Polimetrics II: Advanced Quantitative Methods for Political Scientists*
School of Politics & Global Studies
Arizona State University

Babak RezaeeDaryakenari
Data & codes available at
`https://github.com/babakrezaee/POS-604`

Contents

* WORK IN PROGRESS: Please do not circulate or cite without permission

1 Matrix algebra

A *matrix* is simply a rectangular array of numbers. So, any table of data is a matrix. The size of a matrix is indicated by the number of its rows and the number of its columns. A matrix with k rows and n columns is called a $k \times n$ ("k by n") matrix. The number in row i and column j is called the (i, j)th entry, and is often written a_{ij}. Two matrices are *equal* if they both have the same size and if the corresponding entries in the two matrices are equal.

Matrices are in a sense generalized numbers. When the sizes are right, two matrices can be added, subtracted, multiplied, and even divided.

1.1 Addition

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & a_{ij} & \vdots \\ a_{k1} & \dots & a_{kn} \end{bmatrix} + \begin{bmatrix} b_{11} & \dots & b_{1n} \\ \vdots & b_{ij} & \vdots \\ b_{k1} & \dots & b_{kn} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11} & \dots & a_{1n} + b_{1n} \\ \vdots & b_{ij} + b_{ij} & \vdots \\ a_{k1} + b_{k1} & \dots & a_{kn} + b_{kn} \end{bmatrix}$$

1.2 Subtraction

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & a_{ij} & \vdots \\ a_{k1} & \dots & a_{kn} \end{bmatrix} - \begin{bmatrix} b_{11} & \dots & b_{1n} \\ \vdots & b_{ij} & \vdots \\ b_{k1} & \dots & b_{kn} \end{bmatrix} = \begin{bmatrix} a_{11} - b_{11} & \dots & a_{1n} - b_{1n} \\ \vdots & b_{ij} - b_{ij} & \vdots \\ a_{k1} - b_{k1} & \dots & a_{kn} - b_{kn} \end{bmatrix}$$

1.3 Scalar multiplication

$$r \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & a_{ij} & \vdots \\ a_{k1} & \dots & a_{kn} \end{bmatrix} = \begin{bmatrix} ra_{11} & \dots & ra_{1n} \\ \vdots & ra_{ij} & \vdots \\ ra_{k1} & \dots & ra_{kn} \end{bmatrix}$$

1.4 Matrix multiplication

Not all pairs of matrices can be multiplied together, and the order in which matrices are multiplied can matter.

We can define the matrix product AB if and only if

number of columns of A=number of rows of B

For the matrix product to exist, A must be $k \times m$ and B must be $m \times n$. To obtain the (i, j)th entry of AB, multiply the ith row of A and the jth column of B as follows:

$$r\begin{bmatrix} a_{i1} & a_{i2} & \dots & a_{im} \end{bmatrix} . \begin{bmatrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{mj} \end{bmatrix} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{im}b_{mj}$$

For example,

$$\begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} aA + bC & aB + bD \\ cA + dC & cB + dD \\ eA + fC & eB + fD \end{bmatrix}$$

The $n \times n$ matrix

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \ddots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

has the property that for any $m \times n$ matrix A,

$$AI = A$$

and for any $n \times 1$ matrix B,

$$IB = B$$

The matrix I is called the $n \times n$ *identity matrix* because it is a multiplicative identity for matrices just as the number 1 is for real numbers.

1.5 Laws of matrix algebra

- Associative laws: $(A + B) + C = A + (B + C)$, $(AB)C = A(BC)$
- Commutative law for addition: $A + B = B + A$
- Distributive laws: $A(B + C) = AB + AC$ $(A + B)C = AC + BC$

1.6 Transpose

There is one other operation on matrices which we shall frequently use. The *transpose* of $k \times n$ matrix A is the $n \times k$ matrix obtained by interchanging the rows and columns of A. This matrix is often written as A^T. The first row of A becomes the

first column of A^T. The second row of A becomes the second column of A^T, and so on. For example,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}^T = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{bmatrix}$$

The properties of transpose:

- $(A + B)^T = A^T + B^T$
- $(A - B)^T = A^T - B^T$
- $(A^T)^T = A$
- $(rA)^T = rA^T$
- $(AB)^T = B^T A^T$

2 Computing derivatives

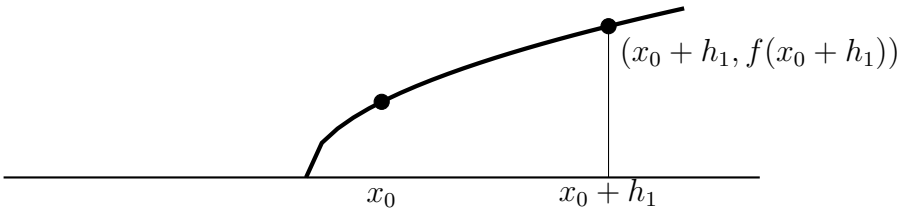

Figure 1: Approximating the tangent line by a sequence of secant lines.

Since l_n passes through the two points $(x_0, f(x_0))$ and $(x_0 + h_n, f(x_n + h_n))$, its slope is

$$\frac{f(x_0 + h_n) - f(x_0)}{(x_0 + h_n) - x_0} = \frac{f(x_0 + h_n) - f(x_0)}{h_n}. \tag{1}$$

Therefore, the slope of the tangent line is the limit of this process as h_n converges to 0.
Definition Let $(x_0, f(x_0))$ be a point on the graph of y=f(x). The derivative of f at x_0, written

$$f'(x_0) \text{ or } \tfrac{df}{dx}(x_0) \text{ or } \tfrac{dy}{dx}(x_0),$$

is the slope of the tangent line to the graph of f at $(x_0, f(x_0))$. Analytically,

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \tag{2}$$

3 Review: Ordinary Least Square (OLS)

As part of Popperian and Lakatosian research paradigm, we need to evaluate our theoretical arguments. While some scholars rely on qualitative methods, others employ quantitative approaches to test their hypotheses[1].

In social science, we are interested to explain whether and how two factor/variables are associated. That is, how variable x affects variable y. For example: how does democracy, x, affect economy, y?, how ethnic diversity, x, affect the risk of civil war, y?, and so forth.We can show it as:

$$x \xrightarrow{?} y$$

One of the main concerns of us in this class,an almost all of social science metrics classes, is modeling this association. You are probably familiar with Acemoglu, Johnson, and Robinson (2001)[2]. The authors estimate the effect of institutions on economic performance.

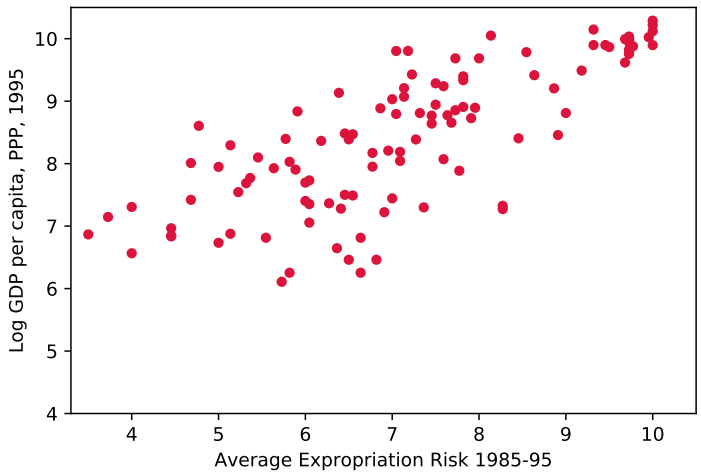

Figure 2: The scatter diagram of *protection against expropriation* and *log GDP per capita*.

- Outcome variable/Dependent variable/Endogenous variable (y): *log GDP per capita* a proxy of economic performance.

[1] Of course, some scholars mix qualitative and quantitative methods in their empirical analysis, known as mixed-methods.

[2] Daron Acemoglu, Simon Johnson, and James A Robinson. The colonial origins of comparative development: an empirical investigation. The American Economic Review, 91(5):1369–1401, 2001. http://www.jstor.org/stable/2677930

- Independent variable/Regressor/Explanatory variable/Exogenous variable (x): *protection against expropriation* a proxy of institutional quality.

We need a function to model this association. Formally,

$$y = f(x) \tag{3}$$

There are literally infinite number of functions that we can nominate to model the association between x and y (See Figure 3). However, the question is which one is the best, or one of the best options?

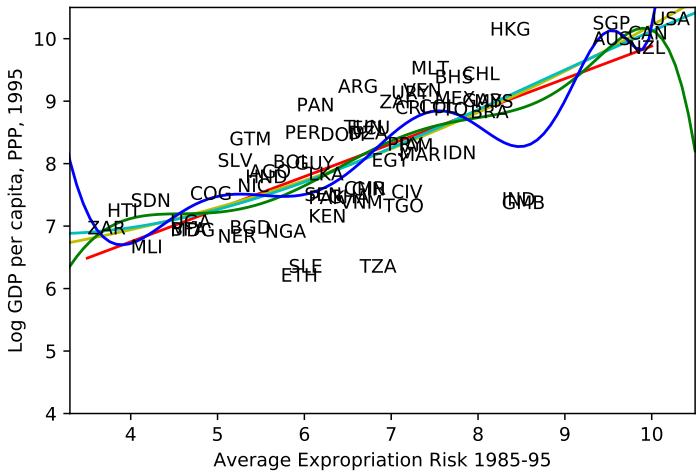

Figure 3: The scatter diagram of *protection against expropriation* and *log GDP per capita*.

In POS 603: Polimetrics I, you learned about *Ordinary Least Square (OLS)* method. *OLS* is a linear fit to the association between x and y (See Figure 4). Why do we use OLS? What is OLS method?

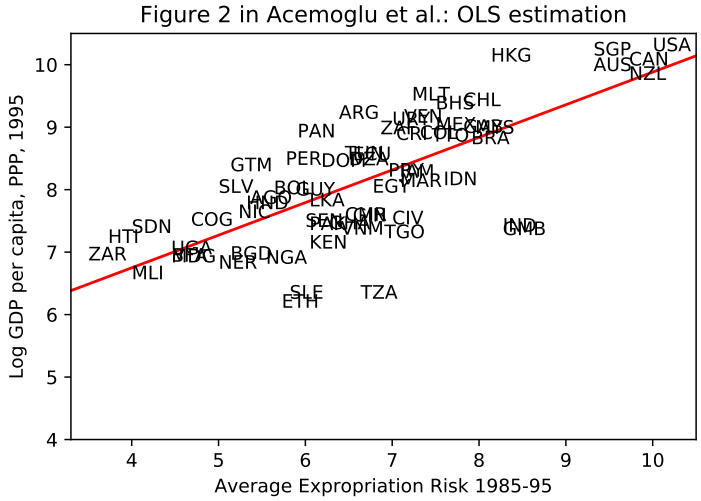

Figure 4: The linear fit of *protection against expropriation* and *log GDP per capita*.

3.1 Bivariate OLS

Our statistical model for a linear bivariate model is

$$y = \beta_0 + \beta_1 x + \epsilon \tag{4}$$

where y is a vector of outcome, x is a vector of independent variable, ϵ is disturbance, β_0[3] is the intercept, and β_1 is the slope in this linear model. y and x are given, meaning we have data on them. We need to estimate the parameters of this model: β_0 and β_1. But, how? Ordinary Least Square estimate the parameters using minimizing the sum of squared residuals (SSR):

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i}^{N} e_i^2 \tag{5}$$

where $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

Exercise 1:

Show that OLS estimator for a bivariate model is as follow:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

[3] Some sources use α to name the intercept of the model.

3.2 OLS in Matrix Form

Let

- X be an $N \times k$ matrix where we have observations on k independent variables for N observations. Since our model will usually contain a constant term, one of the columns in the X matrix will contain only ones. This column should be treated exactly the same as any other column in the X matrix.
- y be an $N \times 1$ vector of observations on the dependent variable.
- ϵ be an $N \times 1$ vector of disturbances or errors.
- β be an $k \times 1$ vector of unknown population parameters that we want to estimate.

Our statistical model will essentially look something like the following:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{N\times 1} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1K} \\ 1 & x_{21} & \dots & x_{2K} \\ \vdots & \dots & x_{nk} & \vdots \\ 1 & x_{n1} & \dots & x_{NK} \end{bmatrix}_{N\times K} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}_{K\times 1} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}_{N\times 1}$$

or, we can write

$$Y = X\beta + \epsilon \tag{6}$$

It is assumed that this is a simplified reflection of the world that we want to model. The model has a systematic component (βX) and a stochastic component (ϵ). Our goal is to obtain estimates of the population parameters in the β vector.

The *sum of squared residuals (RSS)* in matrix format is $\epsilon'\epsilon$[4]:

$$\begin{bmatrix} e_1 & e_2 & \dots & e_N \end{bmatrix}_{1\times N} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}_{N\times 1} = \begin{bmatrix} e_1^2 + e_2^2 + \dots + e_N^2 \end{bmatrix}_{1\times 1} \tag{7}$$

[4] It is important to note that this is very different from $\epsilon'\epsilon$, the variance-covariance matrix of residuals.

We can also re-write the SSR as follow:

$$\begin{aligned}\epsilon'\epsilon &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \end{aligned} \tag{8}$$

where we use the fact that the transpose of a scalar is the scalar i.e. $y'X\hat{\beta} = (y'X\hat{\beta})' = \hat{\beta}'X'y$.

To find OLS estimator, we should following optimization problem:

$$\min_{\hat{\beta}} \epsilon'\epsilon \tag{9}$$

We need to take the derivative of Eq. 3.2 with respect to $\hat{\beta}$. This gives the following equation[5]:

$$\frac{\partial \epsilon'\epsilon}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0 \tag{10}$$

To check this is a minimum, not a maximum, we would take the derivative of Eq. 3.2 with respect to $\hat{\beta}$ again, giving us $2X'X > 0$.

From Eq. 3.2,

$$(X'X)\hat{\beta} = X'y \tag{11}$$

$$\begin{aligned}(X'X)^{-1}(X'X)\hat{\beta} &= (X'X)^{-1}X'y \\ I\hat{\beta} &= (X'X)^{-1}X'y \\ \hat{\beta} &= (X'X)^{-1}X'y \end{aligned} \tag{12}$$

3.3 Properties of OLS

1. The observed values of X are uncorrelated with the residuals.

$$\begin{aligned}(X'X)\hat{\beta} &= X'y \\ (X'X)\hat{\beta} &= X'(X\hat{\beta} + \epsilon) \\ (X'X)\hat{\beta} &= X'X\hat{\beta} + X'\epsilon \\ X'\epsilon &= 0 \end{aligned} \tag{13}$$

[5] $\frac{\partial a'b}{\partial b} = \frac{\partial b'a}{\partial b} = a$ when a and b are $K \times 1$. $\frac{\partial b'Ab}{\partial b} = 2Ab = 2b'A$, where A is any symmetric matrix.

2. The sum of the residuals is zero.

If there is a constant, then the first column in X will be a column of ones. This means that for the first element in the $X'e$ vector (i.e. $X_{11}e_1 + x_{21}e_2 + \cdots + x_{N1}e_N$) to be zero, it must be the case that $\sum_1^N e_i = 0$.

3. The sample mean of the residuals is zero.

This follows straightforwardly from the previous property, $\bar{e} = \frac{\sum_1^N e_i}{N} = 0$.

4. The regression hyperplane, line in a bivariate model, passes through the means of the observed values ($\bar{x}$ and $\bar{y}$).

$\bar{e} = 0 \Rightarrow e = y - X\beta \Rightarrow \frac{e}{N} = \frac{y}{N} - \frac{x}{N}\hat{\beta} \Rightarrow \bar{e} = \bar{y} - \bar{x}\hat{\beta} \Rightarrow$
$0 = \bar{y} - \bar{x}\hat{\beta} \Rightarrow \bar{y} = \bar{x}\hat{\beta}$

5. The predicted values of y are uncorrelated with the residuals.

Exercise 2:

Prove this property of OLS.

These properties always hold true. You should be careful not to infer anything from the residuals about the disturbances.

Note that we know nothing about $\hat{\beta}$ except that it satisfies all of the properties discussed above. We need to make some assumptions about the true model in order to make any inferences regarding β (the true population parameters) from $\hat{\beta}$ (our estimator of the true parameters). Recall that $\hat{\beta}$ comes from our sample, but we want to learn about the true parameters.

3.4 The Gauss-Markov Assumptions

1. Linearity assumption: $y = X\beta + \epsilon$

2. X is an $N \times K$ matrix of full rank.

This assumption states that there is no perfect multicollinearity. In other words, the columns of X are linearly independent. This assumption is known as the identification condition.

3. $E(\epsilon|X) = 0$

This assumption - the zero conditional mean assumption - states that the disturbances average out to 0 for any value of X. Put differently, no observations of the independent variables convey any information about the expected value of the disturbance.

The assumption implies that $E(y) = X\beta$. This is important since it essentially says that we get the mean function right.

4. $E(\epsilon\epsilon'|X) = \sigma^2 I$

This captures the familiar assumption of homoskedasticity and no autocorrelation.

5. $X \perp \epsilon$

X may be fixed or random, but must be generated by a mechanism that is unrelated to ϵ:

6. $\epsilon|X \sim N(0, \sigma^2 I)$.

This assumption is not actually required for the Gauss-Markov Theorem. However, we often assume it to make hypothesis testing easier. The Central Limit Theorem is typically evoked to justify this assumption.

Theorem 1: The Gauss-Markov Theorem

The Gauss-Markov Theorem states that, conditional on assumptions 1-5, OLS estimator is the Best Linear, Unbiased and Effcient estimator (BLUE):

1. $\hat{\beta}$ is an unbiased estimator of β.
2. $\hat{\beta}$ is a linear estimator of β.
3. $\hat{\beta}$ has minimal variance among all linear and unbiased estimators.

3.5 Hypothesis Testing

Assumption 6:

$$\epsilon \sim N(0, \sigma^2 I) \tag{14}$$

And, from BLUE, we know that $\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon$, so we can show:

$$\hat{\beta} \sim N(\beta, \sigma(X'X)^{-1}) \tag{15}$$

Thus, our assumption of normal distribution for ϵ lead to a normal distribution for the estimator of β, i.e. $\hat{\beta}$.

Here is an approximate distribution of $\hat{\beta}_1$ for Acemoglu et al.(2001).

Dep. Variable:	log GDP	**R-squared:**	0.611
Model:	OLS	**Adj. R-squared:**	0.608
Method:	Least Squares	**F-statistic:**	171.4
Date:	Wed, 17 Jan 2018	**Prob (F-statistic):**	4.16e-24
Time:	01:18:50	**No. Observations:**	111
DF	109		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Constant	4.6261	0.301	15.391	0.000	4.030	5.222
Expropriation Risk	0.5319	0.041	13.093	0.000	0.451	0.612

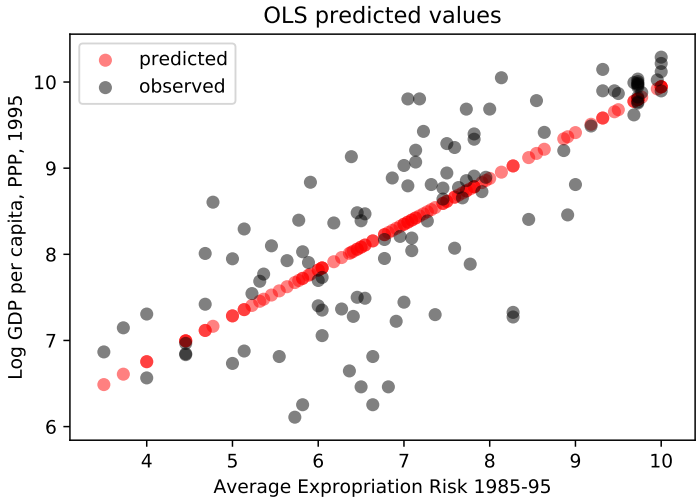

Figure 5

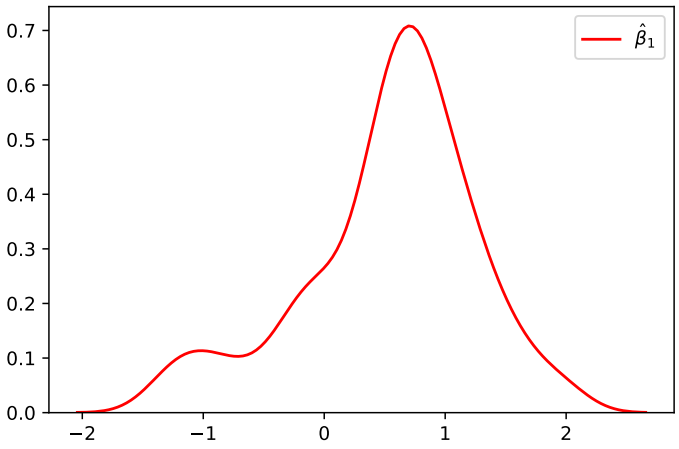

Figure 6: The distribution of $\hat{\beta}_1$ using $\hat{\beta}_1 = \beta_1 + (X'X)^{-1}X'e$ assuming that $\hat{\beta}_1$ is very close to the true value of β_1.

This distributions looks very much similar to the distribution of error terms.

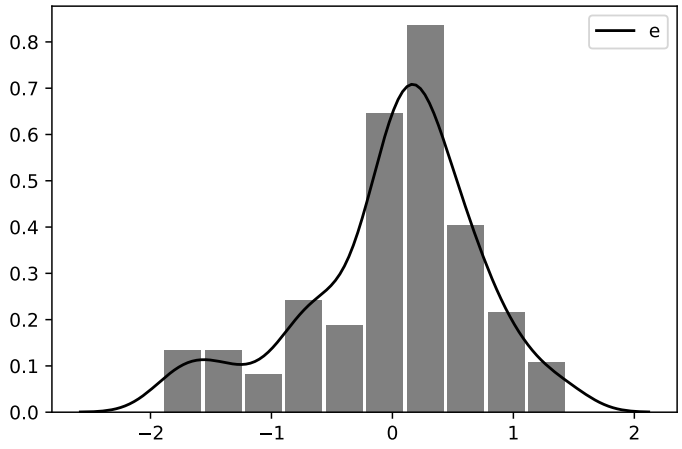

Figure 7: The distribution of $\hat{\beta}_1$ using $\hat{\beta}_1 = \beta_1 + (X'X)^{-1}X'e$ assuming that $\hat{\beta}_1$ is very close to the true value of β_1.

The distribution of $\hat{\beta}_1$ look like a normal distribution, but why do we use *t-test*?

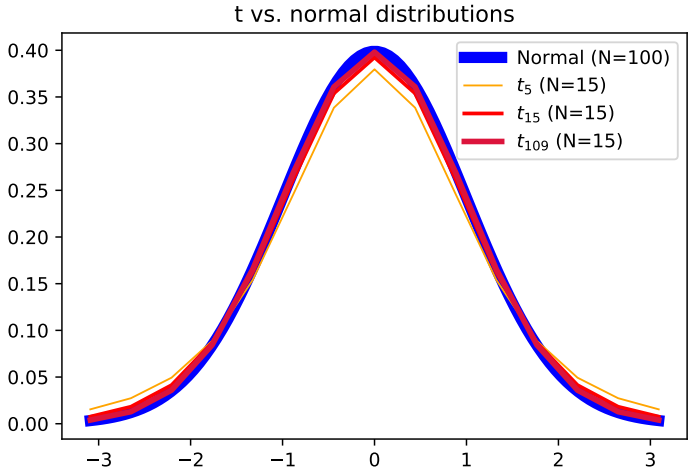

Figure 8: The distribution of $\hat{\beta}_1$ using $\hat{\beta}_1 = \beta_1 + (X'X)^{-1}X'e$ assuming that $\hat{\beta}_1$ is very close to the true value of β_1.

Now, we can run our statistical tests (of course thanks to Guinness Brewery)!

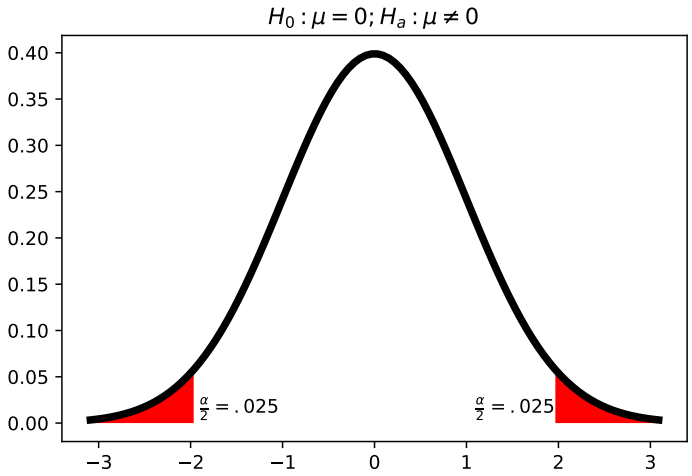

Figure 9: Two tail test.

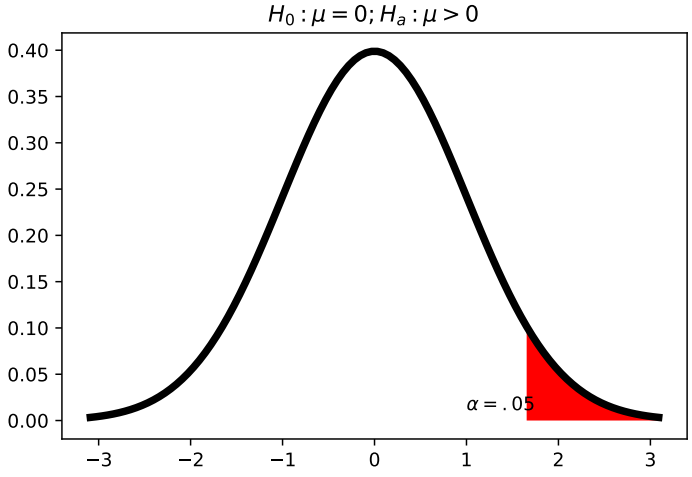

Figure 10: One(right) tail test.

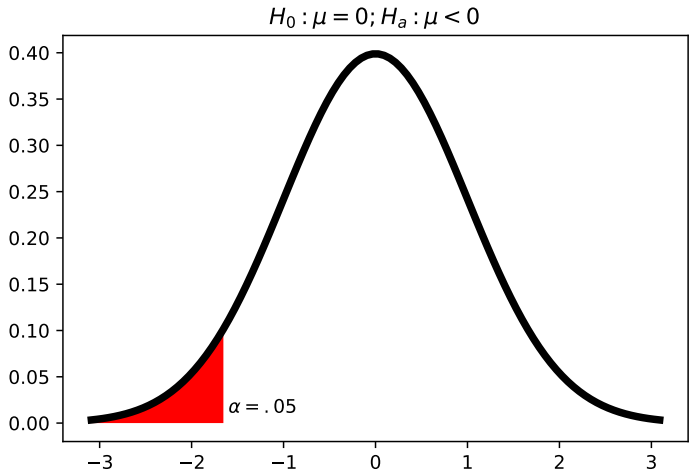

Figure 11: One(left) tail test.

3.6 Estimating Non-Linear Models Using OLS

The OLS linear method of estimation usually works very well for estimating linear association between the dependent and independent variables. However, the world is not necessarily linear. Figure 12 shows a non-linear association between two variables. While

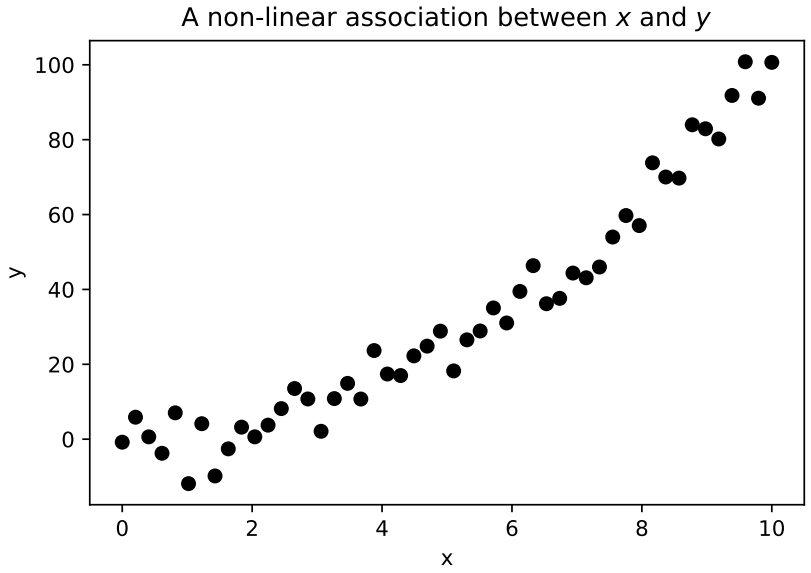

Figure 12: Non-linear association between x and y.

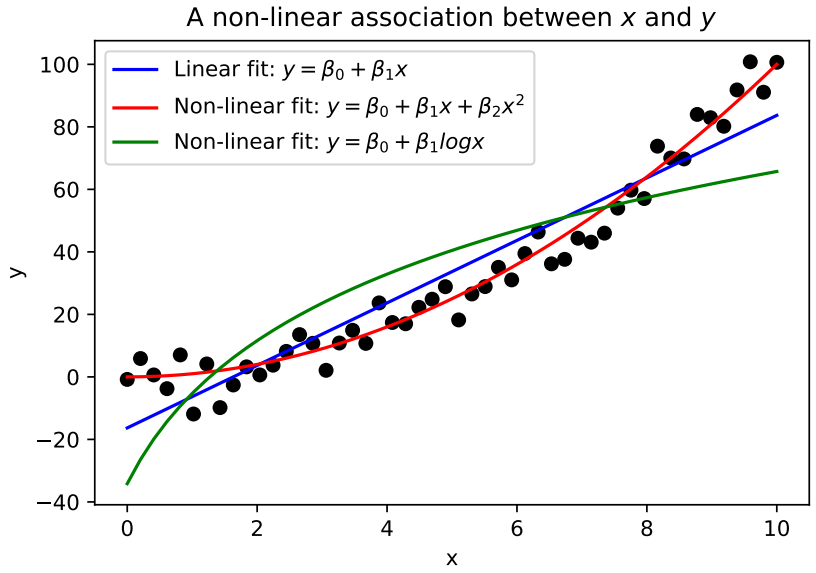

Figure 13: Estimating non-linear association between x and y, using different transformation functions.

4 Binary Outcome Models

It is common that political scientists work on research questions in which the outcome variable is a binary/categorical variable. Do we observe civil war in country x? Does person i vote in an election? Will the UN Security Council impose sanction on country x?

The simplest case of a binary outcomes is a model with two outcomes. Let's look at the scatter graph of the association between x and y in which y is a binary outcome (Figure 16). This graph also shows the fitted OLS models for the association between x and our binary outcome, y. What does the graph tell us about the estimated models?

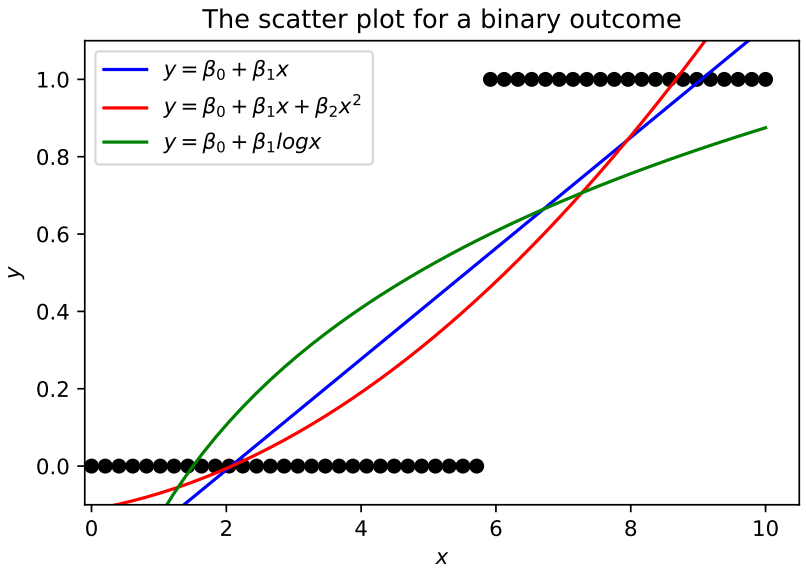

Figure 14: Binary outcome and OLS fit.

4.1 Which OLS assumptions are violated in models with binary outcome?

Estimating a model with binary outcome using OLS model violates three of the Gauss-Markov assumptions:

- **Violation 1: Heteroskedasticity**
  A residuals versus fitted plot in OLS ideally looks like a random scatter plot, as $E(y|X\beta, x) = 0$. This means that $Var(\epsilon) = \sigma_i I \neq \sigma I$.

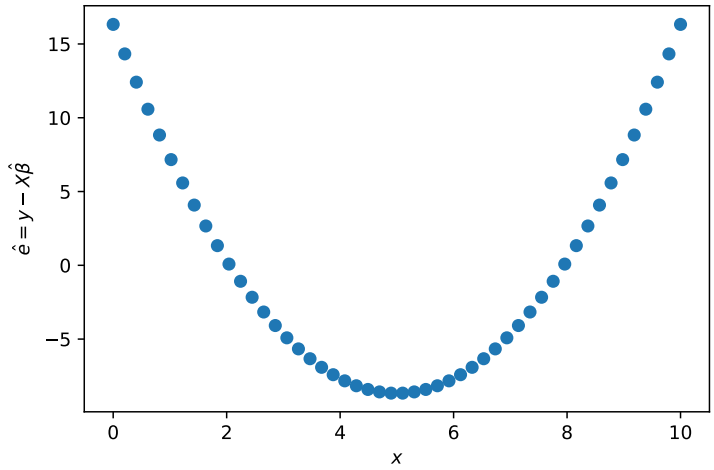

Figure 15: Non-random association between $\hat{e}$ and x

Formal proof: Let $X_i = \{0, 1\}$ and assume $p = pr(x_i = 1)$ then $X_i^2 = X_i$. Thus, $V(x_i) = E(x_i^2) - E(x_i)^2 = E(x_i) - E(x_i)^2 =$

$p - p^2 = p_i(1 - p_i)$.
Is p_i constant across all individuals?

Although Heteroskedasticity is a violation of classic OLS assumptions, it can be resolved using robust standard errors.

- **Violation 2: Error terms are not normally distributed**
  This problem can be discussed in relation to the previous violation.

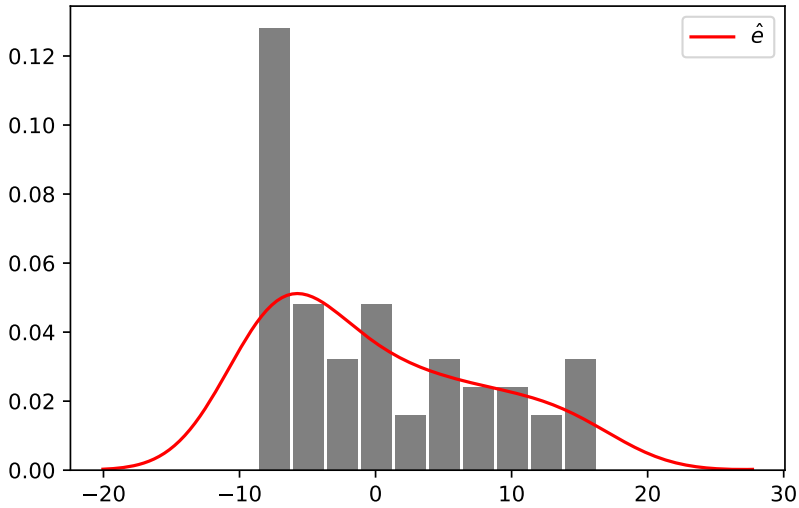

Figure 16: Non-random association between $\hat{e}$ and x

- **Violation 3: Linearity**
  An *OLS* regression of y on x ignores the discreteness of the dependent variable and does not constrain predicted probabilities to be between zero and one.

4.2 *logit* and *probit* models

To address these issues, we can form a regression model by parameterizing the probability p_i to depend on a regressor vector x and a $K \times 1$ parameter vector β. The commonly used models are of single-index form with conditional probability given by

$$p_i \equiv Pr[y_i = 1|X] = F(x_i\beta) \tag{16}$$

where $F(.)$ is a specified function. If we assume that $F(\mathring{u})$ is the *cdf* of the logistic distribution, the estimated model is called *logit*. Similarly, if we use assume that $F(\mathring{u})$ is the standard normal cdf, then the estimated model is *probit*. Note that if $F(\mathring{u})$ is a cdf, then this cdf is only being used to model the parameter p and does not denote the cdf of y itself.

5 Maximum Likelihood (ML) Estimator

Consider the sample: $\{(y_i, x_i), i = 1, ..., N\}$. The maximum likelihood (ML) estimator maximizes the likelihood function, see below. The likelihood function is the joint density, which given independent observations is the product $\prod_i f(y_i|x_i, \beta)$ of the individual densities, where we have conditioned on the regressors.

$$P(\beta|y, X) = \frac{f(y, X|\beta)P(\beta)}{P(y, X)} \tag{17}$$

$$\mathcal{L}(\beta|y, x) \equiv \prod_i f(y_i|x_i, \beta) \tag{18}$$

The question is how we can estimate the parameters of this model such that $P(Y|y, X)$ is maximized.

$$\max_{\hat{\beta}} \mathcal{L}(\beta|y, x) \equiv \prod_i f(y_i|x_i, \beta) \tag{19}$$

5.1 Estimating a normally distributed variable

Assume $y_i \sim N(\mu, \sigma)$, so:

$$P(y_i = Y) = \frac{1}{(2\pi\sigma^2)^{-\frac{1}{2}}} e^{[\frac{-(y_i-\mu_i)^2}{2\sigma^2}]} \tag{20}$$

Assuming that $(y_1, y_2, ..., y_n)$ are distributed independently, we have:

$$\begin{aligned} f(y_1, ..., y_n|\mu, \sigma) = \\ f(y_1|\mu, \sigma)f(y_2|\mu, \sigma)...f(y_n|\mu, \sigma) = \\ \prod_i f(y_i|\mu, \sigma) \end{aligned} \tag{21}$$

20 in 21:

$$\prod_i f(y_i|\mu, \sigma) = (\frac{1}{2\pi\sigma^2})^{\frac{n}{2}} exp(-\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2}) \tag{22}$$

Working with this function is easier, if we transfer it to a *log-liklihood model*, and this transformation does not affect the optima of function, as logarithm functions are strictly increasing.

$$log(\mathcal{L}(\mu,\sigma)) = (-\frac{n}{2})log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 \tag{23}$$

We now can find the optimal parameters of *MLE* by computing the derivatives of the log-likelihood function as follow:

F.O.C.:

$$\frac{\partial}{\partial\mu}log(\mathcal{L}(\mu,\sigma)) = \frac{-2n(\bar{y}-\mu)}{2\sigma^2} = 0 \tag{24}$$

$$\hat{\mu} = \bar{y} = \sum_{i=1}^{n}\frac{y_i}{n} \tag{25}$$

$$\begin{aligned}\frac{\partial}{\partial\sigma}log((\frac{1}{2\pi\sigma^2})^{\frac{n}{2}}exp(-\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2+n(\bar{y}-\mu)^2}{2\sigma^2})) = 0 \Rightarrow \\ \frac{\partial}{\partial\sigma}(\frac{n}{2}log(\frac{1}{2\pi\sigma^2}) - \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2+n(\bar{y}-\mu)^2}{2\sigma^2}) = \end{aligned} \tag{26}$$

$$-\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i-\bar{y})+n(\bar{y}-\mu)^2}{\sigma^3} = 0 \tag{27}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu) \tag{28}$$