Regression III: Advanced Methods

William G. Jacoby

Department of Political Science

Michigan State University

http://polisci.msu.edu/jacoby/icpsr/regress3

Simple Linear Regression

- If the relationship between Y and X is *linear*, then the linear regression provides an elegant summary of the statistical dependence of Y on X. If X and Y are bivariate normal, the summary is a complete description
- Simple linear regression fits a straight line, determined by two parameter estimates—an intercept and a slope
- The general idea is to determine the expectation of Y given X:

$$\begin{aligned} Y|X &= E(Y|X) + [Y|X - E(Y|X)] \\ &= E(Y|X) + \text{residual} \end{aligned}$$

- From here on, we shall label the residual component as *E* (for error)

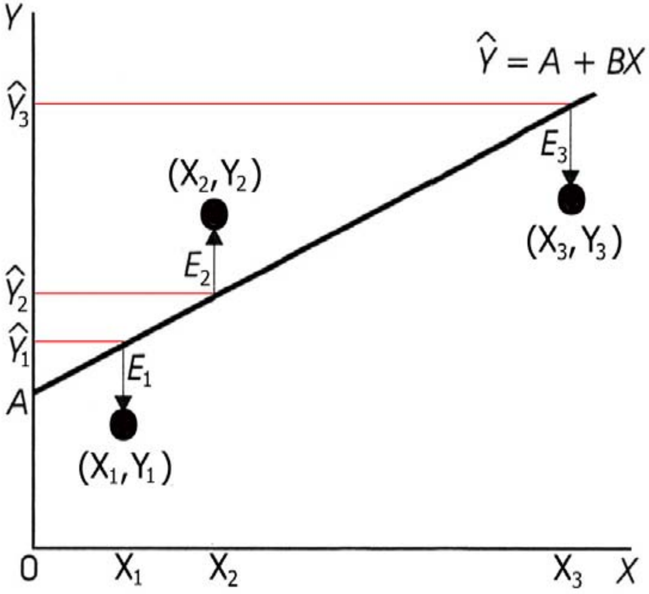
Ordinary Least Squares (OLS)

- OLS fits a straight line to data by minimizing the residuals (vertical distances of observed values from predicted values):

$$
\begin{aligned}
Y_i &= A + BX_i + E_i \\
&= \widehat{Y}_i + E_i \\
E_i &= Y_i - \widehat{Y}_i
\end{aligned}
$$

where

$\widehat{Y}_i$ is the fitted value of Y_i



- The solution for A and B ensures that the *sum of the errors from the mean function* is as small as possible:

$$
\begin{aligned}
\sum_{i=1}^{n} E_i =& \sum (Y_i - \overline{Y}) - B \sum (X_i - \overline{X}) \\
=& 0 - B \times 0 = 0
\end{aligned}
$$

- To avoid the problem of positive and negative residuals cancelling each other out when summed, we use the sum of the *squared* residuals, $\sum E_i^2$
- For a fixed set of data, each possible choice of A and B yields different sums of squares—*i.e.*, the residuals depend on the choice of *A* and *B*. We can express this relationship as the *function* *S*(*A,B*):

$$\begin{aligned}
S(A,B) &= \sum_{i=1}^{n} E_i^2 \\
&= \sum (Y_i - A - BX_i)^2 \\
&= \sum (Y_i - A - BX_i)(Y_i - A - BX_i) \\
&= \sum (Y_i^2 - Y_iA - Y_iBX_i - Y_iA + A^2 \\
&\qquad + ABX_i - Y_iBX_i + ABX_i + B^2X_i^2) \\
&= \sum (Y_i^2 - 2Y_iA - 2Y_iBX_i + 2ABX_i + B^2X^2 + A^2)
\end{aligned}$$

- We can then find the least squares line by taking the partial derivatives of the sum of squares function with respect to the coefficients:

$$S(A,B) = \sum (Y_i^2 - 2Y_iA - 2Y_iBX_i + 2ABX_i + B^2X^2 + A^2)$$

$$\frac{\partial S(A,B)}{\partial A} = \sum (-2Y_i + 2BX_i + 2A)$$

$$= \sum (-2)(Y_i - A - BX_i)$$

$$\frac{\partial S(A,B)}{\partial B} = \sum (-2Y_iX_i + 2AX_i + 2BX_i^2)$$

$$= \sum (-X_i)(2)(Y_i - A - BX_i)$$

- Setting the partial derivatives to 0, we get the simultaneous linear equations (the *normal equations*) for *A* and *B:*

$$An + B\sum X_i = \sum Y_i$$
$$A\sum X_i + B\sum X_i^2 = \sum X_i Y_i$$

- Solving the normal equations gives the least-squares coefficients for *A* and *B*:

$$A = \bar{Y} - B\bar{X}$$
$$B = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2}$$

- See from the denominator of the equation for B, that *if the X values are identical the coefficients are not uniquely defined—i.e.*, if X is a constant an infinite number of slopes can be obtained:

$$\text{If } X \text{ is constant, } \sum(X_i - \overline{X})^2 = 0$$

- The second normal equation also implies that the residuals are uncorrelated with X:

$$\begin{aligned}\sum X_i E_i &= \sum X_i (Y_i - A - BX_i) \\ &= \sum X_i Y_i - A \sum X_i - B \sum X_i^2 \\ &= 0\end{aligned}$$

- ***Interpretation of the coefficients***:
  - ***Slope coefficient, B****:* The average change in *Y* associated with a one unit increase in *X (conditional on linear relationship between X and Y)*
  - ***Intercept, A****:* The fitted value (conditional mean) of *Y* at *X*=0. That is, it is where the line passes through the *Y*-axis of the scatterplot. Often A is used only to find the “start” or “height” of the line—*i.e.*, it is not given literal interpretation.

Multiple regression

- It is relatively straightforward to extend the simple regression model to several predictors. Consider a model with two predictors:

$$\widehat{Y} = A + B_1X_1 + B_2X_2$$

- Rather than fit a straight line, we now fit a ***flat regression plane*** to a three-dimensional plot. The residuals are the vertical distances from the plane:

$$E_i = Y_i - \widehat{Y}_i = Y_i - (A + B_1X_{i1} + B_2X_{i2})$$

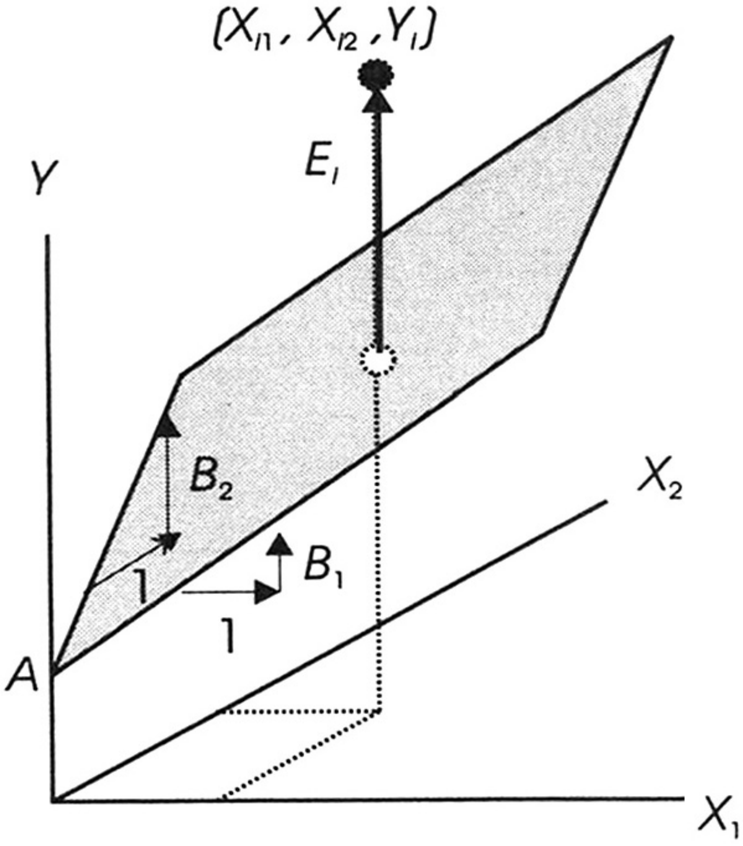
- The goal, then, is to fit the plane that comes as close to the observations as possible—we want the values of A, B_1 and B_2 that minimize the sum of squared errors:

$$S(A, B_1, B_2) = \sum E_i^2 = \sum (Y_i - A - B_1X_{i1} - B_2X_{i2})^2$$

The multiple regression plane

- B_1 and B_2 represent the partial slopes for X_1 and X_2 respectively
- For each observation, the values of X_1, X_2 and Y are plotted in 3-dimensional space
- The regression plane is fit by *minimizing the sum of the squared errors*
- E_i (the residual) is now the *vertical distance* of the observed value Y from the fitted value of Y *on the plane*

Figure 5.5 from Fox (1997)



The Sum-of-Squares Function

- We proceed by differentiating the sum of squares function with respect to the regression coefficients:

$$\frac{\partial S(A, B_1, B_2)}{\partial A} = \sum(-1)(2)(Y_i - A - B_1 X_{i1} - B_2 X_{i2})$$

$$\frac{\partial S(A, B_1, B_2)}{\partial B_1} = \sum(-X_{i1})(2)(Y_i - A - B_1 X_{i1} - B_2 X_{i2})$$

$$\frac{\partial S(A, B_1, B_2)}{\partial B_2} = \sum(-X_{i2})(2)(Y_i - A - B_1 X_{i1} - B_2 X_{i2})$$

- Normal equations for the coefficients are obtained by setting the partial derivatives to 0:

$$An + B_1 \sum X_{i1} + B_2 \sum X_{i2} = \sum Y_i$$

$$A \sum X_{i1} + B_1 \sum X_{i1}^2 + B_2 \sum X_{i1} X_{i2} = \sum X_{i1} Y_i$$

$$A \sum X_{i2} + B_1 \sum X_{i2} X_{i1} + B_2 \sum X_{i2}^2 = \sum X_{i2} Y_i$$

- The solution for the coefficients can be written out easily with the variables in mean-deviation form:

$$A = \overline{Y} - B_1\overline{X}_1 - B_2\overline{X}_2$$

$$B_1 = \frac{\sum X_1^* Y^* \sum X_2^{*2} - \sum X_2^* Y^* \sum X_1^* X_2^*}{\sum X_1^{*2} \sum X_2^{*2} - (\sum X_1^* X_2^*)^2}$$

$$B_2 = \frac{\sum X_2^* Y^* \sum X_1^{*2} - \sum X_1^* Y^* \sum X_1^* X_2^*}{\sum X_1^{*2} \sum X_2^{*2} - (\sum X_1^* X_2^*)^2}$$

- The coefficients are ***uniquely defined as long as the denominator is not equal to zero***—which occurs if one of the X's is invariant (as with simple regression), or if X_1 and X_2 are perfectly *collinear*. For a unique solution,

$$\sum X_1^{*2} \sum X_2^{*2} \neq \left(\sum X_1^* X_2^*\right)^2$$

Marginal versus ***Partial*** Relationships

- Coefficients in a simple regression represent ***marginal effects***
  - They do not control for other variables
- Coefficients in multiple regression represent ***partial effects***
  - Each slope is the effect of the corresponding variable *holding all other independent variables in the model constant*
  - In other words, the B_1 represents the effect of X_1 on Y, controlling for all other X variables in the model
  - Typically the marginal relationship of a given X is larger than the partial relationship after controlling for other important predictors

Matrix Form of Linear Models

- If we substitute α with β_0, the general linear model takes the following form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

- With the inclusion of a 1 for the constant, the regressors can be collected into a row vector, and thus the equation for ***each individual observation*** can be rewritten in *vector form*·

$$Y_i = [1, x_{i1}, x_{i2}, \ldots, x_{ik}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \varepsilon_i$$

$$= \underset{(1\times k+1)}{\boldsymbol{x}_i'} \quad \underset{(k+1\times 1)}{\boldsymbol{\beta}} \quad + \varepsilon_i$$

- Since *each observation has one such equation*, it is convenient to combine these equations in a single *matrix equation:*

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times k+1)}{\boldsymbol{X}} \underset{(k+1\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}$$

- $\boldsymbol{X}$ is Called the ***model matrix***, because it contains all the values of the explanatory variables for each observation in the data

- We assume that ε follows a multivariate-normal distribution with expectation E(ε)=0 and V(ε)=E(εε□)=$\sigma_{\varepsilon}{}^{2}\mathbf{I}_{n}$ That is, ε~$N_{\boldsymbol{n}}$(0, $\sigma_{\varepsilon}{}^{2}\mathbf{I}_{n}$).
- Since the ε are dependent on the conditional distribution of $\mathbf{y}$, $\mathbf{y}$ is also normally distributed with mean and variance as follows (note that this is the conditional value of Y!):

$$\begin{aligned}\boldsymbol{\mu} &\equiv E(\mathbf{y}) \\ &= E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta}\end{aligned}$$

$$\begin{aligned}V(\mathbf{y}) &= E\left[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'\right] \\ &= E\left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\right] \\ &= E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma_{\varepsilon}^{2}\mathbf{I}_{n}\end{aligned}$$

- Therefore, $\mathbf{y}$~$N_{\boldsymbol{n}}(\mathbf{X}\boldsymbol{\beta}, \sigma_{\varepsilon}{}^{2}\mathbf{I}_{n})$

OLS Fit in Matrix Form

- The fitted linear model is then

$$\mathbf{y} = \boldsymbol{X}\mathbf{b} + \mathbf{e}$$

  where **b** is the vector of fitted slope coefficients and **e** is the vector of residuals

- Expressed as a function of **b**, OLS finds the vector **b** that minimizes the ***residual sum of squares***:

$$\begin{aligned} S(\mathbf{b}) &= \sum E_i^2 = \mathbf{e}'\mathbf{e} \\ &= (\mathbf{y} - \boldsymbol{X}\mathbf{b})'(\mathbf{y} - \boldsymbol{X}\mathbf{b}) \\ &= \mathbf{y}'\mathbf{y} - \boldsymbol{X}\mathbf{b} - \mathbf{b}'\boldsymbol{X}'\mathbf{y} + \mathbf{b}'\boldsymbol{X}{\prime}\boldsymbol{X}\mathbf{b} \\ &= \mathbf{y}'\mathbf{y} - (2\mathbf{y}'\boldsymbol{X})\mathbf{b} + \mathbf{b}'(\boldsymbol{X}'\boldsymbol{X})\mathbf{b} \end{aligned}$$

$$S(\mathbf{b}) = \mathrm{y}'\mathrm{y} - (2\mathrm{y}'\boldsymbol{X})\mathbf{b} + \mathbf{b}'(\boldsymbol{X}'\boldsymbol{X})\mathbf{b}$$

We see that with respect to the $\mathbf{b}$ coefficient vector, there is a constant $(\mathrm{y}'\mathrm{y})$, a linear form in $\mathbf{b}$ and a quadratic form in $\mathbf{b}$. To minimize S($\mathbf{b}$), we find the partial derivative with respect to $\mathbf{b}$

$$\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = 0 - 2\boldsymbol{X}'\mathrm{y} + 2\boldsymbol{X}'\boldsymbol{X}\mathbf{b}$$

- The normal equations are found by setting this derivative to 0:

$$\boldsymbol{X}'\boldsymbol{X}\mathbf{b} = \boldsymbol{X}'\mathrm{y}$$

- If $\boldsymbol{X}$□$\boldsymbol{X}$ is nonsingular (rank of k+1) we can uniquely solve for the least-squares coefficients:

$$\mathbf{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{y}$$

Unique solution and the rank of $X'X$

- The rank of $X'X$ is equal to the rank of X. This attribute leads to two criteria that must be met in order to ensure $X'X$ is nonsingular, and thus obtain a unique solution:
  - Since the rank of X can be no larger than the smallest of n and $k+1$ to obtain a unique solution, we need *at least as many observations as there are coefficients in the model*
  - Moreover, the columns of X must not be linearly related—*i.e.*, the X-variables must be independent. Perfect collinearity prevents a unique solution, but even near collinearity can cause statistical problems.
  - Finally, no regressor other than the constant can be invariant—an invariate regressor would be a multiple of the constant.

Fitted Values and the Hat Matrix

- ***Fitted values*** are then obtained as follows:

$$\begin{aligned}\hat{\mathrm{y}} &= \boldsymbol{X}\mathbf{b} \\ &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{y} \\ &= \mathbf{H}\mathbf{y}\end{aligned}$$

- Where $\mathbf{H}$ is the ***Hat Matrix*** that projects the Y's onto their predicted values:

$$\underset{(n\times n)}{\mathrm{H}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

- Properties of the ***Hat Matrix:***
  - It depends solely on the predictor variable $\mathbf{X}$
  - It is square, symmetric and idempotent: $\mathbf{HH}=\mathbf{H}$
  - Finally, the trace of $\mathbf{H}$ is the degrees of freedom for the model

Distribution of the least-squares Estimator

- We now know that **b** is a linear estimator of β:

$$\mathbf{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{y} = \mathbf{M}\mathbf{y}$$

- Establishing the expectation of **b** from the expectation of $\mathbf{y}$**,** we see the **b** is an unbiased estimator of β:

$$E(\mathbf{b}) = E(\mathbf{M}) = \mathbf{M}E(\mathbf{y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta) = \beta$$

- Solving for the variance of **b,** we find that it depends only on the model matrix and the variance of the errors:

$$\begin{aligned} V(\mathbf{b}) &= \mathbf{M}V(\mathbf{y})\mathbf{M}' \\ &= [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\sigma_\varepsilon^2\mathbf{I}_n[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' \\ &= \sigma_\varepsilon^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma_\varepsilon^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned}$$

- Finally, if $\mathbf{y}$ is normally distributed, the distribution of **b** is:

$$\mathbf{b} \sim N_{k+1}[\beta, \sigma_\varepsilon^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]$$

Generality of the Least Squares Fit

- Least squares is desirable because of its simplicity
- The solution for the slope, $\mathbf{b}=(\mathbf{X}\square X)^{-1}X\mathbf{y}$ is expressed in terms of just 2 matrices and three basic operations:
  - *matrix transposition* (simply interchanging the elements in the rows and columns of a matrix)
  - *matrix multiplication* (sum of the products of each row and column combination of two conformable matrices)
  - *matrix inversion* (the matrix equivalent of a numeric reciprocal)
- The multiple regression model is also satisfying because of its generality. It has only two notable limitations:
  - It can be used to examine only a single dependent variable
  - It cannot provide a unique solution when the X's are not independent