

Multiple Regression: Estimation and Inference

Dr. Michael Fix

mfix@gsu.edu

Georgia State University

15 February 2024

Note: The slides are distributed for use by students in POLS 8810. Please do not reproduce or redistribute these slides to others without express permission from Dr. Fix.

The Model

- Multiple regression:

$$\mathbf{Y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{u}$$

$N \times 1$ $N \times K$ $K \times 1$ $N \times 1$

- or:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + u_i$$

- or:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{K1} \\ 1 & X_{12} & X_{22} & \cdots & X_{K2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1N} & X_{2N} & \cdots & X_{KN} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}$$

Estimating β

- Residuals:

$$\mathbf{u} = \mathbf{Y} - \mathbf{X}\beta$$

- The inner product of \mathbf{u} :

$$\begin{aligned}\mathbf{u}\mathbf{u}' &= [u_1 \quad u_2 \quad \cdots \quad u_N] \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix} \\ &= u_1^2 + u_2^2 + \dots + u_N^2 \\ &= \sum_{i=1}^N u_i^2\end{aligned}$$

Estimating β

- We want to minimize the squared errors, so start with:

$$\begin{aligned}\mathbf{u}'\mathbf{u} &= (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) \\ &= \mathbf{Y}'\mathbf{Y} - 2\beta'\mathbf{X}'\mathbf{Y}' + \beta'\mathbf{X}'\mathbf{X}\beta\end{aligned}$$

- Now get:

$$\frac{\partial \mathbf{u}'\mathbf{u}}{\partial \beta} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta$$

Estimating β

- Solve:

$$-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta = 0$$

$$-\mathbf{X}'\mathbf{Y} + \mathbf{X}'\mathbf{X}\beta = 0$$

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}$$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- **Important Note: Unlike bivariate OLS, we do not compute the estimates using $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$**

Assumptions of the CLRM

1. Linearity

- The CLRM as specified in the form $Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + u_i$ specifies a linear relationship between y and x_1, x_2, \dots, x_k .

2. Full Rank (No Perfect Multicollinearity)

- All columns in \mathbf{X} are linearly independent
- $N > K$

Assumptions of the CLRM

3. $E(\mathbf{u}) = 0$

- 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:

$$E(\mathbf{u}|\mathbf{X}) = \begin{bmatrix} E[u_1|\mathbf{X}] \\ E[u_2|\mathbf{X}] \\ \vdots \\ E[u_n|\mathbf{X}] \end{bmatrix} = 0 \quad (1)$$

- The assumption in equation [1] is essential, as it implies that:

$$E(\mathbf{y}|\mathbf{X}) = \mathbf{X}\beta \quad (2)$$

Assumptions of the CLRM

4. Spherical Disturbances (Homoskedasticity and Nonautocorrelation)
 - $\text{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2$, for all $i = 1, \dots, n$,
 - and
 - $\text{Cov}(u_i, u_j|\mathbf{X}) = 0$, for all $i \neq j$
 - State that the disturbance terms in the CLRM possess constant variance and that they are uncorrelated across observations

Assumptions of the CLRM

- Additionally, these assumptions imply that:

$$E(\mathbf{uu}'|\mathbf{X}) = \begin{bmatrix} E[u_1 u_1 | \mathbf{X}] & E[u_1 u_2 | \mathbf{X}] & \dots & E[u_1 u_n | \mathbf{X}] \\ E[u_2 u_1 | \mathbf{X}] & E[u_2 u_2 | \mathbf{X}] & \dots & E[u_2 u_n | \mathbf{X}] \\ \vdots & \vdots & \vdots & \vdots \\ E[u_n u_1 | \mathbf{X}] & E[u_n u_2 | \mathbf{X}] & \dots & E[u_n u_n | \mathbf{X}] \end{bmatrix}$$
$$= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

- Which we neatly summarize as:

$$E(\mathbf{uu}'|\mathbf{X}) = \sigma^2 \mathbf{I} \quad (3)$$

Assumptions of the CLRM

5. Nonstochastic Regressors

- This assumption simply holds that all values in the matrix \mathbf{X} are fixed
- Or: $\text{Cov}(\mathbf{X}, \mathbf{u}) = 0$
- 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
- Also assumes *no measurement error*

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:

$$\mathbf{u} \sim N[0, \sigma^2 \mathbf{I}] \quad (4)$$

OLS: [Still] Unbiased

- Start with:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

- Substitute OLS $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\end{aligned}$$

OLS: [Still] Unbiased

- and so:

$$\hat{\beta} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}.$$

- By $\text{Cov}(\mathbf{X}, \mathbf{u}) = 0$, we have $E(\hat{\beta}) = \beta$.

$\hat{\beta}$ is a Consistent Estimator of β

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (5)$$

- Since $\mathbf{Y} = \mathbf{X}\beta + \mathbf{u}$:

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \end{aligned} \quad (6)$$

$\hat{\beta}$ is a Consistent Estimator of β

- Taking expected value:

$$E[\hat{\beta} - \beta] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{u}|\mathbf{X}] \quad (7)$$

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

$$\begin{aligned} E[\hat{\beta} - \beta] &= 0 \\ E[\hat{\beta}] &= \beta \end{aligned} \quad (8)$$

$\hat{\beta}$ 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, as we saw last week

Two Approaches

- F-test
 - Compares the model as specified (the unrestricted model) to a restricted model
 - Default in all (?) software is to effectively compare to a null model
 - This doesn't tell us much
 - Mathematically, it is pretty straightforward (we'll omit that here)

Two Approaches

- R^2
 - Often discussed as a measure of the amount of variance explained
 - Effectively calculated as $1 - \frac{\text{Residual Sum of Squared errors}}{\text{Total Sum of Squares}}$
 - Bounded by 0 (no points on regression line) and 1 (perfect fit, all points on regression line)
 - ~~Can be manipulated by~~ Increases when adding additional independent variables

Goodness-of-Fit Summarized

- All (?) software will provide an F-test, R^2 and R^2_{adj}
- Always report these
- Don't pretend that they mean more than they do.

Let's start with a toy model

```
### Load necessary packages ----
# Use install.packages() if you do not have this package
library(tidyverse) # Data manipulation
library(stargazer) # Creates nice regression output tables
library(lmtest)    # Breusch-Pagan test
library(psych)     # Histograms and correlations for a data matrix

### Load your data ----
# We are using V-Dem version 12
my_data <- readRDS("data/vdem12.rds")

# Let's change names of some of these variables for the sake of simplicity
# I am also subsetting it to only US
us_data <- my_data |>
  filter(country_name == "United States of America") |>
  rename(democracy = v2x_polyarchy,
         gdp_per_capita = e_gdppc,
         urbanization = e_miurbani)

### Bivariate OLS ----
# Let's fit a bivariate and multivariate models
simple <- lm(democracy ~ gdp_per_capita, data = us_data)
multiple <- lm(democracy ~ gdp_per_capita + urbanization, data = us_data)

# View model summary
```

Summary of two models

Factors explaining democracy in the US		
Dependent variable:		
	Democracy	
	Simple OLS (1)	Multiple OLS (2)
GDP per capita	0.012 (0.0003) $p = 0.000^*$	0.013 (0.0003) $p = 0.000^*$
Urbanization		0.253 (0.056) $p = 0.00001^*$
Constant	0.332 (0.006) $p = 0.000^*$	0.264 (0.010) $p = 0.000^*$
Observations	231	201
R2	0.904	0.942
Adjusted R2	0.904	0.942
Residual Std. Error	0.063 (df = 229)	0.044 (df = 198)
F Statistic	2,164.562* (df = 1; 229)	1,615.404* (df = 2; 198)
Notes	$p < 0.05$. Standard errors are in parentheses.	

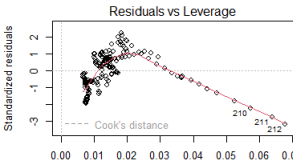
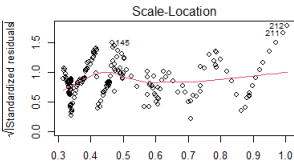
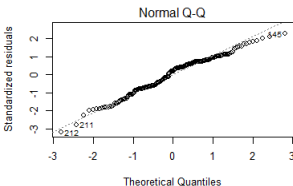
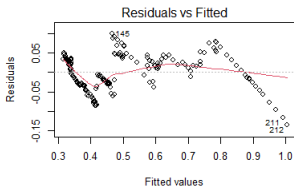
Gauss-Markov assumptions with plot()

```

● ● ●

### Gauss-Markov assumptions using plot() ----
# Let's start with the easiest way
# plot(insert_model_name_here) function will help us to understand these assumptions
plot(multiple)

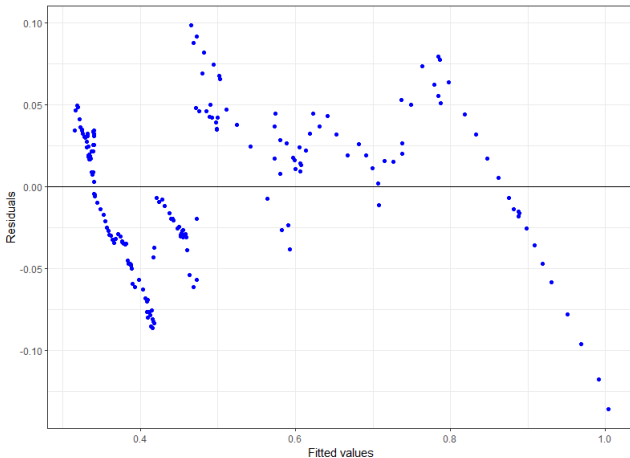
# First plot (left top corner) helps us with homoscedasticity + linearity (red line)
# Second plot (left down corner) helps us with homogeneity of variance
# Third plot (right top corner) helps us with normality of residuals
# Fourth plot (right down corner) shows outliers and high leverage points
    
```



Gauss-Markov assumptions: Homoscedasticity

```
### Gauss-Markov assumptions using other functions ----  
# You can use visuals or tests  
# Looking for heteroskedasticity - plotting residuals ~ fitted.values  
multiple |>  
  ggplot(aes(x = .fitted, y = .resid)) +  
  geom_point(col = 'blue') +  
  geom_abline(slope = 0) +  
  labs(x = "Fitted values", y = "Residuals") +  
  theme_bw()  
  
# Perform Breusch-Pagan test  
bptest(multiple)  
  
# Since the p-value is less than 0.05, we reject the null hypothesis.  
# We have sufficient evidence to say that heteroscedasticity is present in the model.
```

Gauss-Markov assumptions: Homoscedasticity

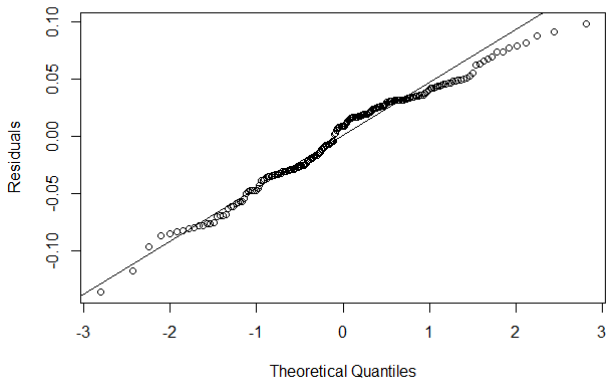


```
> # Perform Breusch-Pagan test  
> bptest(multiple)  
  
studentized Breusch-Pagan test  
  
data: multiple  
BP = 13.71, df = 2, p-value = 0.001054
```


Gauss-Markov assumptions: Linearity of residuals

```
● ● ●  
# Linearity of Relationship  
# Use qqnorm and qqline to examine linearity assumption  
qqnorm(residuals(multiple), ylab = "Residuals")  
qqline(residuals(multiple))
```

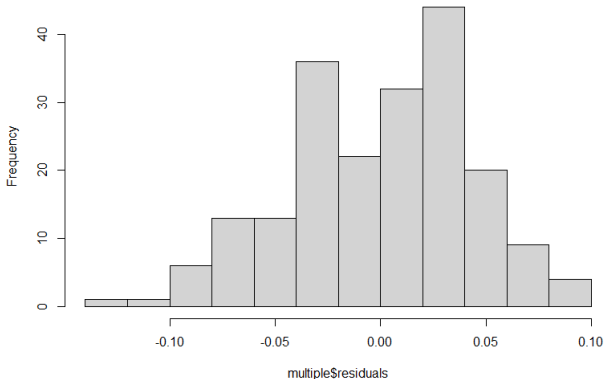
Normal Q-Q Plot



Gauss-Markov assumptions: Normality of residuals

```
# Testing for multicollinearity
us_data |>
  select(democracy, gdp_per_capita, urbanization) |>
  pairs.panels(lm = T,
              method = "pearson")
```

Histogram of multiple\$residuals



Gauss-Markov assumptions: Autocorrelation



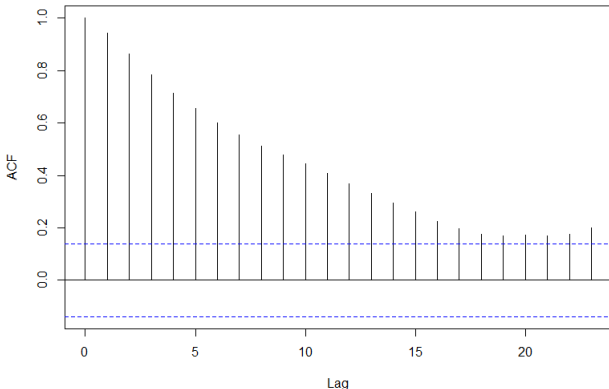
```
# Serial/autocorrelation
# You can visualize or use a test
# Durbin-Watson test for autocorrelation
dwtest(multiple, data = us_data)

# Or, plot residuals for autocorrelation
stats::acf(multiple$residuals, type = "correlation")
```

```
> detest(multiple, data = us_data)
Durbin-Watson test

data: multiple
DW = 0.065127, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Series multiple\$residuals



First, A Pretty Table

Table: A Toy Model

	Coefficient	p -Value
GDP per Capita	0.011 (0.000)	0.000
Urbanization	0.253 (0.056)	0.000
Intercept	0.264 (0.010)	0.000
N	201.	
R^2	0.942	
R^2_{adj}	0.942	

Note: Dependent variable is Democracy. Standard errors in parentheses.

TEXCode for Table

```

\begin{table}[h!]
  \begin{center}
    \caption{A Toy Model}
    \begin{tabular}{l r@{.} l r@{.} l }
      \hline
      \hline
      & \multicolumn{2}{c}{Coefficient}& \multicolumn{2}{c}{$p$-Value}\\
      \hline
      GDP per Capita & 0 & 011 & 0 & 000 \\
      & (0 & 000) & \multicolumn{2}{c}{} \\
      Urbanization & 0 & 253 & 0 & 000 \\
      & (0 & 056) & \multicolumn{2}{c}{} \\
      Intercept & 0 & 264 & 0 & 000 \\
      & (0 & 010) & \multicolumn{2}{c}{} \\
      \hline
      N & 201 & & \multicolumn{2}{c}{} \\
      $R^2$ & 0 & 942 & \multicolumn{2}{c}{} \\
      $R^2_{-adj}$ & 0 & 942 & \multicolumn{2}{c}{} \\
      \hline
      \hline
    \end{tabular} \\
  \end{center}
  \medskip
  Note: Dependent variable is XXX. Standard errors in parentheses.
\end{table}

```

Interpreting Estimates

- Beyond begin BLUE, OLS is nice because of the ability to interpret coefficient estimates as independent effects
- We can write the information in the table as an equation to help think about interpretation:

$$DV = 0.264 + 0.011GDP + 0.253Urban$$

- We can thus say “a one unit increase in GDP corresponds with a 0.011 unit increase in DV.”

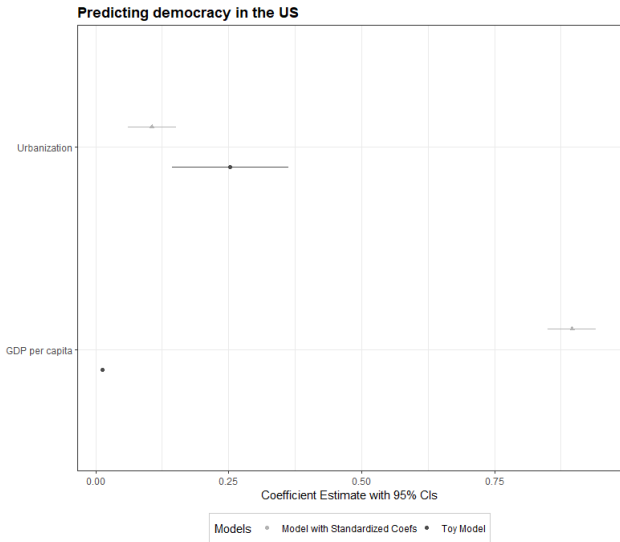
“Standardized” Coefficients

- An alternative to presenting our estimates of $\hat{\beta}$ is to present “standardized” coefficients
- The logic is to be able to compare effect sizes for things that are not on a common scale. E.g. can we say that GDP or Urbanization has a greater *substantive* effect on DV?

“Standardized” Coefficients

```
### Standardized coefficients (beta coeffs) ----  
  
# Standardized Regression Coefficients => b_i(S_Xi/S_Y)  
  
# There are different ways to get your beta coefficients, but use base R  
streg_multiple <- lm(data.frame(scale(multiple$model))) # Get standardized regression  
coefficients  
  
# Or you could scale each variable when running lm() - like this  
# lm(scale(democracy) ~ scale(gdp_per_capita) + scale(urbanization), data = us_data)  
  
# Let's make a coefficient plot of these two models to see the difference  
# I am going to use tidy() function from broom to create a nice plot  
m1_tidy <- tidy(multiple) |> mutate(model = "Toy Model")  
m2_tidy <- tidy(streg_multiple) |> mutate(model = "Model with Standardized Coefs")  
all_models <- bind_rows(m1_tidy, m2_tidy) # combine these models  
  
# Let's plot  
dwplot(all_models,  
  show_intercept = F,  
  dot_args = list(aes(colour = model, shape = model)),  
  size = 3) |>  
  relabel_predictors(c(urbanization = "Urbanization", gdp_per_capita = "GDP per capita")) +  
  theme_bw() +  
  theme(plot.title = element_text(face="bold"),  
    legend.position = "bottom",  
    legend.background = element_rect(colour="grey80"),  
    legend.title.align = .5) +  
  labs(x = "Coefficient Estimate with 95% CIs",  
    y = "", title = "Predicting democracy in the US") +  
  scale_shape_discrete(name = "Models", breaks = c(0, 1)) + # breaks assign shapes  
  scale_colour_grey(start = .3, end = .7, name = "Models")
```


“Standardized” Coefficients using dot-and-whisker plot



“Standardized” Coefficients

- The issue is that standardizing coefficient alters our interpretation.
- Now we can only say that “a one *standard deviation* increase in GDP the DV increases by *XXXX standard deviations*”
- Great! Now we can *sort of* directly compare effects sizes, but...
 - What does this mean?
 - For other issues, see [King \(1986, 669–674\)](#)