Review of OLS

Week 2
POLS 8830: Advanced Quantitative Methods

Ryan Carlin
Georgia State University
rcarlin@gsu.edu

The Classic Regression Equation

- 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:

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

The Classic Regression Equation

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

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

- Therefore, with no alterative in meaning, we can rewrite equation [1] with the following notation:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \quad (4)$$

Assumptions of the CLRM

1. Linearity
   - The CLRM as specified in the form $Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i$ specifies a linear relationship between y and $x_1, x_2, \ldots, x_k$.
2. Full Rank (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
   - Thus, there are no exact linear relationships

Assumptions of the CLRM

3. $E[\epsilon_i|\mathbf{X}] = 0$

- This assumption implies that the disturbance term should have a conditional expected value of 0 at every observation.
- For the full set of observations, we can write this as:

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

- The assumption in equation [5] is essential, as it implies that:

$$E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta \tag{6}$$

Assumptions of the CLRM

4. Spherical Disturbances (Homoscedasticity and Nonautocorrelation)

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

Assumptions of the CLRM

- Additionally, these assumptions imply that:

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

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

- Which we neatly summarize as:

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

Assumptions of the CLRM

5. Nonstochastic Regressors

- This assumption simply holds that all values in the matrix $\mathbf{X}$ are fixed
- In practice, this assumption does not match the reality of social science data where many of our independent variables of theoretical interest are random
- Thus our assumption is more about the data generating process that produces $\mathbf{x_i}$ as being fixed

Assumptions of the CLRM

6. Normality

- Here we simply add to the list of assumptions about the disturbances by assuming they are normally distributed
- Formally, we state:

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

Implementation

- Base Packages:
  - **glm** or **lm**
    - Generalized linear models, or linear model
- Primary Packages:
  - **lmtest**
    - Tests and Diagnostics for OLS
  - **sandwich**
    - Robust standard errors

Implementation: GLM Syntax

- GLM Implementation
  - `glm(formula, family = gaussian, data, weights, subset, na.action, start = NULL, etastart, mustart, offset, control = list(...), model = TRUE, method = "glm.fit", x = FALSE, y = TRUE, singular.ok = TRUE, contrasts = NULL, ...)`

Implementation: GLM Syntax

- GLM Implementation
  - `glm(formula, family = gaussian, data, weights, subset, na.action, start = NULL, etastart, mustart, offset, control = list(...), model = TRUE, method = "glm.fit", x = FALSE, y = TRUE, singular.ok = TRUE, contrasts = NULL, ...)`
- Main Components:
  - formula: $Y \sim X_1 + X_2 + X_3 \ldots$
  - family: ‘gaussian’ for linear regression
  - data: call to your dataframe, list, or environment

Implementation: GLM Implementation

- `mRate <- glm(Murder ~ Population + Income + Illiteracy, family = gaussian, data = state)`

Implementation: GLM Implementation

- `mRate <- glm(Murder ~ Population + Income + Illiteracy, family = gaussian, data = state)`
- mRate: glm object
- Murder: Outcome Variable
- Population, Income, Illiteracy: Independent Variables
- state: Data Frame or coercable object
- summary(mRate)

Note: state comes from the **datasets** package built into R.

Implementation: GM Assumptions

1. Linearity in the relationship under study
2. Error term is independently and identically distributed normally about 0 with standard deviation of σ^2
3. No perfect multicollinearity between independent variables
4. Spherical errors (υ_i neither correlated with the independent variables nor one another)

Implementation: GM Assumptions

1. Linearity in the relationship under study
   - This is generally going to be a theoretical assumption made in model selection
   - Can use a version of scatterplots to check
     - `qqnorm(residuals(g/lm object))`
     - `qqline(residuals(g/lm object))`

Implementation: GM Assumptions

1. Linearity in the relationship under study
   - `qqnorm(residuals(g/lm object))`
   - `qqline(residuals(g/lm object))`

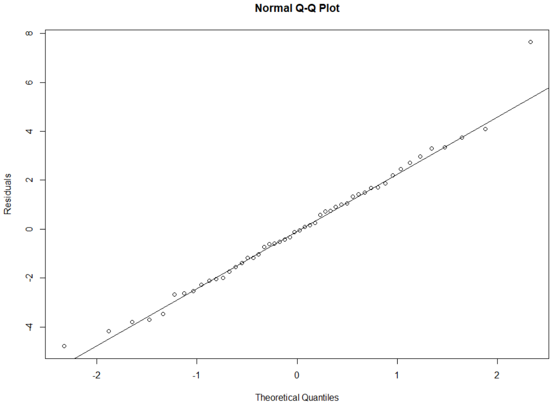

Figure: Sample Q-Q Plot

Implementation: GM Assumptions

2. Error term is independently and identically distributed normally about 0 with standard deviation of σ^2

- `hist(mRate$residuals)`
- `sd(mRate$residuals)`

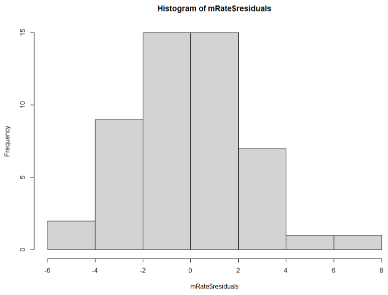

Figure: Distribution of Residuals

Implementation: GM Assumptions

3. No perfect multicollinearity between independent variables
   - Three ways - correlations, tolerance, variable inflation factor

Implementation: GM Assumptions

3. No perfect multicollinearity between independent variables

- Three ways - correlations, tolerance, variable inflation factor
- Correlation
  - `cor.test(IV1, IV2, method = c("pearson", "kendall", "spearman"), exact = NULL, conf.level = 0.95, continuity = FALSE, use = "complete.obs")`

Implementation: GM Assumptions

3. No perfect multicollinearity between independent variables

- Three ways - correlations, tolerance, variable inflation factor
- Tolerance
  - `object=(1-(model$deviance/model$null.deviance))`

Implementation: GM Assumptions

3. No perfect multicollinearity between independent variables

- Three ways - correlations, tolerance, variable inflation factor
- VIF
  - `vif(model)`
- Any IV with a vif greater than 10 needs to be addressed; greater than 5 indicates potential issues

Implementation: GM Assumptions

4. Spherical errors (v_i neither correlated with the independent variables nor one another)

- Heteroskedasticity: Breusch-Pagan Test
    - `bptest(model)`
    - `coeftest(model, vcov = vcovHC(model, "HC1"))`
        - can use sandwich and other SE calculation variants: `"HC0"`, `"HC1"`, `"HC2"`, `"HC3"`, `"arellano"`, etc.

Implementation: GM Assumptions

4. Spherical errors (υ_i neither correlated with the independent variables nor one another)

- Heteroskedasticity: Breusch-Pagan Test
  - `bptest(model)`
  - `coeftest(model, vcov = vcovHC(model, "HC1"))`
    - can use sandwich and other SE calculation variants: `"HC0"`, `"HC1"`, `"HC2"`, `"HC3"`, `"arellano"`, etc.
- Auto/serial correlation: Durbin-Watson Test
  - `dwtest(DV ~ IV1 + IV2 + IV3 ...)`
- Significant results indicate the existence of heteroskedastic errors or serial correlation respectively.