Review of the Classical Linear Regression Model

David A. Hughes, Ph.D.

Auburn University at Montgomery

david.hughes@aum.edu

24 January 2020

Overview

The Classic Linear Regression Model

- Assume the following equation to be true for the population:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i \tag{1}$$

- Which we can rewrite as a series of equations:

$$\begin{array}{c} Y_1 = \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \ldots + \beta_k X_{k1} + \epsilon_1 \\ \underline{Y_2 = \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \ldots + \beta_k X_{k2} + \epsilon_2} \\ Y_n = \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \ldots + \beta_k X_{kn} + \epsilon_n \end{array} \tag{2}$$

The Classic Regression Model

- Looking at equation [2], we can see that really all we have here is a matrix:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \dots & X_{k1} \\ 1 & X_{22} & X_{32} & \dots & X_{k2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{2n} & X_{3n} & \dots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{bmatrix} \tag{3}$$

- Therefore, with no alternative in meaning, we can rewrite equation [1] using matrix notation (note the bold-face type):

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{4}$$

Assumptions of the CLRM

1. Linearity
   - The CLRM is linear in the *parameters* (not necessarily linear in the variables).
2. No Perfect Multicollinearity
   - $\mathbf{X}$ is an n x k matrix of rank K
   - This means that all columns in $\mathbf{X}$ are linearly independent and there are at least K observations
   - There can be no exact linear relationships between two or more variables

Assumptions of the CLRM

3. No endogeneity among data and error term (i.e., $E[\epsilon_i|\mathbf{X}] = 0$)

- The disturbance term should have a conditional expected value of 0 at every observation.
- We can write this as:

$$E[\epsilon|\mathbf{X}] = \begin{bmatrix} E[\epsilon_1|\mathbf{X}] \\ E[\epsilon_2|\mathbf{X}] \\ \vdots \\ E[\epsilon_n|\mathbf{X}] \end{bmatrix} = 0 \qquad (5)$$

- Equation [5] therefore implies:

$$E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta, \qquad (6)$$

which is the linear predictor.

Assumptions of the CLRM

4. Homoskedasticity and Non-autocorrelation

- Var$[\epsilon_i|\mathbf{X}] = \sigma^2,$ for all $i = 1, \ldots, n,$
- Cov$[\epsilon_i, \epsilon_j|\mathbf{X}] = 0,$ for all $i \neq j$
- This tells us that residuals in the CLRM possess consistent variance and that they are uncorrelated across observations

Assumptions of the CLRM

- These assumptions imply that the variance covariance matrix of disturbance terms simplify as follows:

$$E[\epsilon\epsilon'|\mathbf{X}] = \begin{bmatrix} E[\epsilon_1\epsilon_1|\mathbf{X}] & E[\epsilon_1\epsilon_2|\mathbf{X}] & \dots & E[\epsilon_1\epsilon_n|\mathbf{X}] \\ E[\epsilon_2\epsilon_1|\mathbf{X}] & E[\epsilon_2\epsilon_2|\mathbf{X}] & \dots & E[\epsilon_2\epsilon_n|\mathbf{X}] \\ \vdots & \vdots & \vdots & \vdots \\ E[\epsilon_n\epsilon_1|\mathbf{X}] & E[\epsilon_n\epsilon_2|\mathbf{X}] & \dots & E[\epsilon_n\epsilon_n|\mathbf{X}] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

- Which we neatly summarize as:

$$E[\epsilon\epsilon'|\mathbf{X}] = \sigma^2\mathbf{I}, \tag{7}$$

such that $\mathbf{I}$ is the "identity matrix."

Assumptions of the CLRM

5. Nonstochastic Regressors
    - $\mathbf{X}$ must be generated by some means unrelated to ϵ.
6. Normality

- The residuals are normally distributed.
- Formally, we state:

$$\epsilon|\mathbf{X} \sim N[0, \sigma^2\mathbf{I}] \tag{8}$$

OLS is BLUE

- If all of the above assumptions are met, then OLS is the Best Linear Unbiased Estimator (BLUE) we can utilize to estimate the population parameters β and ϵ from equation [4].
- We will denote the statistics we estimate as $\mathbf{b}$ and $\mathbf{e}$ respectively.
- Where:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{9}$$

$\mathbf{b}$ is an Unbiased Estimator of β

$$\begin{aligned}\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad (10)\\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \epsilon)\\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon\\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon\end{aligned}$$

- Because we are interested is the expected value of $\mathbf{b}$ over repeated samples, we take the expected value of $\mathbf{b}$ over repeated iterations of $\mathbf{X}$:

$$E[\mathbf{b}|\mathbf{X}] = \beta + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon|\mathbf{X}] \qquad (11)$$

$\mathbf{b}$ is an Unbiased Estimator of β

- Since $E[\epsilon|\mathbf{X}] = 0$ (by assumption), then:

$$E[\mathbf{b}|\mathbf{X}] = \beta \tag{12}$$

- Or, averaged over all possible values of $\mathbf{X}$

$$E[\mathbf{b}] = \beta \tag{13}$$

$\mathbf{b}$ is a Consistant Estimator of β

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{14}$$

- Since $\mathbf{y} = \mathbf{X}\beta + \epsilon$:

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \epsilon) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon \end{aligned} \tag{15}$$

$\mathbf{b}$ is a Consistant Estimator of β

- Taking expected value:

$$E[\mathbf{b} - \beta] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\epsilon|\mathbf{X}] \tag{16}$$

- Since $E[\epsilon|\mathbf{X}] = 0$ (by assumption):

$$\begin{aligned} E[\mathbf{b} - \beta] &= 0 \\ E[\mathbf{b}] &= \beta \end{aligned} \tag{17}$$

$\mathbf{b}$ is an Efficient Estimator of β

- In addition to Unbiasedness and Consistency, the least squares estimator is also the minimum variance, or most efficient of all unbiased linear estimators
- This can be shown via the Gauss-Markov Theorem
- Moreover, it holds even if $\mathbf{X}$ is stochastic, so long as all other assumptions are met

Violations of the Gauss-Markov Theorem

- When our data exhibit endogeneity, heteroskedasticity, autocorrelation, etc., the assumptions of Gauss-Markov are violated.
- This means that either our $\hat{\beta}$s or $\hat{\sigma}_{\hat{\beta}}$s are biased/inefficient.
- Oftentimes, we can perform a work-around by transforming variables, calculating robust standard errors, etc. But this won't always be the case.

Gauss-Markov and categorical dependent variables

- Categorical and limited dependent variables pose a grave risk both to the Gauss-Markov assumptions and our interpretation of OLS results.
- On the one hand, we're almost certainly violating assumptions of homoskedasticity, normality, etc., which means that we're either getting biased or inefficient results.
- On the other hand, interpreting the values of $\hat{\beta}_k$ can be downright weird.

The linear probability model

- Suppose we code judges' votes on the US Courts of Appeals as being either liberal or conservative ("libvote=1" if yes, "0" else)
- Now suppose we want to predict the likelihood a judge casts a liberal vote solely as a function of his or her ideology.
- We'll let the ideal point of a judge's appointing president stand in for their own ("potus_ideal $\in [-1, 1]$") such that increasing values represent greater conservatism.
- Imagine we estimated the following linear regression:

$$\widehat{\text{libvote}}_i \quad = \quad \hat{\beta}_0 + \hat{\beta}_1 \text{potus\_ideal}_i.$$

The linear probability model: An example

```
. reg libvote potus_ideal
. predict yhat2
. predict res, res

      Source |       SS           df       MS      Number of obs   =    42,156
-------------+----------------------------------   F(1, 42154)     =    287.99
       Model |  68.2337579         1  68.2337579   Prob > F        =    0.0000
    Residual |  9987.62816    42,154  .23693192    R-squared       =    0.0068
-------------+----------------------------------   Adj R-squared   =    0.0068
       Total |  10055.8619    42,155  .238544939   Root MSE        =    .48676

------------------------------------------------------------------------------
     libvote |      Coef.   Std. Err.      t    P>|t|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
 potus_ideal |  -.0838547   .0049413   -16.97   0.000    -.0935397   -.0741696
       _cons |   .3978374   .0023882   166.59   0.000     .3931565    .4025184
------------------------------------------------------------------------------
```

Do we have homoskedasticity?

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of potus_ideal

         chi2(1)      =     7.48
         Prob > chi2  =   0.0062
```

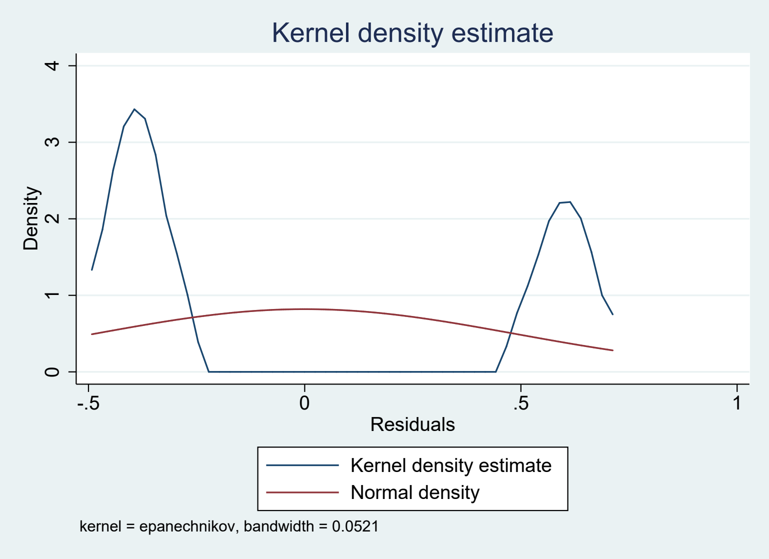
No.

Do we have normally distributed errors?

```
. swilk res

                   Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
         res |     42,156    0.70748   4745.478    23.407    0.00000

Note: The normal approximation to the sampling distribution of W'
      is valid for 4<=n<=2000


. kdensity res, normal
```

No.

Do we have normally distributed errors? (Visually)



Interpreting the LPM

- Suppose we came up with a LPM that adhered to all of the Gauss-Markov asssumptions.
- We still have a problem insofar as we don't really know how to interpret model parameters like slope coefficients.
- All of these problems tell us that OLS is not the appropriate estimator.

Discussion

- OLS is the workhorse of regression analysis, but there are numerous scenarios under which it is an inappropriate regression technique.
- The remainder of this course will focus upon the likelihood theory of inference, the conditions under which it is an appropriate estimation technique, and the conditions under which its own assumptions are violated.