

# PS C236A/ Stat C239A

## Regression

Erin Hartman

September 1, 2009

### 1 OLS in Matrix Form

In algebraic notation, the regression model for unit  $i$  can be written as:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

Therefore, with  $n$  observations, all of the observations can be written as a system of equations:

$$\begin{aligned} y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1p} + \epsilon_1 \\ y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2p} + \epsilon_2 \\ &\vdots \\ y_n &= \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{np} + \epsilon_n \end{aligned}$$

The above system of equations can be expressed in matrix form as:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

- On the left hand side of the equation, we have  $\mathbf{Y}$  which is an  $n \times 1$  vector of *observable* random variables, also known as the *dependent* or *response* variable. Each unit of observation corresponds to a  $Y_i$
- On the right hand side is  $\mathbf{X}$ , an  $n \times p$  matrix of *observable* random variables.  $\mathbf{X}$  is also called the *design matrix*. Each column of  $\mathbf{X}$  is a variable, also called *explanatory*, *dependent* variables or *covariates*. Each row is a unit of observation
- Next to  $\mathbf{X}$  is  $\beta$ , a  $p \times 1$  vector of *parameters*. These parameters are typically unknown and must be estimated from the data

- Also on the right hand side is  $\epsilon$ , an  $n \times 1$  vector of *unobservable* random variables.  $\epsilon$  is also referred to as the *random error* or *disturbance* term.
- Note that if we wish to include an intercept, then we simply make the first column of  $X$  be a  $n \times 1$  vector of 1s

Therefore, we can write the OLS regression equation as such:

$$Y = X\beta + \epsilon$$

## 1.1 Deriving $\hat{\beta}$

We know:

$$e = Y - X\beta$$

Then the sum of squared residuals can be defined as:

$$\begin{aligned} e^T e &= \sum_{i=1}^n e_i^2 \\ &= (Y - X\beta)^T (Y - X\beta) \\ &= Y^T Y - 2Y^T X\beta + \beta X^T X\beta \end{aligned}$$

Making the first order conditions for  $\hat{\beta}$ :

$$\begin{aligned} \delta(e^T e) / \delta\beta &= 0 \\ \Rightarrow -2X^T Y + 2X^T X\hat{\beta} &= 0 \\ \Rightarrow \hat{\beta} &= (X^T X)^{-1} X^T Y \end{aligned}$$

so long as the matrix  $X^T X$  is invertible.

## 1.2 Deriving $\hat{\sigma}^2$

Recall:

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T Y \\ &= (X^T X)^{-1} X^T (X\beta + \epsilon) \\ \Rightarrow \hat{\beta} - \beta &= (X^T X)^{-1} X^T \epsilon \end{aligned}$$

Plugging this into the covariance equation:

$$\begin{aligned}
\text{cov}(\hat{\beta}|X) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X] \\
&= E\left[\left((X^T X)^{-1} X^T \epsilon\right)\left((X^T X)^{-1} X^T \epsilon\right)'|X\right] \\
&= E\left[(X^T X)^{-1} X^T \epsilon \epsilon^T X (X^T X)^{-1}|X\right] \\
&= (X^T X)^{-1} X^T E(\epsilon \epsilon^T | X) X (X^T X)^{-1} \\
&\quad \text{where } E(\epsilon \epsilon^T | X) = \sigma^2 I_{p \times p} \\
&= (X^T X)^{-1} X^T \sigma^2 I_{p \times p} X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1}
\end{aligned}$$

We estimate  $\sigma^2$  dividing the residuals squared by the degrees of freedom because the  $e_i$  are generally smaller than the  $\epsilon_i$  due to the fact that  $\hat{\beta}$  was chosen to make the sum of square residuals as small as possible.

$$\hat{\sigma}^2 = \frac{1}{n - p} \sum_{i=1}^n e_i^2$$

### 1.3 The Hat Matrix

The *hat matrix*, or *projection matrix*

$$H = X(X^T X)^{-1} X^T \text{ with } \tilde{H} = I - H$$

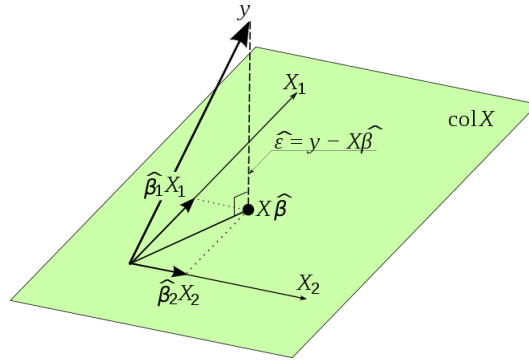
where  $HY$  is the part of  $Y$  that projects into  $X$ . We use the hat matrix to find the fitted values,  $\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY$

We can now write

$$\begin{aligned}
e &= (I - H)Y = (I - H)(X\beta + \epsilon) = (I - H)\epsilon \\
\|e\|^2 &= e^T e = \epsilon^T \tilde{H}^2 \epsilon = \epsilon^T \tilde{H} \epsilon
\end{aligned}$$

If  $HY$  yields part of  $Y$  that projects into  $X$ , this means that  $\tilde{H}Y$  is the part of  $Y$  that does not project into  $X$ , which is the *residual* part of  $Y$ . Therefore,  $\tilde{H}Y$  makes the residuals.

It is important to note that this is just the mechanics of OLS, and all that is required in order for  $Y$  to be split into the part that projects into  $X$  and the residual part of  $Y$  is that  $X$  has full rank. In order to derive meaning from this mechanical procedure, we require a set of assumptions, which are laid out in the following section.



## 2 The OLS assumptions

1. *Linear in Parameters*:  $Y$  is related to the independent variables and the error term as  $Y = X\beta + \epsilon$
2. The  $X$ 's are fixed at take on  $\geq 2$  values
3. *Full Rank* (in multiple regression): There is no perfect collinearity among any of the independent variables
4. *Zero Conditional Mean*:  $E(\epsilon|X) = 0$
5. *Homoskedasticity*:  $Var(\epsilon|X) = \sigma^2$
6. *Random Sampling*:  $Y_i$  is an *iid* random sample, although this can be relaxed to  $cov(y_i, y_j) = 0 = cov(\epsilon_i, \epsilon_j) \quad i \neq j$
7. *Normal Errors* (optional):  $Y \sim N(X\beta, \sigma^2)$

## 3 Gauss-Markov Theorem

**Gauss-Markov Theorem:** Under assumptions 1-6 above,  $\hat{\beta}$  is the best linear unbiased estimator (BLUE) of  $\beta$

### 3.1 What does BLUE mean?

**B** est is defined as having the *smallest variance* among the class of linear unbiased estimators

**L** inear means that the estimator  $\tilde{\beta}$  of  $\beta$  is linear  $\iff$  it can be expressed as a linear function of the data on the dependent variable

**U** nbiased means that  $E(\tilde{\beta}) = \beta$

**E** stimator is a rule that can be applied to any sample of data to produce an estimate

## 3.2 The “U”

### 3.2.1 Unbiasedness of $\hat{\beta}$

Recall:

$$\hat{\beta} = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + \epsilon) = (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T \epsilon = \beta + (X^T X)^{-1} X^T \epsilon$$

We know that  $\hat{\beta}$  is unbiased if  $E(\hat{\beta}) = \beta$

$$\begin{aligned} E(\hat{\beta}) &= E(\beta + (X^T X)^{-1} X^T \epsilon | X) \\ &= E(\beta | X) + E((X^T X)^{-1} X^T \epsilon | X) \\ &= \beta + (X^T X)^{-1} E(\epsilon | X) \end{aligned}$$

$$\text{where } E(\epsilon | X) = E(\epsilon) = 0$$

$$E(\hat{\beta}) = \beta$$

### 3.2.2 Unbiasedness of $\hat{\sigma}^2$

Now, note that:

$$\begin{aligned} E(e^T e | X) &= E(\epsilon^T \tilde{H} \epsilon | X) \\ &= E(\sum_i \sum_j \epsilon_i \tilde{H}_{ij} \epsilon_j | X) \\ &= \sum_i \sum_j E(\epsilon_i \tilde{H}_{ij} \epsilon_j | X) \\ &\quad \text{if } i \neq j, E(\epsilon_i \tilde{H}_{ij} \epsilon_j | X) = 0 \\ &\quad \text{if } i = j, E(\epsilon_i \tilde{H}_{ij} \epsilon_j | X) = \sigma^2 \tilde{H}_{ii} \\ &= \sigma^2 \text{Tr}(\tilde{H}) \\ &= \sigma^2(n - p) \end{aligned}$$

We know that  $\hat{\sigma}^2$  is unbiased if  $E(\hat{\sigma}^2|X) = \sigma^2$

$$\begin{aligned}
 E(\hat{\sigma}^2|X) &= E\left(\frac{1}{n-p}e^T e|X\right) \\
 &= \frac{1}{n-p}E(e^T e|X) \\
 &= \frac{1}{n-p}\sigma^2(n-p) \\
 E(\hat{\sigma}^2|X) &= \sigma^2
 \end{aligned}$$

### 3.3 The “B”

We know we want an estimator of the form  $\tilde{\beta} = m + MY$  with  $E(\tilde{\beta}|X) = \beta$

$$E(\tilde{\beta}|X) = E(m + MY|X) = E(m + M(X\beta + \epsilon)|X) = m + MX\beta$$

$$\Rightarrow m = 0 \text{ and } MX = I_{p \times p} \quad (1)$$

Therefore, this implies we want  $\tilde{\beta} = MY$ , so WLOG we can say  $M = (X^T X)^{-1} X^T + c$

Thus,

$$MX = ((X^T X)^{-1} X^T + c)X = X^T X)^{-1} X^T X + cX = I_{p \times p} + CX = I_{p \times p} \text{ by (1)}$$

$$\Rightarrow CX = 0 \quad (2)$$

Also note that

$$\tilde{\beta} = MY = M(X\beta + \epsilon) = \beta + M\epsilon \text{ by } MX = I_{p \times p}$$

$$\Rightarrow \tilde{\beta} - \beta = M\epsilon \quad (3)$$

Now, as noted in section 3.1, “best” means having the smallest variance, therefore we want to minimize  $cov(\tilde{\beta}|X)$

$$\begin{aligned}
cov(\tilde{\beta}|X) &= E((\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^T|X) \\
&= E((M\epsilon)(M\epsilon)^T|X) \\
&= E(M\epsilon\epsilon^T M^T|X) \\
&= ME(\epsilon\epsilon^T|X)M^T \\
&= \sigma^2 MM^T
\end{aligned}$$

where,

$$\begin{aligned}
MM^T &= ((X^T X)^{-1} X^T + c)((X^T X)^{-1} X^T + c)^T \\
&= (X^T X)^{-1} X^T X (X^T X)^{-1} + (X^T X)^{-1} X^T C^T + CX (X^T X)^{-1} + CC^T \\
&\quad \text{where, from (2), } CX = 0 \text{ and } C^T X^T = 0 \\
MM^T &= (X^T X)^{-1} + CC^T
\end{aligned}$$

Thus,

$$cov(\tilde{\beta}|X) = \sigma^2 (X^T X)^{-1} + \sigma^2 CC^T$$

$MM^T$  is variance matrix of the estimator, and it is minimized when  $CC^T = 0$ . Therefore, any estimator besides  $\hat{\beta}$  has strictly greater variance because only  $\hat{\beta}$  will have  $CC^T = 0$ .

Finally, we get that the “best” estimator is where  $C = 0$ , thus getting

$$\tilde{\beta} = (X^T X)^{-1} X^T Y = \hat{\beta}$$

## 4 The Sherlock Holmes approach to Regression

How do you decide what to put into the model?

- Say that you have a theoretical reason to believe your model is correctly specified. Then, that should only leave measurement, sampling, and observation challenges.
  - Weakness: Critics will claim that your model is wrong, that some number of variables are or are not included, and it is hard to say that they are wrong in the social sciences
- Then, since theoretical reasons for including covariates may not be enough, one can ask: “Can the *data* generate an all-cause, complete specification (i.e. correctly specified) model?”

- Leamer’s (1978) “The Axiom of Correct Specification”
  - a The set of explanatory variables that are thought to determine (linearly) the dependent variable must be (1) unique, (2) complete, (3) small in number, and (4) observable.
  - b Other determinants of the dependent variable must have a probability distribution with at most a few unknown parameters
  - c All unknown parameters must be constant
  
- However, Leamer immediately turns around to say:
 

“If this axiom were, in fact, accepted, we would find one equation estimated for every phenomenon, and we would have books that compiled these estimates published with the same scientific fanfare that accompanies estimates of the speed of light or the gravitational constant. Quite the contrary, we are literally deluged with regression equations, all offering to “explain” the same event, and instead of a book of findings we have volumes of competing estimates.”
  
- Instead, most regressions are produced using “data-instigated specification search”, or the “Sherlock Holmes” approach.
  - Instead of using a theoretical based model, a lot of times people will include some covariates because they are available, some because they have a theoretical basis, even some post-treatment variables
  - Very often, then, people run the regression, see what is significant or not (especially the variables of interest), then either remove or add more covariates to change the standard errors
  
- Leamer’s response is that this approach is terrible
 

“ . . . if theories are constructed after having studied the data, it is difficult to establish by how much, if at all, the data favor the data-instigated hypothesis. For example, suppose I think that a certain coefficient ought to be positive, and my reaction to the anomalous result of a negative estimate is to find another variable to include in the equation so that the estimate is positive. Have I found evidence that the coefficient is positive?” (1983)
  
- There are new approaches to data-mining that avoid some of these pitfalls, but it is still a tricky business.