1.ses2 |   4.193894   .8454511     4.96   0.000     2.532795    5.854992
       _cons |   53.45002    .630632    84.76   0.000     52.21098    54.68905
------------------------------------------------------------------------------
```

How should we interpret the intercept and coefficients?

Multiple regression in Stata

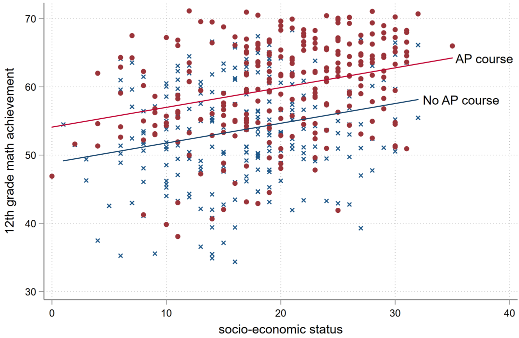
AP course example using (continuous) SES as a control variable:

```
. reg achmat12 approg ses

      Source |       SS           df       MS      Number of obs   =       493
-------------+----------------------------------   F(2, 490)       =     66.29
       Model |  6558.71774         2  3279.35887   Prob > F        =    0.0000
    Residual |  24240.1041       490  49.4696003   R-squared       =    0.2130
-------------+----------------------------------   Adj R-squared   =    0.2097
       Total |  30798.8219       492  62.5992314   Root MSE        =    7.0335

------------------------------------------------------------------------------
    achmat12 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      approg |   5.227734   .6528237     8.01   0.000     3.945055    6.510413
         ses |   .2895499    .047092     6.15   0.000     .1970227    .3820771
       _cons |   48.86905   .9103236    53.68   0.000     47.08043    50.65767
------------------------------------------------------------------------------
```

Note: this regression constrains the slope on *ses* to be the same for AP and non-AP students. That is, it finds the best fit regression equation where the slope is the same for these two groups.

Multiple regression in Stata

Using continuous SES as control variable:



Multiple regression: measures of fit

The measures of fit in Lecture 10 are the same for multiple regression, and have the same interpretation:

$$R^2 = \frac{SSM}{TSS} \text{ or } 1 - \frac{SSE}{TSS}$$

The SER (Root MSE in Stata) has a minor modification in that we divide by $n-k-1$ (where k is the number of slope coefficients):

$$SER = \sqrt{\frac{\sum_{i=1}^{n} \widehat{u_i}^2}{n-k-1}} = \sqrt{\frac{SSE}{n-k-1}}$$

Adjusted R^2

Note that R^2 will *always increase* with an additional explanatory variable, unless the slope on that variable is zero. The **adjusted R-squared** adjusts for the number of regressors:

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\frac{SSE}{TSS}$$

If you add explanatory variables, SSE goes down, *increasing* your R^2. However, an additional explanatory variable increases k, which *decreases* your R^2.

Adjusted R^2 is almost always a bit lower than R^2. In some extreme cases of poor fit it can be less than zero.

Multiple regression and causality

Can multiple regression coefficients can be interpreted as causal? In most cases, unfortunately not. While the regression controls for some omitted variables, there are likely others. This is where careful research design comes in (see later courses!)

Multiple regression inference

Much of what you learned about regression inference (hypothesis tests, confidence intervals) holds here as well. Under a few assumptions, with a large sample size n, the OLS estimators of each slope coefficient:

- Are **unbiased** (on average, you get the true slope).
- Have an approximate **normal** distribution.
- Have a somewhat complicated formula for their standard error, but it gets smaller as n gets large.

This means you can proceed as usual with your t-statistics, p-values, etc., from the Stata output! Note: as in Lecture 10, the standard error formula depends on homoskedasticity assumption. (Can use `robust` standard errors if you don't want to assume this).

Example using California school district data

Using *caschool* data:

1. Estimate the simple regression of test scores (*testscr*) on class size (*str*) and interpret the slope on class size.
2. Add the percent of students who are English learners (*el_pct*) as a regressor and interpret both slopes.
3. Add the number of computers per student (*comp_stu*) and the percent of students eligible for free or reduced price meals (*meal_pct*) and interpret all slopes. *What are these control variables trying to accomplish?*
4. Which of the predictor variables above are *statistically significant*?
5. Interpret the R^2 and adjusted R^2 in the above regressions. How do these change from (1)-(3)?

Partial relationships

Multiple regression with $k = 2$

Consider the multiple regression with two explanatory variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

When x_1 and x_2 are uncorrelated

When x_1 and x_2 are uncorrelated, the OLS estimators of β_1 and β_2 are:

$$\hat{\beta}_1 = r_{y1} \frac{s_y}{s_1}$$

$$\hat{\beta}_2 = r_{y2} \frac{s_y}{s_2}$$

where r_{y1} is the correlation between y and x_1, and r_{y2} is the correlation between y and x_2. (s_1 is the standard deviation of x_1, and s_2 is the standard deviation of x_2). Notice these are equivalent to the formula for $\hat{\beta}$ in the simple regression case. The r_{y1} and r_{y2} are sometimes called **zero-order correlations**.

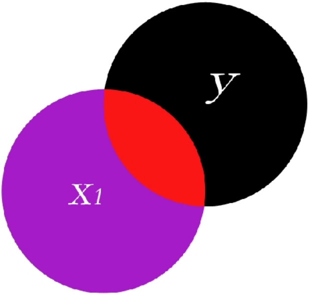
When x_1 and x_2 are uncorrelated

R^2 has the same interpretation in multiple regression: the fraction of overall variation in y that is explained by the x variables. When x_1 and x_2 are uncorrelated, R^2 is simply:

$$R^2 = r_{y1}^2 + r_{y2}^2$$

(the sum of the two squared zero-order correlations). R—the square root of R^2—is called the **multiple correlation**

Venn diagram with one explanatory variable

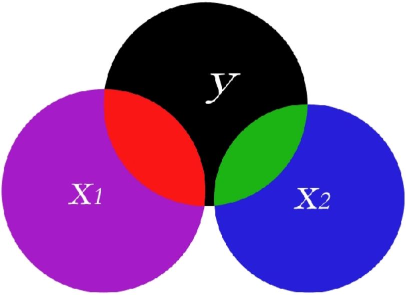


Venn diagram with one explanatory variable

The circle labeled y represents variation in y, and the circle labeled x_1 represents variation in x_1.

- Think of the overlap (red) as variation in y "explained" by variation in x_1
- The red area represents correlation between x_1 and y: information used by the regression to estimate $\hat{\beta}_1$
- The black area is variation in y *unexplained* by variation in x_1 ("residual" variation)
- The proportion of y covered by x_1 represents the R^2

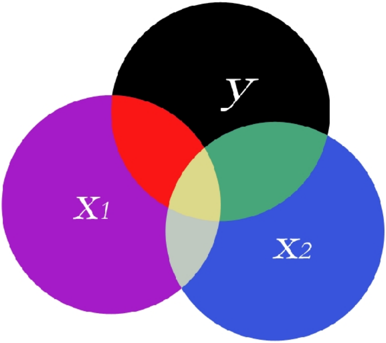
Venn diagram with two explanatory variables - 1



Venn diagram with two explanatory variables - 1

- Think of the overlap between y and x_1 (red) as variation in y "explained" by variation in x_1
- Think of the overlap between y and x_2 (green) as variation in y "explained" by variation in x_2
- x_1 and x_2 do not overlap (they are uncorrelated), so it is easy to attribute variation in y separately to x_1 and x_2
- The red area represents information used by the regression to estimate $\hat{\beta}_1$
- The green area represents information used by the regression to estimate $\hat{\beta}_2$
- The black area is variation in y unexplained by x_1 or x_2
- The proportion of y covered by x_1 and x_2 represents R^2

Venn diagram with two explanatory variables - 2



Venn diagram with two explanatory variables - 2

- In this case x_1 and x_2 overlap—they are *correlated* (represented by the yellow area), thus it is not as clear how to attribute variation in y separately to x_1 and x_2
- The red area represents the unique information used by the regression to estimate $\hat{\beta}_1$
- The green area represents the unique information used by the regression to estimate $\hat{\beta}_2$
- Both the red and green areas are *smaller* than those in example 1—we have less certainty about how much of y can be attributed to each explanatory variable

When x_1 and x_2 are correlated

When x_1 and x_2 are correlated, the OLS estimators of β_1 and β_2 can be written:

$$\hat{\beta}_1 = \left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_y}{s_1}\right)$$

$$\hat{\beta}_2 = \left(\frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_y}{s_2}\right)$$

where r_{12} is the correlation between x_1 and x_2 (and other terms were defined previously). Notice what happens if $r_{12} = 0$ (i.e. if there is no correlation between x_1 and x_2).

Example: private school attendance and math achievement

```
. reg achmat12 private

      Source |       SS           df       MS      Number of obs   =       500
-------------+----------------------------------   F(1, 498)       =     10.92
       Model |  665.476286         1  665.476286   Prob > F        =    0.0010
    Residual |  30351.3075       498  60.9464005   R-squared       =    0.0215
-------------+----------------------------------   Adj R-squared   =    0.0195
       Total |  31016.7838       499  62.1578833   Root MSE        =    7.8068

------------------------------------------------------------------------------
    achmat12 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     private |   2.630139   .7959513     3.30   0.001     1.066303    4.193976
       _cons |   56.22278   .4058571   138.53   0.000     55.42538    57.02019
------------------------------------------------------------------------------

. reg achmat12 private ses

      Source |       SS           df       MS      Number of obs   =       500
-------------+----------------------------------   F(2, 497)       =     29.23
       Model |  3264.62388         2  1632.31194   Prob > F        =    0.0000
    Residual |  27752.1599       497  55.8393559   R-squared       =    0.1053
-------------+----------------------------------   Adj R-squared   =    0.1017
       Total |  31016.7838       499  62.1578833   Root MSE        =    7.4726

------------------------------------------------------------------------------
    achmat12 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     private |   .5417838    .821064     0.66   0.510    -1.071401    2.154968
         ses |   .3552102   .0520643     6.82   0.000     .2529169    .4575035
       _cons |   50.21781   .9620882    52.20   0.000     48.32755    52.10807
------------------------------------------------------------------------------
```


Example: private school attendance and math achievement

```
. corr achmat12 private ses
(obs=500)

             | achmat12  private      ses
-------------+---------------------------
    achmat12 |   1.0000
     private |   0.1465   1.0000
         ses |   0.3232   0.3728   1.0000

. summ achmat12 private ses

    Variable |        Obs        Mean    Std. Dev.
-------------+-------------------------------------
    achmat12 |        500    56.90662    7.884027
     private |        500         .26    .4390735
         ses |        500      18.434    6.924271
```

$$\hat{\beta}_1 = \left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_y}{s_1}\right)$$

$$\hat{\beta}_1 = \left(\frac{0.1465 - 0.3232 * 0.3728}{1 - 0.3728^2}\right)\left(\frac{7.884}{0.439}\right) = 0.542$$

Example: private school attendance and math achievement

```
. corr achmat12 private ses
(obs=500)

             |  achmat12  private      ses
-------------+---------------------------
    achmat12 |    1.0000
     private |    0.1465   1.0000
         ses |    0.3232   0.3728   1.0000

. summ achmat12 private ses

    Variable |        Obs        Mean    Std. Dev.
-------------+-----------------------------------
    achmat12 |        500    56.90662    7.884027
     private |        500         .26    .4390735
         ses |        500      18.434    6.924271
```

$$\hat{\beta}_2 = \left(\frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_y}{s_2}\right)$$

$$\hat{\beta}_2 = \left(\frac{0.3232 - 0.1465 * 0.3728}{1 - 0.3728^2}\right)\left(\frac{7.884}{6.924}\right) = 0.3552$$

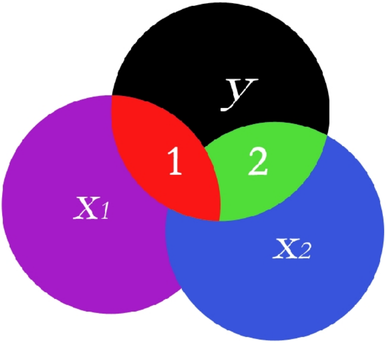
When x_1 and x_2 are correlated

With two explanatory variables the R^2 can be written as:

$$R^2 = r_{y1}^2 + r_{y2|1}^2$$

- First part: the proportion of variation in y explained by x_1
- Second part: the proportion of variation in y explained by x_2 *beyond that explained by* x_1 (a **semi-partial** correlation)

When x_1 and x_2 are correlated



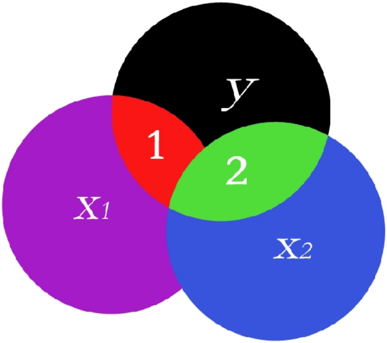
When x_1 and x_2 are correlated

Equivalently, the R^2 can be written as:

$$R^2 = r_{y2}^2 + r_{y1|2}^2$$

- First part: the proportion of variation in y explained by x_2
- Second part: the proportion of variation in y explained by x_1 *beyond that explained by* x_2 (a *semi-partial* correlation)

When x_1 and x_2 are correlated



Semi-partial correlations

The correlations $r^2_{y2|1}$ and $r^2_{y1|2}$ are called *semi-partial* or *part* correlations. They represent the correlation observed between y and that part of x_1 (or x_2) that is uncorrelated with x_2 (or x_1).

Multiple regression and R^2

More on multiple regression and R^2:

- R^2 still ranges between 0 and 1
- R^2 will be high when the x's are highly correlated with y
- R^2 will not fall below the highest R^2 with an individual x
- R^2 cannot *decrease* when additional xs are added to the regression equation (see earlier slide on Adjusted R^2)
- R^2 will be larger when the explanatory variables are not *redundant*—i.e. their intercorrelation is low
- There is usually diminishing returns to additional explanatory variables (a greater chance of redundancy)

Multicollinearity

Multicollinearity is the condition when explanatory variables in a regression are highly correlated. The consequence of this is that it becomes more difficult to discern how much of the variation in y is "due to" each individual x. This is a bigger problem the smaller the sample size.

Semi-partial (part) correlations

The **semi-partial (or part) correlation** between y and x_1 is the correlation observed between y and that part of x_1 that is uncorrelated with the other x variables.

- The *square* of the semi-partial correlation is the amount by which R^2 decreases when that explanatory variable is excluded.
- It is also the proportion of the variation in y that is explained by x_1 only
- This can be used to assess the relative importance of the explanatory variables (in terms of independent predictive power).
- Could be used to guide model specification.

Can obtain semi-partial correlations in Stata using `pcorr y x1 x2`

Semi-partial (part) correlations

Try this using the math achievement and private school example above.

- Regress *achmat08* on *ses* and *private*, note R^2 (0.1053)
- Regress *achmat08* on *ses* alone, note R^2 (0.1045)
- Regress *achmat08* on *private* alone, note R^2 (0.0215)
- Get squared semi-partial correlations `pcorr achmat08 private ses`

```
. pcorr achmat12 private ses
(obs=500)

Partial and semipartial correlations of achmat12 with
```

Variable	Partial Corr.	Semipartial Corr.	Partial Corr.^2	Semipartial Corr.^2	Significance Value
private	0.0296	0.0280	0.0009	0.0008	0.5097
ses	0.2926	0.2895	0.0856	0.0838	0.0000