12. Multivariate relationships

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)
- Interpreting regression slope and intercept
- 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

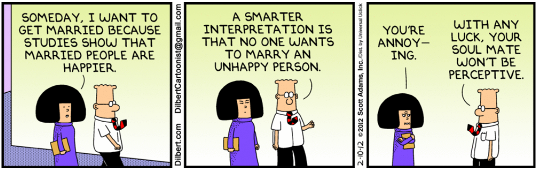
Correlation vs. causality, revisited

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

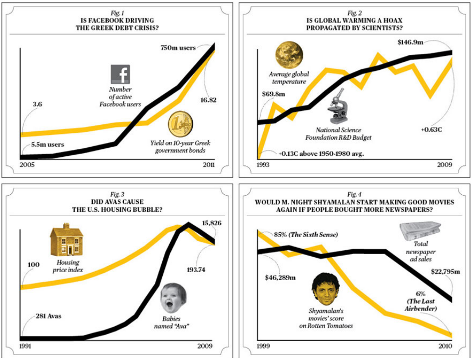
- Russian cholera epidemic: peasants observed that in communities with lots of doctors, there were lots of cholera cases; doctors were murdered.
- SAT prep courses: in 1988 Harvard interviewed its freshmen and found that those who took SAT coaching courses scored 63 points lower than those who did not.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

Causal questions imply an "all else equal" assumption.

Correlation vs. causality, revisited

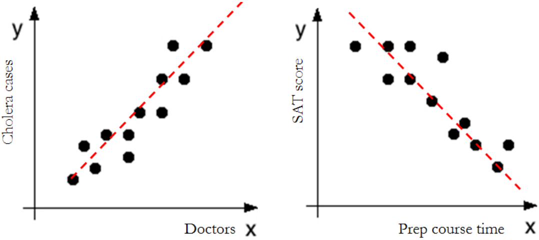


Correlation vs. causality, revisited



Correlation vs. causality, revisited

Imagine collecting data and conducting a simple regression analysis for each case:



In the lefthand figure, each data point is a community. In the righthand figure, each data point is a college applicant.

Correlation vs. causality, revisited

There is clearly an *association* between these pairs of variables, but can we say that variation in X is causing variation in Y ($\mathbf{X} \rightarrow \mathbf{Y}$)? For any correlation between two variables, there are three possible explanations: X is causing Y, Y is causing X, or some other factor is causing both.

Criteria for a causal relationship:

- Association between the two variables
- An appropriate time ordering
- Elimination of alternative explanations

Correlation vs. causality, revisited

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 ($Y \rightarrow X$).
- SAT prep courses: it is *possible* that the prep course worsened SAT performance (the time ordering is appropriate). But it is more likely a third factor explains *both* enrollment in the prep course *and* low SAT scores (e.g., test anxiety, poor prior academic preparation). One would need to eliminate alternative explanations before making a causal connection.
  - The association between SAT performance and prep course participation may be **spurious**.

Correlation vs. causality, revisited


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 and observational data

Eliminating 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. Consider again these questions:

- 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?
- Does participation in an SAT prep course improve SAT performance?

Experimental and observational data

This is in contrast to the medical researcher who can randomly assign subjects to receive a new drug or a placebo. With this study design, she controls the time ordering, and can confidently attribute any systematic differences in the subjects' outcomes to the drug (and not due to some third factor). There are fewer opportunities for such designs in social science.

Eliminating alternative explanations

In the absence of random assignment, the elimination of alternative explanations 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.

Outliers and anecdotal 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.

Controlling for other variables

In practice how does one eliminate alternative explanations for the association between X and Y? Typically one tries to find ways of removing the effects of other variables from this association. This is called **controlling** for the effects of other variables. It is the statistical equivalent of a lab researcher "holding other variables constant."

Suppose we are interested in the relationship between X_1 and Y. The variables we wish to remove the effects of are called **control variables** or **covariates** (e.g., $X_2, X_3, ..., X_k$).

- We statistically control for a third variable X_2 by examining the relationship between X_1 and Y *conditional on* X_2 (i.e., for fixed values of X_2). With a relatively small number of values, this can be done with *partial tables* that show the conditional mean of Y given X

Controlling for other variables

Does computer ownership benefit 8th grade math achievement?

```
. 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
------------------------------------------------------------------------------
```

Controlling for other variables

The association between computer ownership and math achievement could be spurious, and explained by a third factor correlated with computer ownership *and* math achievement (e.g., SES). Let's control for SES, using 2 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	ses2 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 both low and high SES students, the "effect" is 2.32 points.

Types of multivariate relationships

Ways in which the response variable Y may be related to explanatory and control variables:

- An association between Y and X_1 that is fully attributable to a third variable X_2 is said to be **spurious**. The association disappears after controlling for X_2. (E.g., a common time trend).
- Y may have **multiple causes** X_1, X_2, etc. Controlling for X_2 may change but not eliminate the association between Y and X_1 (and vice versa). (E.g., SES and computer ownership)
- The effect of X_1 on Y may be *indirect*, through an intervening variable (or **mediator**) X_2. (E.g., education $\rightarrow$ income $\rightarrow$ health, perhaps).

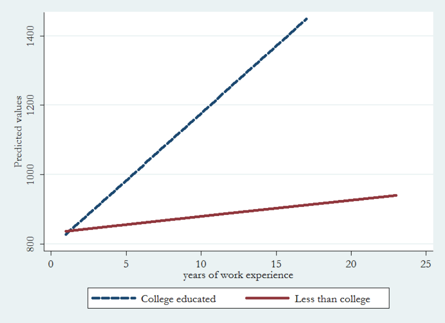
The first two examples are often called **confounders**. Not accounting for other X variables provides a distorted view of the relationship between X_1 and Y.

Types of multivariate relationships

- A third variable may mask (i.e. understate) the association between Y and X_1. This is sometimes called a **suppressor variable**.
  - Example: Head Start participation and achievement (suppressor variable: poverty).
- When the effect of a variable X_1 on Y varies with the level of a third variable X_2, this is called a **statistical interaction** or an **interaction effect**.

Interaction effect

The relationship between monthly earnings (Y) and years of work experience (X_1) depends on the level of education (X_2) (wage2.dta):



Interaction effect

```
. corr wage exper
(obs=935)

             |     wage    exper
-------------+------------------
        wage |   1.0000
       exper |   0.0022   1.0000


. corr wage exper if educ>=16
(obs=247)

             |     wage    exper
-------------+------------------
        wage |   1.0000
       exper |   0.2733   1.0000


. corr wage exper if educ<16
(obs=688)

             |     wage    exper
-------------+------------------
        wage |   1.0000
       exper |   0.0563   1.0000
```


Multiple regression

Multiple regression allows one to statistically control for other explanatory variables that are ignored in simple regression. With 2 explanatory variables the best fit "line" is:

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

With k explanatory variables, one can consider the best fit "line":

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + ... + b_k x_k$$

There is now an intercept and k slope coefficients to compute.

Multiple regression

As before, the best fit "line" is the one where the intercept a and slope coefficients $b_1, b_2, ..., b_k$ minimize the sum of the squared deviations between the actual data points and the regression line:

$$\min_{a,b} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

$$\min_{a,b} \sum_{i=1}^{n} (y_i - a - b_1 x_1 - b_2 x_2 - ... - b_k x_k)^2$$

(Of course with multiple X variables the best fit "line" is no longer a line, but multi-dimensional.)

Example: multiple regression

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

```
. regress wage exper

      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  1,   933) =    0.00
       Model |  732.242855     1  732.242855           Prob > F      =  0.9467
    Residual |   152715436   933  163682.139           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0011
       Total |   152716168   934  163507.675           Root MSE      =  404.58

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .2024031   3.026148     0.07   0.947    -5.736443    6.141249
       _cons |   955.6049    37.4111    25.54   0.000     882.1853    1029.025
------------------------------------------------------------------------------

. regress wage exper educ

      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  2,   932) =   73.26
       Model |  20747023.1     2  10373511.5           Prob > F      =  0.0000
    Residual |   131969145   932  141597.795           R-squared     =  0.1359
-------------+------------------------------           Adj R-squared =  0.1340
       Total |   152716168   934  163507.675           Root MSE      =  376.29

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   17.63777   3.161775     5.58   0.000     11.43275    23.84279
        educ |   76.21639   6.296604    12.10   0.000     63.85922    88.57355
       _cons |  -272.5279   107.2627    -2.54   0.011    -483.0323   -62.02344
------------------------------------------------------------------------------
```

Example: multiple regression

The slope coefficients are now interpreted as **marginal** or **partial** effects: the linear relationship between Y and X_1, *conditional on* (or "holding constant") X_2 and any other included control variables.

- Conditional on years of education (holding constant years of education), we predict that an additional year of work experience is associated with \$17.64 additional monthly earnings.
- Conditional on years of work experience (holding constant work experience), we predict that an additional year of education is associated with \$76.22 additional monthly earnings.

The *prediction equation* can be used to find the "best prediction" of Y given values of $X_2, ..., X_K$.

Example: multiple regression

For example, let years of experience be $X_1 = 10$ and years of education completed be $X_2 = 14$. Our best prediction of monthly earnings is:

$$\hat{y} = -272.53 + 17.64 * 10 + 76.22 * 14 = 970.95$$

When x_1 and x_2 are uncorrelated

When x_1 and x_2 are uncorrelated, the OLS estimators of b_1 and b_2 are:

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

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

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 $\hat{b}$ in the simple regression case. The r_{yk} are sometimes called **zero-order correlations**.

When x_1 and x_2 are uncorrelated

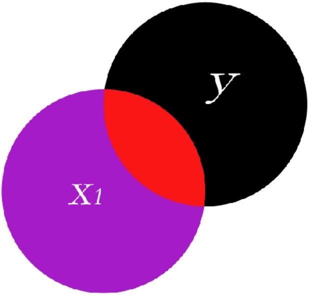
In multiple regression, R^2 can still be used as a measure of fit, interpreted in the same way: the fraction of overall variation in y that is explained by the regression. When x_1 and x_2 are uncorrelated, R^2 is simply:

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

(the sum of the two squared zero-order correlations)

- R^2 is the *coefficient of determination*
- R–the square root of R^2–is the *multiple correlation*

Venn diagram with one explanatory variable

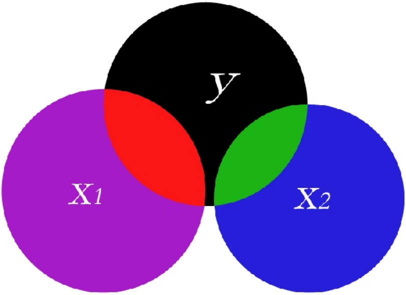


Venn diagram with one explanatory variable

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 information used by the regression to estimate $\hat{b}_1$
- 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

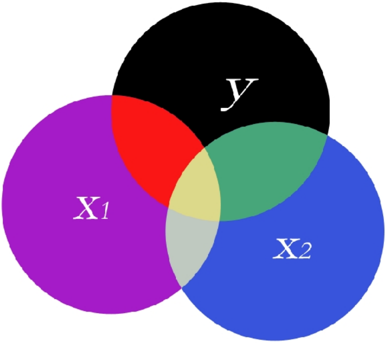
Venn diagram with two explanatory variables - 1



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 $\hat{b}_1$
- The green area represents information used by the regression to estimate $\hat{b}_2$
- 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 $\hat{b}_1$
- The green area represents the unique information used by the regression to estimate $\hat{b}_2$
- 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 b_1 and b_2 are:

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

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

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
```

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

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

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
```


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

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

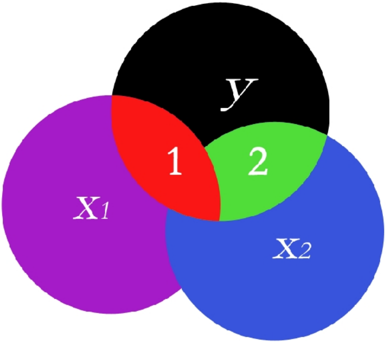
When x_1 and x_2 are correlated

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

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

- 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



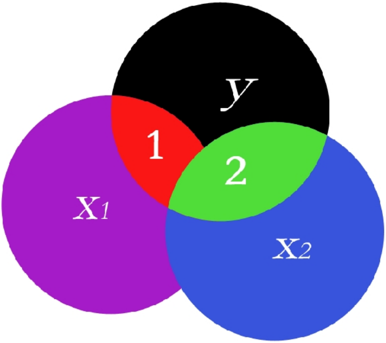
When x_1 and x_2 are correlated

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. The 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

Some facts about 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
- 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)

Adjusted R^2

The calculated R^2 tends to overestimate the population R^2 (it is upwardly biased). The smaller is N relative to the number of explanatory variables K, the more R^2 will be inflated. An **adjusted $\mathbf{R^2}$** is often used instead:

$$R^2_{ADJ} = 1 - \left(1 - R^2\right) \frac{(N-1)}{N-K-1}$$

Holding N constant, the adjusted R^2 "penalizes" you for including additional explanatory variables K.

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

Partial regression equations

Suppose you are using the multiple regression equation:

$$y = a + b_1x_1 + b_2x_2$$

The estimated prediction equation is not graphable in two dimensions since there are two x variables. However, consider fixing x_2 at a particular value, $X2$. Now the regression equation is:

$$y = (a + (b_2X2)) + b_1x_1$$

This can be graphed in two dimension. See example using math achievement and private school, SES.

Partial plots

A partial regression plot ("added variable plot") for y versus x_1 uses the following *residuals*:

- $\hat{u}$ from a regression of y on *all other* xs
- $\hat{v}$ from a regression of x_1 on *all other* xs

Note the slope of the best fit lit in this added variable plot equals the slope from the multiple regression. Likewise, the slope of the best fit line equals the slope of a regression of y on $\hat{v}$.

F-statistic

The F-statistic reported in regression output is calculated as:

$$F = \frac{R^2/k}{(1 - R^2)(n - k - 1)}$$

F follows a F sampling distribution with k and $n - k - 1$ degrees of freedom. (k is the number of explanatory variables in the regression). This F-statistic can be used to test the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_1 : \text{at least one } \beta_j \neq 0$$

It is a test of the *joint significance* of the explanatory variables.