

Regression III

Lecture 2: OLS

Dave Armstrong

University of Wisconsin – Milwaukee
Department of Political Science

e: armstrod@uwm.edu
w: www.quantoid.net/ICPSR.php

1 / 62

The Problem: Describing the Data

Ordinary Least Squares: Scalar Form

Ordinary Least Squares: Matrix Form

Properties of the OLS Estimator

Statistical Inference for OLS

Example: Duncan Data

Predictions and Inference

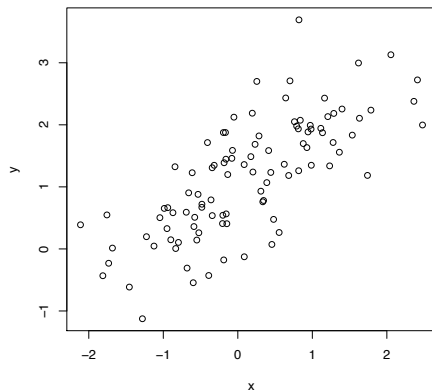
Model Fit

Model (Mis)Specification

2 / 62

Data

- We start with the problem of how to characterize the relationship between two variables.



3 / 62

Why do we use a line?

- It's accurate - linear modeling only makes sense if the dependence between the observed variables can be described by a line.
- It's easy to estimate - if it's accurate, then this type of modeling is easy - i.e., it has a closed-form, relatively intuitive solution.
- It's easy to describe - we can use a single number (the slope of the line) to describe the relationship between the variables.

4 / 62

How do we choose?

- There are many possible lines that “look” equally good
- Let’s try to fit some lines to these points
- Now, how do we adjudicate between these different lines
 - Rock, paper, scissors?
 - Arm wrestling?
 - ???

5 / 62

The Problem: Describing the Data

Ordinary Least Squares: Scalar Form

Ordinary Least Squares: Matrix Form

Properties of the OLS Estimator

Statistical Inference for OLS

Example: Duncan Data

Predictions and Inference

Model Fit

Model (Mis)Specification

6 / 62

Linearity

- If our goal is to fit a line that describes the data, it is assumed that the data can be described by a line.

$$Y_i = A + BX_i$$

- However, we know that not all of the points will fall on the line, there will be some distance between each point which can be expressed by E_i .

$$Y_i = A + BX_i + E_i \quad (1)$$

- This brings about the first (and, perhaps most important, assumption of OLS regression) - Linearity, which can be expressed by equation 1

7 / 62

Residuals

- We can re-express the residual (E_i) as a function of Y_i and X_i .

$$E_i = Y_i - A - BX_i$$

- A line that fits “best” can be thought of as one whose parameters (A and B) make E_i as small as possible.
- What are some options for “small”
 - Minimize $\sum_i E_i$
 - Minimize $\sum_i |E_i|$
 - Minimize $\sum_i E_i^2$

8 / 62

Least Squares

- We want to find A and B , such that we minimize the squares residuals:

$$S(A, B) = \sum_{i=1}^n E_i^2 = \sum (Y_i - A - BX_i)^2$$

- For any fixed $\{X_i, Y_i\}$, every different choice of regression coefficients A and B , produce a different residual sum of squares ($\sum_i E_i^2$).
- How, then, do we get from this equation, to a solution for the parameters of A and B ?

9 / 62

The Solution: Step 1

Take the partial first derivatives of $S(A, B)$ with respect to A and B .

$$\frac{\partial S(A, B)}{\partial A} = \sum_{i=1}^n (-1)(2)(Y_i - A - BX_i)$$

$$\frac{\partial S(A, B)}{\partial B} = \sum_{i=1}^n (-X_i)(2)(Y_i - A - BX_i)$$

Then, set them equal to zero and solve. This gives us:

$$nA + B \sum_i^n X_i = \sum_i^n Y_i$$

$$A \sum_i^n X_i + B \sum_i^n X_i^2 = \sum_i^n X_i Y_i$$

10 / 62

Solution for A

$$\sum_i^n Y_i = nA + B \sum_i^n X_i$$

$$nA = \sum_i^n Y_i - B \sum_i^n X_i$$

$$A = \frac{1}{n} \sum_i^n Y_i - B \frac{1}{n} \sum_i^n X_i$$

$$A = \bar{Y} - B\bar{X}$$

11 / 62

Solution for B

$$\sum_i^n X_i Y_i = A \sum_i^n X_i + B \sum_i^n X_i^2$$

$$= (\bar{Y} - B\bar{X}) \sum_i^n X_i + B \sum_i^n X_i^2$$

$$= \left(\frac{1}{n} \sum_i^n Y_i - B \frac{1}{n} \sum_i^n X_i \right) \sum_i^n X_i + B \sum_i^n X_i^2$$

$$= \frac{1}{n} \sum_i^n Y_i \sum_i^n X_i - B \frac{1}{n} \sum_i^n X_i \sum_i^n X_i + B \sum_i^n X_i^2$$

$$n \sum_i^n X_i Y_i = \sum_i^n Y_i \sum_i^n X_i - B \sum_i^n X_i \sum_i^n X_i + Bn \sum_i^n X_i^2$$

$$n \sum_i^n X_i Y_i - \sum_i^n Y_i \sum_i^n X_i = Bn \sum_i^n X_i^2 - B \sum_i^n X_i \sum_i^n X_i$$

$$= B \left(n \sum_i^n X_i^2 - \sum_i^n X_i \sum_i^n X_i \right)$$

$$B = \frac{n \sum_i^n X_i Y_i - \sum_i^n Y_i \sum_i^n X_i}{n \sum_i^n X_i^2 - \sum_i^n X_i \sum_i^n X_i}$$

$$= \frac{\sum_i^n X_i Y_i - \frac{1}{n} \sum_i^n Y_i \sum_i^n X_i}{\sum_i^n X_i^2 - \frac{1}{n} \sum_i^n X_i \sum_i^n X_i}$$

12 / 62

Re-expressing the numerator

$$\begin{aligned}
 \sum_i^n X_i Y_i - \frac{1}{n} \sum_i^n Y_i \sum_i^n X_i &= \sum_i^n X_i Y_i - \frac{1}{n} \sum_i^n Y_i \sum_i^n X_i - \underbrace{\frac{1}{n} \sum_i^n X_i \sum_i^n Y_i + \frac{1}{n} \sum_i^n Y_i \sum_i^n X_i}_0 \\
 &= \sum_i^n X_i Y_i - \bar{Y} \sum_i^n X_i - \bar{X} \sum_i^n Y_i + \bar{Y} n \bar{X} \\
 &= \sum_i^n X_i Y_i - \bar{Y} \sum_i^n X_i - \bar{X} \sum_i^n Y_i + \sum_i^n \bar{Y} \bar{X} \\
 &= \sum_i^n (X_i Y_i - X_i \bar{Y} - Y_i \bar{X} + \bar{X} \bar{Y}) \\
 &= \sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})
 \end{aligned}$$

13 / 62

Re-expressing the denominator

$$\begin{aligned}
 \sum_i^n X_i^2 - \frac{1}{n} \left(\sum_i^n X_i \right)^2 &= \sum_i^n X_i^2 - 2 \frac{1}{n} \left(\sum_i^n X_i \right)^2 + \frac{1}{n} \left(\sum_i^n X_i \right)^2 \\
 &= \sum_i^n X_i^2 - 2 \frac{1}{n} \sum_i^n X_i \sum_i^n X_i + \frac{n}{n^2} \left(\sum_i^n X_i \right)^2 \\
 &= \sum_i^n X_i^2 - 2 \frac{1}{n} \sum_i^n X_i \sum_i^n X_i + n \frac{\left(\sum_i^n X_i \right)^2}{n^2} \\
 &= \sum_i^n X_i^2 - 2 \frac{1}{n} \sum_i^n X_i \sum_i^n X_i + n \left(\frac{\sum_i^n X_i}{n} \right)^2 \\
 &= \sum_i^n X_i^2 - 2 \frac{1}{n} \sum_i^n X_i \sum_i^n X_i + n \bar{X}^2 \\
 &= \sum_i^n X_i^2 - 2 \frac{1}{n} \sum_i^n X_i \sum_i^n X_i + \sum_i^n \bar{X}^2 \\
 &= \sum_i^n (X_i^2 - 2 \bar{X} X_i + \bar{X}^2) \\
 &= \sum_i^n (X_i - \bar{X})^2
 \end{aligned}$$

14 / 62

Putting it back together

$$B = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i^n (X_i - \bar{X})^2} \quad (2)$$

More on what this tells us in a bit, but let's look at the matrix versions of this same thing.

15 / 62

The Problem: Describing the Data

Ordinary Least Squares: Scalar Form

Ordinary Least Squares: Matrix Form

Properties of the OLS Estimator

Statistical Inference for OLS

Example: Duncan Data

Predictions and Inference

Model Fit

Model (Mis)Specification

16 / 62

Matrix Form of Linear Models (1)

- If we substitute A for B_0 , the general linear model takes the form:

$$Y_i = B_0 + B_1X_{i1} + B_2X_{i2} + \dots + B_kX_{ik} + \varepsilon_i$$

- 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:

$$Y_i = [1, x_{i1}, x_{i2}, \dots, x_{ik}] \begin{bmatrix} B_0 \\ B_1 \\ B_2 \\ \vdots \\ B_k \end{bmatrix} + \varepsilon_i$$

$$= \underset{(1 \times k+1)}{\mathbf{x}'_i} + \underset{(k+1 \times 1)}{\mathbf{b}} + \varepsilon_i$$

17 / 62

Matrix Form of Linear Models(2)

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

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

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

- \mathbf{X} is called the model matrix, because it contains all the values of the explanatory variables for each observation in the data

18 / 62

OLS Fit in Matrix Form

- The fitted linear model is then:

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

where \mathbf{b} is the vector of fitted slope coefficients and \mathbf{e} is the vector of residuals.

- Expressed as a function of \mathbf{b} , OLS finds the vector \mathbf{b} that minimizes the residual sum of squares:

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

19 / 62

OLS Fit in Matrix Form (2)

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

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

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

- The normal equations are found by setting this derivative to zero.

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

- If $\mathbf{X}'\mathbf{X}$ is non-singular (of rank $k + 1$) we can uniquely solve for the least squares coefficients:

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

20 / 62

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:
 1. We need at least as many observations as there are coefficients in the model. Since the rank of X can be no larger than the smallest of n and $k + 1$ to obtain a unique solution.
 2. The columns of X must not be perfectly linearly related. Perfect collinearity prevents a unique solution, but even near collinearity can cause statistical problems. Moreover, no regressor other than the constant can be invariant - an invariant regressor would be a multiple of the constant.

21 / 62

Fitted Values and the Hat Matrix

- Fitted values are then obtained as follows:

$$\begin{aligned}\hat{y} &= Xb \\ &= X(X'X)^{-1}X'y \\ &= Hy\end{aligned}$$

- Where H is the *Hat Matrix* that projects the Y 's onto their predicted values:

$$H = X(X'X)^{-1}X'$$

- Properties of the Hat Matrix:
 1. It depends solely on the predictor variables X
 2. It is square, symmetric and idempotent: $HH = H$
 3. Finally, the trace of H is the degrees of freedom for the model

22 / 62

Where are we now?

- We've gone this far and what have we assumed?
 1. Linearity
 2. No perfect collinearity
- What have we not assumed yet?
 1. e independent from X .
 2. $e \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ - iid errors.
- We can find a unique solution to the problem we faced by only making the two assumptions mentioned above.
- What we don't know anything about is whether this tells us only something about a sample (in particular, this sample) or whether it tells us something about a larger set of observations from which this one is drawn.
- We turn to this other set of questions now.

23 / 62

The Problem: Describing the Data

Ordinary Least Squares: Scalar Form

Ordinary Least Squares: Matrix Form

Properties of the OLS Estimator

Statistical Inference for OLS

Example: Duncan Data

Predictions and Inference

Model Fit

Model (Mis)Specification

24 / 62

Reformulation of the Model

- When we consider the statistical properties of \mathbf{b} , it's as an estimator of something, specifically the population parameter vector $\boldsymbol{\beta}$.
- Here, we have to assume either that \mathbf{X} is fixed before the data collection, or more appropriately for this class, that \mathbf{X} is independent of the errors in the population $\boldsymbol{\varepsilon}$.
- Given this, we need to re-express our model as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (3)$$

- We can take the expectation of both sides:

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

- If we add the assumption $E(\boldsymbol{\varepsilon}) = 0$, then we get:

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$$

25 / 62

Unbiased-ness

- For an estimator $\tilde{\theta}$ of a population parameter θ to be unbiased, it must be the case that $E(\tilde{\theta}) = \theta$.
- We want to assess the bias in \mathbf{b} , so we need to see whether $E(\mathbf{b}) = \boldsymbol{\beta}$.

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\varepsilon} \\ &= \mathbf{I}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\varepsilon} \\ E(\mathbf{b}|\mathbf{X}) &= \boldsymbol{\beta} + E((\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\varepsilon}|\mathbf{X}) \\ &= \boldsymbol{\beta} + 0 \end{aligned}$$

26 / 62

BLUE

- To this point, we have made a set of assumptions that allowed us to show the \mathbf{b} as an estimator of $\boldsymbol{\beta}$ is both linear and unbiased.
- This is great, but it doesn't get us out of the problem with which we started class - namely, how do we know this is the "best" line?
- To do this, we need to add another concept to our arsenal - efficiency.
- If the OLS line is "best", it should be more efficient (i.e., have smaller variance) than any other linear, unbiased estimate.
- How do we get there?

27 / 62

Variance of the errors

- To get to the point where we can make statements about OLS being the "best", we need to make another assumption about the errors.
- Homoskedasticity: $V(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2 \mathbf{I}_n$ or, the variance of the errors conditional on \mathbf{X} is the same.
- Put another way, the variance of the residuals is $\sigma^2 \mathbf{I}_n$.
- We know that \mathbf{y} is different from $\mathbf{X}\boldsymbol{\beta}$ because of $\boldsymbol{\varepsilon}$. Thus, the distribution of \mathbf{y} (i.e., the spread of points of \mathbf{y} around its expectation, $\mathbf{X}\boldsymbol{\beta}$, will have the same variance as $\boldsymbol{\varepsilon}$, or $\sigma^2 \mathbf{I}_n$.
- Now, we can see what this does for us.

28 / 62

Variance of Linear Estimators

Let's consider the linear estimator $b_0 = Cy$ where C is a $K \times n$ matrix. If b_0 is unbiased, then $E(Cy|X) = E((CX\beta + C\varepsilon)|X) = \beta$. Thus, $CX = I$.

We also know that $\text{var}(b_0|X) = \sigma^2 CC'$. If we define $D = C - ((X'X)^{-1}X')$, then:

$$\begin{aligned}\text{Var}(b_0|X) &= \sigma^2 [(D + ((X'X)^{-1}X')) (D + ((X'X)^{-1}X'))'] \\ &= \sigma^2 [DD' + DX(X'X)^{-1} + (X'X)^{-1}X'D' + (X'X)^{-1}]\end{aligned}$$

We also know that $CX = I$ and the $CX = (D + ((X'X)^{-1}X'))X$, so

$$\begin{aligned}I &= (D + ((X'X)^{-1}X'))X \\ &= DX + (X'X)^{-1}X'X \\ &= DX + I\end{aligned}$$

Thus, $DX = 0$.

29 / 62

Variance of Linear Estimators (2)

If $DX = 0$, this implies that

$$\begin{aligned}\text{Var}(b_0|X) &= \sigma^2 (X'X)^{-1} + \sigma^2 DD' \\ &= \text{Var}(b|X) + \sigma^2 DD'\end{aligned}$$

Since DD' is a quadratic in D , it must be non-negative definite (i.e., greater than or equal to zero). Thus, $\text{Var}(b_0|X) \geq \text{Var}(b|X)$.

30 / 62

Recap

Now we know that the OLS estimator b is linear, unbiased, and efficient. What assumptions did we have to make along the way:

- Linearity
 - $y = X\beta + \varepsilon$.
 - No perfect collinearity (or X of full-rank).
- Unbiasedness
 - ε independent from X
 - $E(\varepsilon) = 0$
- Efficiency
 - Homoskedasticity: $V(\varepsilon|X) = \sigma^2$.
 - $V(\varepsilon|X) = \sigma^2 I_n$

What we have not done yet is talk about inference. To do this, we need to make one more assumption:

$$\varepsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_n)$$

31 / 62

The Problem: Describing the Data

Ordinary Least Squares: Scalar Form

Ordinary Least Squares: Matrix Form

Properties of the OLS Estimator

Statistical Inference for OLS

Example: Duncan Data

Predictions and Inference

Model Fit

Model (Mis)Specification

32 / 62

Statistical Inference for OLS

If we know that $\boldsymbol{\varepsilon}$ is distributed normally, then that implies \mathbf{b} is also normally distributed. Specifically:

$$\mathbf{b} \sim \mathcal{N}_{k+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

This is called the *sampling distribution* of \mathbf{b} . $\boldsymbol{\beta}$ is fixed in the population, it is a constant (scalar). However, because we have a “random” sample, \mathbf{b} will differ in each sample according to the distribution above.

Notice, $\boldsymbol{\beta}$ and σ are population quantities, so this is a theoretical distribution.

33 / 62

Inference for Individual Coefficients (1)

- Any individual coefficient B_j is distributed normally with expectation β_j and sampling variance $\sigma^2 v_{jj}$, where v_{jj} is the j^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.
- So, we can test the hypothesis $H_0 : \beta_j = \beta_j^{(0)}$ with:

$$Z_0 = \frac{B_j - \beta_j^{(0)}}{\sigma \sqrt{v_{jj}}}$$

This, however, doesn't help much since β and σ are unknown.

- $S_E^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k-1}$ is an unbiased estimator of σ^2 , so

$$\hat{V}(\mathbf{b}) = \frac{\mathbf{e}'\mathbf{e}}{n-k-1} (\mathbf{X}'\mathbf{X})^{-1}; \quad \text{SE}(B_j) = \sqrt{\hat{V}(\mathbf{b})}$$

34 / 62

Inference for Individual Coefficients (2)

Because B_j and S_E^2 are independent, their ratio is distributed t with $n - k - 1$ degrees of freedom:

$$t_0 = \frac{B_j - \beta_j^{(0)}}{\text{SE}(B_j)}$$

Thus:

$$\beta_j \in B_j \pm t_{95, n-k-1} \text{SE}(B_j)$$

35 / 62

Inference for Multiple Coefficients: F -test

- Assume we have an OLS model with k explanatory variables that produces residual sum of squares RSS for the *full* model.
- Now, place q linear restrictions on the model coefficients (e.g., set some of them to zero) and generate a new residual sum of squares RSS_0 for the *restricted* model.

$$F_0 = \frac{\frac{RSS_0 - RSS}{q}}{\frac{RSS}{n-k-1}}$$

- The statistic F_0 is distributed F with q and $n - k - 1$ degrees of freedom.

36 / 62

Incremental F -test Without Estimating Both Models

If \mathbf{b} is the coefficient vector for the *full* model, the incremental F -test can be accomplished by:

1. Take the q elements you want to restrict from \mathbf{b} such that

$$\mathbf{b}_1 = \begin{bmatrix} B_1 \\ \vdots \\ B_q \end{bmatrix}$$

2. Take the q rows and columns of $(\mathbf{X}'\mathbf{X})^{-1}$ such that:

$$\mathbf{V}_{11} = \begin{bmatrix} v_{1,1} & \cdots & v_{1,q} \\ \vdots & \ddots & \vdots \\ v_{q,1} & \cdots & v_{q,q} \end{bmatrix}$$

3. Then:

$$F_0 = \frac{(\mathbf{b}_1 - \boldsymbol{\beta}_1^0)' \mathbf{V}_{11}^{-1} (\mathbf{b}_1 - \boldsymbol{\beta}_1^0)}{qS_E^2}; \quad F_0 \sim F_{q, n-k-1}$$

37 / 62

General Linear Hypotheses

- Imagine, now, that we want to test $H_0 : \beta_1 = \beta_2$ or $H_0 : \beta_3 = \beta_4 = 0$.
- More generally, we might want to test:

$$H_0 : \underset{(q \times k+1)}{\mathbf{L}} \underset{(k+1 \times 1)}{\boldsymbol{\beta}} = \underset{(q \times 1)}{\mathbf{c}}$$

Here, \mathbf{L} and \mathbf{c} contain pre-specified constants.

- \mathbf{L} , the hypothesis matrix, is of full row rank (i.e., $q \leq k + 1$)
- The resulting F -statistic is:

$$F_0 = \frac{(\mathbf{Lb} - \mathbf{c})' [\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1} (\mathbf{Lb} - \mathbf{c})}{qS_E^2}$$

38 / 62

General Linear Hypotheses (2)

Suppose we have the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

If we wanted to test $H_0 : \beta_1 = \beta_2 = 0$, we would set

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

To test $H_0 : \beta_1 = \beta_2$ (equivalent to $H_0 : \beta_1 - \beta_2 = 0$ you would do:

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 0 \end{bmatrix}$$

39 / 62

More General Linear Hypothesis Examples

To test $H_0 : \beta_1 + \beta_2 = 1$, set:

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}, \quad \mathbf{c} = [1]$$

To test $H_0 : \beta_1 = \beta_2 = \beta_3$ or equivalently $H_0 : \beta_1 = \beta_2$ & $\beta_2 = \beta_3$, set

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Why wouldn't we want to test $H_0 : \beta_1 = \beta_2$ & $\beta_2 = \beta_3$ & $\beta_1 = \beta_3$?

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 1 & 0 & -1 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Here, \mathbf{L} is not of full row rank as row 1 + row 2 = row 3, so the third constraint is implied by the first two.

40 / 62

The Problem: Describing the Data

Ordinary Least Squares: Scalar Form

Ordinary Least Squares: Matrix Form

Properties of the OLS Estimator

Statistical Inference for OLS

Example: Duncan Data

Predictions and Inference

Model Fit

Model (Mis)Specification

41 / 62

Example: Duncan Data

```
> library(car)
> data(Duncan)
> mod <- lm(prestige ~ income + education, data = Duncan)
> summary(mod)

Call:
lm(formula = prestige ~ income + education, data = Duncan)

Residuals:
    Min       1Q   Median       3Q      Max
-29.5380  -6.4174   0.6546   6.6051  34.6412

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.06466     4.27194  -1.420   0.163
income       0.59873     0.11967   5.003 1.05e-05 ***
education    0.54583     0.09825   5.555 1.73e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.37 on 42 degrees of freedom
Multiple R-squared:  0.8282, Adjusted R-squared:  0.82
F-statistic: 101.2 on 2 and 42 DF,  p-value: < 2.2e-16
```

44 / 62

Regression "by hand"

```
> X <- with(Duncan, cbind(1, income, education))
> y <- matrix(Duncan[["prestige"]], ncol = 1)
> b <- solve(t(X) %*% X) %*% t(X) %*% y
> b

      [,1]
income -6.0646629
education 0.5987328
education 0.5458339

> coef(mod)

(Intercept)      income      education
-6.0646629    0.5987328    0.5458339

> e <- matrix(y - X %*% b, ncol = 1)
> Vb <- c((t(e) %*% e)/(nrow(X) - 2 - 1)) * solve(t(X) %*% X)
> Vb

      income      education
income  18.2494814 -0.15184501 -0.150706025
education -0.1518450 0.01432027 -0.008518551
education -0.1507060 -0.00851855 0.009653582

> vcov(mod)

(Intercept)      income      education
(Intercept)  18.2494814 -0.15184501 -0.150706025
income      -0.1518450 0.01432027 -0.008518551
education   -0.1507060 -0.00851855 0.009653582
```

43 / 62

F-test Example

```
> restricted.mod <- lm(prestige ~ 1, data = Duncan)
> anova(restricted.mod, mod, test = "F")

Analysis of Variance Table

Model 1: prestige ~ 1
Model 2: prestige ~ income + education
  Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1      44 43688
2      42  7507  2    36181 101.22 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> xtx <- solve(t(X) %*% X)
> s2e <- (t(e) %*% e)/(nrow(X) - 3)
> F0 <- (t(b[2:3]) %*% solve(xtx[2:3, 2:3]) %*% b[2:3])/(2 * s2e)
> F0

      [,1]
[1,] 101.2162

> pf(F0, 2, nrow(X) - 3, lower.tail = F)

      [,1]
[1,] 8.647636e-17
```

44 / 62

General Linear Hypothesis Example

Test $H_0 : \beta_1 = \beta_2 = 0$:

```
> L <- matrix(c(0, 0, 1, 0, 0, 1), nrow = 2)
> cmat <- matrix(c(0, 0), ncol = 1)
> FOb <- (t(L) %*% b - cmat) %*% solve(L %*% solve(t(X) %*% X) %*%
+ t(L)) %*% (L %*% b - cmat)/(2 * s2e)
> FOb
      [,1]
[1,] 101.2162
> pf(FOb, 2, nrow(X) - 3, lower.tail = F)
      [,1]
[1,] 8.647636e-17
```

45 / 62

The Problem: Describing the Data

Ordinary Least Squares: Scalar Form

Ordinary Least Squares: Matrix Form

Properties of the OLS Estimator

Statistical Inference for OLS

Example: Duncan Data

Predictions and Inference

Model Fit

Model (Mis)Specification

46 / 62

Inference for Predictions

- Often times, we want to know what the prediction is for a certain set of \mathbf{X} values, say \mathbf{x}_0 , where:

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ X_{02} \\ X_{03} \\ \vdots \\ X_{0k} \end{bmatrix}$$

- The prediction can be generated as follows:

$$\hat{Y} = \mathbf{x}'_0 \mathbf{b}$$

47 / 62

Variance of Predictions

- The variance of a mean prediction at \mathbf{x}_0 is:

$$\text{var}(\hat{Y}_0|\mathbf{x}_0) = S_E^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0$$

- The variance of an individual prediction at \mathbf{x}_0 is:

$$\text{var}(\hat{Y}_0|\mathbf{x}_0) = S_E^2 [1 + \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0]$$

- What if we want the variances of three different individual predictions for the Duncan model from above:

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 21 \\ 26 \end{bmatrix}; \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 64 \\ 84 \end{bmatrix}; \quad \mathbf{x}_3 = \begin{bmatrix} 1 \\ 41.87 \\ 52.56 \end{bmatrix}$$

48 / 62

Example: Variance of Predictions

```
> x1 <- matrix(c(1, 21, 26), ncol = 1)
> x2 <- matrix(c(1, 64, 84), ncol = 1)
> x3 <- matrix(c(1, 41.87, 52.56))
> pred1 <- t(x1) %*% b
> pred1
      [,1]
[1,] 20.70041
> pred2 <- t(x2) %*% b
> pred2
      [,1]
[1,] 78.10429
> pred3 <- t(x3) %*% b
> pred3
      [,1]
[1,] 47.69331
> v1 <- s2e * (1 + c(t(x1) %*% solve(t(X) %*% X) %*% x1))
> sqrt(v1)
      [,1]
[1,] 13.64936
> v2 <- s2e * (1 + c(t(x2) %*% solve(t(X) %*% X) %*% x2))
> sqrt(v2)
      [,1]
[1,] 13.68962
> v3 <- s2e * (1 + c(t(x3) %*% solve(t(X) %*% X) %*% x3))
> sqrt(v3)
      [,1]
[1,] 13.51676
```

49 / 62

Example: Variance of Predictions (2)

```
> vhat1 <- s2e * (c(t(x1) %*% solve(t(X) %*% X) %*% x1))
> sqrt(vhat1)
      [,1]
[1,] 2.752105
> vhat2 <- s2e * (c(t(x2) %*% solve(t(X) %*% X) %*% x2))
> sqrt(vhat2)
      [,1]
[1,] 2.945296
> vhat3 <- s2e * (c(t(x3) %*% solve(t(X) %*% X) %*% x3))
> sqrt(vhat3)
      [,1]
[1,] 1.992937
> newdat <- t(cbind(x1, x2, x3))
> colnames(newdat) <- c("int", "income", "education")
> newdat <- as.data.frame(newdat)
> predict(mod, newdat, se.fit = T)

$fit
      1      2      3
20.70041 78.10429 47.69331

$se.fit
      1      2      3
2.752105 2.945296 1.992937

$df
[1] 42

$residual.scale
[1] 13.36903
```

50 / 62

The Problem: Describing the Data

Ordinary Least Squares: Scalar Form

Ordinary Least Squares: Matrix Form

Properties of the OLS Estimator

Statistical Inference for OLS

Example: Duncan Data

Predictions and Inference

Model Fit

Model (Mis)Specification

51 / 62

Model Fit: R^2 or RMSE

- We have thus far figured out just about everything except how well the model fits.
- There are a couple of different measures readily available for this (e.g., without going to the maximum likelihood framework).
- The R^2 and the Standard Error of the Residuals (aka RMSE)

52 / 62

Definitions

The R^2 is the square of the correlation between the predicted values (\hat{y}_i) and the observed y_i values

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} = \frac{\text{RegSS}}{\text{TSS}}$$

Since this correlation cannot go down when variables are added, an adjustment must be made to the R^2 as follows:

$$\tilde{R}^2 = 1 - \frac{\frac{\text{RSS}}{n-k-1}}{\frac{\text{TSS}}{n-1}}$$

The Standard Error of the Residuals is as you would suspect.

$$\text{SER} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-k}}$$

Notice, these are on two different scales - the SER is in units of the dependent variable. The R^2 is on the scale of the coefficients, so $R^2 \in [0, 1]$.

53 / 62

The Problem: Describing the Data

Ordinary Least Squares: Scalar Form

Ordinary Least Squares: Matrix Form

Properties of the OLS Estimator

Statistical Inference for OLS

Example: Duncan Data

Predictions and Inference

Model Fit

Model (Mis)Specification

54 / 62

Model (Mis)Specification

- We'll talk in more depth about model selection (e.g., selecting which variables belong in the model) later on in the course.
- Finally, for today, we'll address the *potential* problem of omitted variable bias.
- First we must decide whether our model is one of empirical relationship or causal relationship. While we all probably want to make causal claims, there are some definite problems that come along with that.
 - A model of empirical relationship simply traces the empirical covariance of one variable with a set of other variables.
 - A causal model proposes that we have been able to ascertain and include all of the relevant causal determinants of y and as such, we can interpret coefficients as the *effect of x on y* .

55 / 62

Misspecification

Suppose we're interested in the population model:

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1^*\boldsymbol{\beta}_1 + \mathbf{X}_2^*\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

where variables with a * superscript are mean deviated versions of the original variables. \mathbf{X}_1^* and \mathbf{X}_2^* are matrices of regressors.

Now, define $\mathbf{X}_2^* + \boldsymbol{\varepsilon} \equiv \tilde{\boldsymbol{\varepsilon}}$ such that

$$\mathbf{y}^* = \mathbf{X}_1^*\boldsymbol{\beta}_1 + \tilde{\boldsymbol{\varepsilon}}$$

Now we can see what happens to \mathbf{b}_1 .

56 / 62

b_1 and Mis-specification

$$\begin{aligned} b_1 &= (X_1^{*'} X_1^*)^{-1} X_1^{*'} y^* \\ &= \left(\frac{1}{n} X_1^{*'} X_1^* \right)^{-1} \frac{1}{n} X_1^{*'} y^* \\ &= \left(\frac{1}{n} X_1^{*'} X_1^* \right)^{-1} \frac{1}{n} X_1^{*'} (X_1^* \beta_1 + X_2^* \beta_2 + \varepsilon) \\ &= \beta_1 + \left(\frac{1}{n} X_1^{*'} X_1^* \right)^{-1} \frac{1}{n} X_1^{*'} X_2^* \beta_2 + \left(\frac{1}{n} X_1^{*'} X_1^* \right)^{-1} \frac{1}{n} X_1^{*'} \varepsilon \end{aligned}$$

Taking probability limits produces:

$$\begin{aligned} \text{plim } b_1 &= \beta_1 + \Sigma_{11}^{-1} \Sigma_{12} \beta_2 + \Sigma_{11}^{-1} \sigma_{1\varepsilon} \\ &= \beta_1 + \Sigma_{11}^{-1} \Sigma_{12} \beta_2 \end{aligned}$$

Where: $\Sigma_{11} \equiv \text{plim} \left(\frac{1}{n} \right) X_1^{*'} X_1^*$, $\Sigma_{12} \equiv \text{plim} \left(\frac{1}{n} \right) X_1^{*'} X_2^*$, and $\sigma_{1\varepsilon} \equiv \text{plim} \left(\frac{1}{n} \right) X_1^{*'} \varepsilon = 0$ (by assumption).

57 / 62

Mis-specification Bias

We assumed $\sigma_{1\varepsilon} = 0$, but what about $\sigma_{1\tilde{\varepsilon}}$?

$$\begin{aligned} \text{plim} \frac{1}{n} X_1^{*'} \tilde{\varepsilon} &= \text{plim} \frac{1}{n} X_1^{*'} (X_2^* \beta_2 + \varepsilon) \\ &= \Sigma_{12} \beta_2 + \sigma_{1\varepsilon} \end{aligned}$$

So, $\sigma_{1\tilde{\varepsilon}}$ can only be 0 if Σ_{12} (the correlation between X_1^* and X_2^* is 0 or β_2 the effect of X_2^* in the population is 0.

This is the classic *omitted variable bias*.

58 / 62

Omitted Variable Bias: The Phantom Menace

- We have the idea that the bias in our coefficients is monotonically decreasing (and approaching zero) as the proportion of relevant controls in our model increases.
- That is to say, if in the true data generating process (DGP), there are 100 variables - a model that includes 75 of them is *better* (coefficients have smaller bias) than a model that includes only 50 of them.
- Clarke (2005) shows that this is not necessarily the case.
 - Adding a subset of controls does not necessarily make the model *better*.
 - It could actually make the model *worse*.
- His recommendations:
 1. Focus on research design and look for natural experiments.
 2. Test theories on smaller, narrower domains (e.g., spatially or temporally narrower).

59 / 62

The Problem: Describing the Data

Ordinary Least Squares: Scalar Form

Ordinary Least Squares: Matrix Form

Properties of the OLS Estimator

Statistical Inference for OLS

Example: Duncan Data

Predictions and Inference

Model Fit

Model (Mis)Specification

60 / 62

Readings

Today:

*Clarke (2005)

*Gill (1999)

*Fox (2008) Chapters 5, 6 & 9

*Fox (2002) Chapter 4

-Gelman and Stern (2006) -Abbott (1998, 1988), Achen (1990),

Lewis-Beck and Skalaban (1990)

Tomorrow:

*Fox (2002) Chapter 7

*Jacoby (1997, 1998, 2006)

*Murrell (2006) Chapters 1-4

-Kastellec and Leoni (2007)

-Sarkar (2008), Venables and Ripley (2002) Chapter 4

Abbott, Andrew. 1988. "Transcending General Linear Reality." *Sociological Theory* 6(2):169-186.

Abbott, Andrew. 1998. "The Causal Devolution." *Sociological Methods and Research* 27:148-181.

Achen, Christopher H. 1990. "What Does "Explained Variance" Explain?: Reply." *Political Analysis* 2(1):173-184.

Clarke, Kevin. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 22(4):341-352.

Fox, John. 2002. *An R and S-Plus Companion to Applied Regression*. Thousand Oaks: Sage Publications.

Fox, John. 2008. *Applied Regression Analysis and Generalized Linear Models, 2nd edition*. Thousand Oaks, CA: Sage, Inc.

Gelman, Andrew and Hal Stern. 2006. "The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant." *The American Statistician* 60(4):328-331.

Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(3):647-674.

Jacoby, William G. 1997. *Statistical Graphics for Univariate and Bivariate Data*. Thousand Oaks, CA: Sage.

Jacoby, William G. 1998. *Statistical Graphics for Visualizing Multivariate Data*. Thousand Oaks, CA: Sage.

Jacoby, William G. 2006. "The Dot Plot: A Graphical Display for Labeled QUantitative Values." *The Political Methodologist* 14(1):6-14.

Kastellec, Jonathan P and Eduardo L. Leoni. 2007. "Using Graphs Instead of Tables in Political Science." *Perspectives on Politics* 5(4):755-771.

Lewis-Beck, Michael S. and Andrew Skalaban. 1990. "The R-squared: Some Straight Talk." *Political Analysis* 2(1):153-171.

Murrell, Paul. 2006. *R Graphics*. Boca Raton, FL: Chapman & Hall/CRC.

Sarkar, Deepayan. 2008. *Lattice: Multivariate Data Visualization with R*. New York: Springer Science + Business Media, LLC.

Venables, William N. and Brian D. Ripley. 2002. *Modern Applied Statistics with S, 4th edition*. 3 ed. New York: Springer.