

Ordinary Least Squares and Maximum Likelihood Estimation

David Carlson

November 23, 2021

What are Regression Models Useful For

- To summarize data

What are Regression Models Useful For

- To summarize data
- To make predictions

What are Regression Models Useful For

- To summarize data
- To make predictions
- To measure causal effects

Causal Inference

- Randomized controlled experiments: Assignment into “treatment” and “control” groups is knowingly randomized

Causal Inference

- Randomized controlled experiments: Assignment into “treatment” and “control” groups is knowingly randomized
- Natural experiments: Assignment into “treatment” and “control” is as if randomized by nature

Causal Inference

- Randomized controlled experiments: Assignment into “treatment” and “control” groups is knowingly randomized
- Natural experiments: Assignment into “treatment” and “control” is as if randomized by nature
- Observational studies: We do not know how assignment into “treatment” and “control” was achieved

Non-Parametric Approaches

- Are smokers (T) more likely to develop lung cancer (Y)?

Non-Parametric Approaches

- Are smokers (T) more likely to develop lung cancer (Y)?
- We can turn this into a question about $p(Y|T, X)$, i.e., about a conditional probability

Non-Parametric Approaches

- Are smokers (T) more likely to develop lung cancer (Y)?
- We can turn this into a question about $p(Y|T, X)$, i.e., about a conditional probability
- Because the analysis is based on observational data, we need to control for potential confounders: “Age group” and “genetic marker,” both dichotomous variables (X)

Non-Parametric Approaches

- Are smokers (T) more likely to develop lung cancer (Y)?
- We can turn this into a question about $p(Y|T, X)$, i.e., about a conditional probability
- Because the analysis is based on observational data, we need to control for potential confounders: “Age group” and “genetic marker,” both dichotomous variables (X)
- This is now a question about $p(Y|T, X)$

Non-Parametric Approaches

- Are smokers (T) more likely to develop lung cancer (Y)?
- We can turn this into a question about $p(Y|T, X)$, i.e., about a conditional probability
- Because the analysis is based on observational data, we need to control for potential confounders: “Age group” and “genetic marker,” both dichotomous variables (X)
- This is now a question about $p(Y|T, X)$

Non-Parametric Approaches

- Are smokers (T) more likely to develop lung cancer (Y)?
- We can turn this into a question about $p(Y|T, X)$, i.e., about a conditional probability
- Because the analysis is based on observational data, we need to control for potential confounders: “Age group” and “genetic marker,” both dichotomous variables (X)
- This is now a question about $p(Y|T, X)$

Smokers			Non-smokers		
	Old	Young		Old	Young
GM	\hat{y}_1	\hat{y}_2	GM	\hat{y}_5	\hat{y}_6
$\sim GM$	\hat{y}_3	\hat{y}_4	$\sim GM$	\hat{y}_7	\hat{y}_8

Non-Parametric Approaches (cont.)

We need not make too many assumptions about how T affects Y after controlling for X . We could simply assume

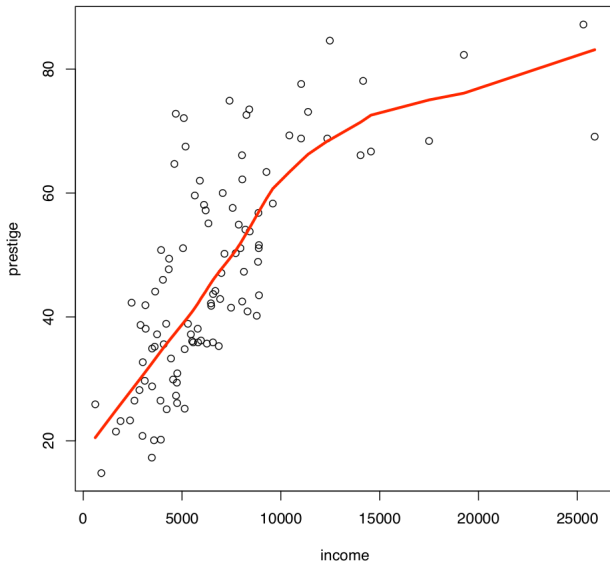
$$p(Y|T, X) = g(T, X)$$

and use sample data to estimate $p(Y|T, X)$ at different combinations of T and X .

Smokers			Non-smokers		
	Old	Young		Old	Young
GM	0.9	0.6	GM	0.5	0.5
$\sim GM$	0.6	0.6	$\sim GM$	0.5	0.5

Non-Parametric Approaches (cont.)

How can we learn anything about $p(Y|X)$ when X is continuous?



Non-Parametric Approaches (cont.)

- NPR depends on less extraneous assumptions (i.e., it does not require linearity), but

Non-Parametric Approaches (cont.)

- NPR depends on less extraneous assumptions (i.e., it does not require linearity), but
- NPR is computationally expensive

Non-Parametric Approaches (cont.)

- NPR depends on less extraneous assumptions (i.e., it does not require linearity), but
- NPR is computationally expensive
- NPR cannot be easily transmitted

Non-Parametric Approaches (cont.)

- NPR depends on less extraneous assumptions (i.e., it does not require linearity), but
- NPR is computationally expensive
- NPR cannot be easily transmitted
- NPR collapses under “curse of dimensionality”

Non-Parametric Approaches (cont.)

- NPR depends on less extraneous assumptions (i.e., it does not require linearity), but
- NPR is computationally expensive
- NPR cannot be easily transmitted
- NPR collapses under “curse of dimensionality”
- We need to move from “natural” NPR to parametric approaches

Substantive Statements as Probability Models

Outcome (Y)	Predictor or Cause (X)
Votes for Party A	Platforms of parties A and B
Frequency of wars	Political regimes of neighboring countries
Campaign spending (\$)	Incumbent strength
Survival of democracy	Country's income level
Cancer rates	Number of phone lines

Probability Models of Data Generating Processes

The **generalized linear model** notation makes it clear that we are building a model of the probability of Y conditional on X :

$$Y_i \sim f(\theta_i)$$

$$\theta_i = g(\mathbf{X}_i)$$

Probability Models of Data Generating Processes

The **generalized linear model** notation makes it clear that we are building a model of the probability of Y conditional on X :

$$Y_i \sim f(\theta_i)$$

$$\theta_i = g(\mathbf{X}_i)$$

- Stochastic component: $f(\cdot)$

Probability Models of Data Generating Processes

The **generalized linear model** notation makes it clear that we are building a model of the probability of Y conditional on X :

$$Y_i \sim f(\theta_i)$$

$$\theta_i = g(\mathbf{X}_i)$$

- Stochastic component: $f(\cdot)$
- Systematic component: \mathbf{X}

Probability Models of Data Generating Processes

The **generalized linear model** notation makes it clear that we are building a model of the probability of Y conditional on X :

$$Y_i \sim f(\theta_i)$$

$$\theta_i = g(\mathbf{X}_i)$$

- Stochastic component: $f(\cdot)$
- Systematic component: \mathbf{X}
- Link function: $g(\cdot)$

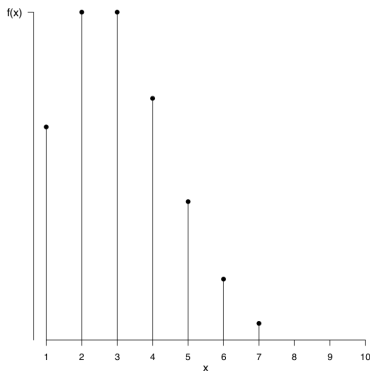
Distribution of Random Variables

- **Random variable:** A real-valued function that is defined on a sample space

Distribution of Random Variables

- **Random variable:** A real-valued function that is defined on a sample space
- Random variable X is characterized by a probability distribution over all possible values x that X can take

Discrete Random Variables



- ▶ The probability function of X is the function f such that for every x

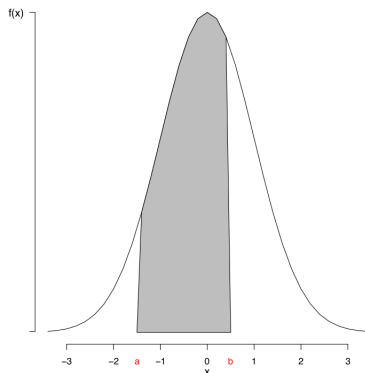
$$f(x) = \Pr(X = x)$$

1. If x is outside the sample space, then $f(x) = 0$
2. If x_1, x_2, \dots includes all values in the sample space, then

$$\sum_{i=1}^{\infty} f(x_i) = 1$$

3. $\Pr(X \in A) = \sum_{x_i \in A} f(x_i)$

Continuous Random Variables



- The probability density function $f(x)$ specifies the probability of X taking values on subsets of the sample space;
e.g., for subset (a, b)

1. $f(x) \geq 0, \forall x$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $\Pr(a < x \leq b) = \int_a^b f(x)dx$

Joint Distribution of Discrete X, Y

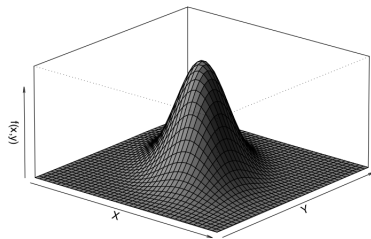
	X		
Y	1	2	3
1	0.1	0.3	0
2	0	0	0.2
3	0.1	0.1	0
4	0	0.2	0

- ▶ If X and Y are discrete random variables, their distribution is also discrete
- ▶ The joint p.f. of (X, Y) is the function f such that for every point (x, y)

$$f(x, y) = \Pr(X = x \text{ and } Y = y)$$

1. If x, y are outside the sample space, then
 $f(x, y) = 0$
2. $\sum_{(x,y)} f(x, y) = 1$

Joint Distribution of Continuous X, Y



- ▶ If X and Y are continuous random variables, their distribution is also continuous
- ▶ The joint pdf of (X, Y) is the function f such that for every region A

$$\Pr(X, Y \in A) = \int_A \int f(x, y) dx dy$$

1. $f(x, y) \geq 0$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) = 1$

Marginal Distribution of X

Y	X			$f_2(Y)$
	1	2	3	
1	0.1	0.3	0	0.4
2	0	0	0.2	0.2
3	0.1	0.1	0	0.2
4	0	0.2	0	0.2
$f_1(X)$	0.2	0.6	0.2	1

The distribution of X computed from the joint distribution of (X, Y) is the marginal distribution of X

- ▶ Discrete distributions:

$$f_1(x) = \sum_y f(x, y)$$

- ▶ Continuous distributions:

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

Conditional Distributions

After observing $Y = y$, the probability that $X = x$ is specified by the conditional probability

$$g_1(x|y) = p(X = x|Y = y) = \frac{p(X = x \text{ and } Y = y)}{p(Y = y)} = \frac{f(x, y)}{f(y)}$$

where $g_1(x|y) \geq 0$ and $\sum_y g_1(x|y) = 1$

Conditional Distributions (cont.)

What's $g_1(X = 1|Y = 3)$?

Y	X			$f_2(Y)$
	1	2	3	
1	0.1	0.3	0	0.4
2	0	0	0.2	0.2
3	0.1	0.1	0	0.2
4	0	0.2	0	0.2
$f_1(X)$	0.2	0.6	0.2	1

$$\frac{f(x = 1 \text{ and } y = 3)}{f_2(y = 3)} = \frac{0.1}{0.2}$$

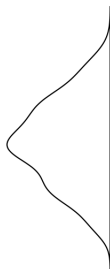
- ▶ $f_2(y = 3) = 0.2$
- ▶ $f(x = 1 \text{ and } y = 3) = 0.1$
- ▶ $f(x = 2 \text{ and } y = 3) = 0.1$
- ▶ $f(x = 3 \text{ and } y = 3) = 0$

$$= \frac{1}{2}$$

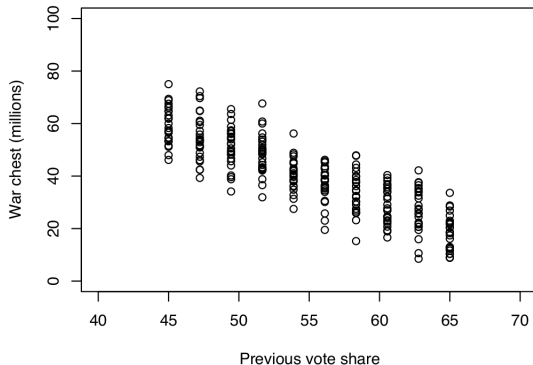
Linear Regression as a Probability Model

War chest as a function of *support in previous election* The regression line joins $E(Y|X)$ at different values of X

Marginal distribution



Conditional distribution



Alternative Notations for OLS Regression

GLM notation

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

$$\mu_i = \alpha + \beta X_i$$

OLS notation

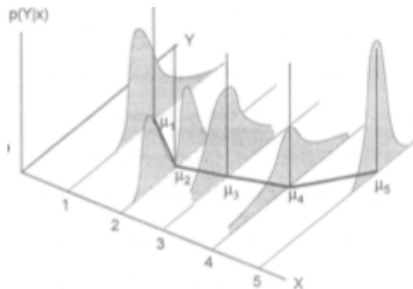
$$Y_i = \alpha + \beta X_i + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

In both cases, note the three main assumptions we impose:

- Normality
- Linearity
- Constant variance

Potential Pitfalls in Parametric Models



1. **Failure of normality:**
 $p(Y|x=1)$, $p(Y|x=2)$,
and $p(Y|x=3)$ are not
normal
2. **Failure of linearity:**
 $E[p(Y|X)] \neq \alpha + \beta X$
3. **Failure of constant variance:**
 $p(Y|x=4)$ and $p(Y|x=5)$
do not have the same spread:
 $(\sigma_y|x=4 \neq \sigma_y|x=5)$

Mechanics of OLS Point Estimates

- We build the line that best fits the data

Mechanics of OLS Point Estimates

- We build the line that best fits the data
- This line could be constructed in many different ways, but we make three consequential decisions:

Mechanics of OLS Point Estimates

- We build the line that best fits the data
- This line could be constructed in many different ways, but we make three consequential decisions:
 - ▶ Vertical distances

Mechanics of OLS Point Estimates

- We build the line that best fits the data
- This line could be constructed in many different ways, but we make three consequential decisions:
 - ▶ Vertical distances
 - ▶ Sum of errors

Mechanics of OLS Point Estimates

- We build the line that best fits the data
- This line could be constructed in many different ways, but we make three consequential decisions:
 - ▶ Vertical distances
 - ▶ Sum of errors
 - ▶ Sum of squared errors

Multiple Regression

- In multiple regression, we fit a least squares “hyperplane” in $k + 1$ -dimensional space

Multiple Regression

- In multiple regression, we fit a least squares “hyperplane” in $k + 1$ -dimensional space
- $\hat{\beta}_0$ is the expected value of Y when all variables X are jointly 0

Multiple Regression

- In multiple regression, we fit a least squares “hyperplane” in $k + 1$ -dimensional space
- $\hat{\beta}_0$ is the expected value of Y when all variables X are jointly 0
- For any slope coefficient estimate $\hat{\beta}_k$: A unit increase in $X_{i,k}$ will yield on average a $\hat{\beta}_k$ increase in Y_i , *all else constant*

Multiple Regression (cont.)

We define the multiple regression model as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

In this model, we have:

- $i = 1, \dots, n$ observations
- k independent variables
- $k + 1$ slope and intercept parameters β_j
- one variance parameter σ^2

Statistical Theory for Linear Models

The scalar notation is cumbersome, but the model can be simplified by defining the following vectors and matrices:

- $\mathbf{y} = [y_1, y_2, \dots, y_n]'$ is a vector of observations on the dependent variable
- $\mathbf{X} = [\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k]$ is a matrix with a column of 1's and k columns of independent variables
- $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \dots, \beta_k]'$
- $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \dots, \epsilon_n]'$ is a vector of random errors

Statistical Theory for Linear Models (cont.)

The linear model can then be represented succinctly as:

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ E(\boldsymbol{\epsilon}) &= [0, 0, \dots, 0]' = \mathbf{0} \\ \text{var}(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') &= \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}_n\end{aligned}$$

Statistical Theory for Linear Models (cont.)

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

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{n \times 1} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}}_{n \times (k+1)} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}}_{(k+1) \times 1} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{n \times 1}$$

For observation 2:

$$y_2 = \beta_0 \cdot 1 + \beta_1 x_{2,1} + \beta_2 x_{2,2} + \dots + \beta_k x_{2,k} + \epsilon_2$$

Derivation of OLS Estimators

- Define the vector of residuals as

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b}$$

- The sum of squared errors is defined as

$$\mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b})$$

- Compute $\frac{\partial}{\partial \mathbf{b}}(\mathbf{e}'\mathbf{e})$

$$\begin{aligned}\frac{\partial}{\partial \mathbf{b}}(\mathbf{e}'\mathbf{e}) &= \frac{\partial}{\partial \mathbf{b}}(\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b}) \\ &= \frac{\partial}{\partial \mathbf{b}}(\mathbf{Y}' - \mathbf{b}'\mathbf{X}')(\mathbf{Y} - \mathbf{X}\mathbf{b}) \\ &= \frac{\partial}{\partial \mathbf{b}}(\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{Y}' + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}) \\ &= \frac{\partial}{\partial \mathbf{b}}(\mathbf{Y}'\mathbf{Y} - 2\mathbf{b}'\mathbf{X}'\mathbf{Y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}) \\ &= -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}\end{aligned}$$

Derivation of OLS Estimators (cont.)

- Set the first derivative of $\mathbf{e}'\mathbf{e}$ with respect to \mathbf{b} equal to 0

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

$$2\mathbf{X}'\mathbf{X}\mathbf{b} = 2\mathbf{X}'\mathbf{Y}$$

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

- Compute the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ of $\mathbf{X}'\mathbf{X}$ and use it to pre-multiply both sides of the previous equation:

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

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

- \mathbf{b} is uniquely defined as long as $\mathbf{X}'\mathbf{X}$ is a full-rank matrix

Derivation of OLS Estimators (cont.)

Second order condition: The matrix of second derivatives of $\mathbf{e}'\mathbf{e}$ (Hessian matrix) should be positive definite, hence a global minimum:

$$\frac{\partial^2(\mathbf{e}'\mathbf{e})}{\partial \mathbf{b} \partial \mathbf{b}'} = \begin{bmatrix} \frac{\partial^2 \mathbf{e}'\mathbf{e}}{\partial \mathbf{b}_0^2} & \frac{\partial^2 \mathbf{e}'\mathbf{e}}{\partial \mathbf{b}_0 \partial \mathbf{b}_1} & \frac{\partial^2 \mathbf{e}'\mathbf{e}}{\partial \mathbf{b}_0 \partial \mathbf{b}_2} & \cdots & \frac{\partial^2 \mathbf{e}'\mathbf{e}}{\partial \mathbf{b}_0 \partial \mathbf{b}_k} \\ \frac{\partial^2 \mathbf{e}'\mathbf{e}}{\partial \mathbf{b}_0 \partial \mathbf{b}_1} & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 \mathbf{e}'\mathbf{e}}{\partial \mathbf{b}_0 \partial \mathbf{b}_n} & \cdots & \cdots & \cdots & \frac{\partial^2 \mathbf{e}'\mathbf{e}}{\partial \mathbf{b}_k^2} \end{bmatrix} = 2\mathbf{X}'\mathbf{X}$$

Derivation of the Moments of \mathbf{b}

We can also find the variance of \mathbf{b} :

$$\begin{aligned}\text{var}(\mathbf{b}) &= \text{var}(\mathbf{A}\mathbf{Y}) \\ &= \mathbf{A}\text{var}(\mathbf{Y})\mathbf{A}' \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\text{var}(\mathbf{Y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\sigma^2\mathbf{I}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Basic Assumptions of OLS (“fixed X” approach)

- No multicollinearity

Basic Assumptions of OLS (“fixed X” approach)

- No multicollinearity
- Variability in X: $\text{var}(X) > 0$ but finite

Basic Assumptions of OLS (“fixed X” approach)

- No multicollinearity
- Variability in X: $\text{var}(X) > 0$ but finite
- Linearity: $Y_i = \alpha + \beta X_i + \epsilon_i$

Basic Assumptions of OLS (“fixed X” approach)

- No multicollinearity
- Variability in X: $\text{var}(X) > 0$ but finite
- Linearity: $Y_i = \alpha + \beta X_i + \epsilon_i$
- Zero mean: $E(\epsilon_i) = 0$

Basic Assumptions of OLS (“fixed X” approach)

- No multicollinearity
- Variability in X: $\text{var}(X) > 0$ but finite
- Linearity: $Y_i = \alpha + \beta X_i + \epsilon_i$
- Zero mean: $E(\epsilon_i) = 0$
- Homoskedasticity: $\text{var}(\epsilon_i) = \sigma^2$

Basic Assumptions of OLS (“fixed X” approach)

- No multicollinearity
- Variability in X: $\text{var}(X) > 0$ but finite
- Linearity: $Y_i = \alpha + \beta X_i + \epsilon_i$
- Zero mean: $E(\epsilon_i) = 0$
- Homoskedasticity: $\text{var}(\epsilon_i) = \sigma^2$
- Non-autocorrelation: $\text{cov}(\epsilon_i, \epsilon_j) = 0, \forall i \neq j$

Basic Assumptions of OLS (“fixed X” approach)

- No multicollinearity
- Variability in X: $\text{var}(X) > 0$ but finite
- Linearity: $Y_i = \alpha + \beta X_i + \epsilon_i$
- Zero mean: $E(\epsilon_i) = 0$
- Homoskedasticity: $\text{var}(\epsilon_i) = \sigma^2$
- Non-autocorrelation: $\text{cov}(\epsilon_i, \epsilon_j) = 0, \forall i \neq j$
- Non-stochastic X: X is fixed in repeated sampling

Basic Assumptions of OLS (“conditional” approach)

- No perfect multicollinearity
- Variability in X : $\text{var}(X) > 0$ but finite
- Linearity: $Y_i = \alpha + \beta X_i + \epsilon_i$
- Conditional zero mean: $E(\epsilon_i | \mathbf{X}) = 0$
- Conditional homoskedasticity: $\text{var}(\epsilon_i | \mathbf{X}) = \sigma^2$
- Conditional non-autocorrelation: $\text{cov}(\epsilon_i, \epsilon_j | \mathbf{X}) = 0, \forall i \neq j$

Maximum-Likelihood Estimation

- We use data to make inferences about a set θ of parameters (ex., $\theta = (\beta_0, \beta_1, \dots, \beta_k)$)
- We observe

$$\mathbf{Y} = (y_1, y_2, \dots, y_n)'$$

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k)$$

and assume

- ▶ that each draw y_i is drawn from the same distribution with parameter θ and
- ▶ that the draws $i = (1, \dots, n)$ are independent
- In short: $y_i \stackrel{iid}{\sim} f(\theta, \mathbf{X}_i)$

Maximum-Likelihood Estimation (cont.)

- The MLE $\hat{\theta}_{ML}$ of parameter θ is the value of θ that is most likely to have produced the observed sample (given assumed model)

Maximum-Likelihood Estimation (cont.)

- The MLE $\hat{\theta}_{ML}$ of parameter θ is the value of θ that is most likely to have produced the observed sample (given assumed model)
- For example, which parameter θ is most likely to have generated the value $y = 3$ if $Y \sim \mathcal{N}(\theta, \sigma^2)$?

Maximum-Likelihood Estimation (cont.)

- The MLE $\hat{\theta}_{ML}$ of parameter θ is the value of θ that is most likely to have produced the observed sample (given assumed model)
- For example, which parameter θ is most likely to have generated the value $y = 3$ if $Y \sim \mathcal{N}(\theta, \sigma^2)$?
- Which parameters $\theta = (\mu, \sigma^2)$ are most likely to have generated the values $\mathbf{y} = (3, 5, 7)$ if $Y \sim \mathcal{N}(\mu, \sigma^2)$?

Maximum-Likelihood Estimation (cont.)

- The MLE $\hat{\theta}_{ML}$ of parameter θ is the value of θ that is most likely to have produced the observed sample (given assumed model)
- For example, which parameter θ is most likely to have generated the value $y = 3$ if $Y \sim \mathcal{N}(\theta, \sigma^2)$?
- Which parameters $\theta = (\mu, \sigma^2)$ are most likely to have generated the values $\mathbf{y} = (3, 5, 7)$ if $Y \sim \mathcal{N}(\mu, \sigma^2)$?
- Which parameters $\theta = (\mu, \sigma^2)$ are most likely to have generated the values $\mathbf{y} = (3, 5, 7, 6, 1, 2, 4, 5, 5, 9, 8)$ if $Y \sim \mathcal{N}(\mu, \sigma^2)$?

Maximum-Likelihood Estimation (cont.)

- We can rewrite the linear regression model as a probability model:

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$$

- Which means that, for any observation:

$$y_i \sim \mathcal{N}(\mathbf{x}_i' \boldsymbol{\beta}, \sigma^2)$$

$$p(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left[-\frac{1}{2\sigma^2} (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 \right]$$

MLE for the Linear Regression Model

Because of our assumption of independence, we can write the joint pdf of \mathbf{Y} as

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) &= p(y_1) \times p(y_2) \times \cdots \times p(y_n) = \prod_{i=1}^n p(y_i) \\ &= \prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left[-\frac{1}{2\sigma^2} (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[-\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \right] \end{aligned}$$

MLE for the Linear Regression Model (cont.)

When the parameters β, σ^2 are expressed as a function of the data \mathbf{Y}, \mathbf{X} , the joint pdf is called the likelihood function:

$$\mathcal{L}(\beta, \sigma^2 | \mathbf{Y}, \mathbf{X}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[-\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\beta)' (\mathbf{Y} - \mathbf{X}\beta) \right]$$

- Note that $\mathcal{L}(\cdot)$ returns a real number for every combination of $\mathbf{Y}, \mathbf{X}, \beta, \sigma^2$
- Maximum likelihood estimates of β are simply the values of $\mathbf{b}, \hat{\sigma}^2$ that maximize $\mathcal{L}(\cdot)$ given \mathbf{Y}, \mathbf{X}

MLE for the Linear Regression Model (cont.)

- In the linear model, $\mathcal{L}(\cdot)$ is a concave function (thus \mathbf{b} and $\hat{\sigma}^2$ exist and are unique) but has a very difficult form
- Fortunately, we can appeal to the invariance property of maximum likelihood estimators to transform $\mathcal{L}(\cdot)$ into something manageable, like the log (ℓ):

$$\begin{aligned}\ell(\boldsymbol{\beta}, \sigma^2 | \mathbf{Y}, \mathbf{X}) &= \ln \mathcal{L}(\cdot) \\ &= \ln \left(\frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[-\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \right] \right) \\ &= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\end{aligned}$$

MLE for the Linear Regression Model (cont.)

- We can now solve the (relatively) simpler problem of choosing β and σ^2 to maximize ℓ :

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{2\sigma^2} (2\mathbf{X}'\mathbf{Y} - 2\mathbf{X}'\mathbf{X}\beta) = 0$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) = 0$$

- The ML estimates are:

$$\mathbf{b}_{ML} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$\hat{\sigma}_{ML}^2 = \frac{(\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b})}{n} = \frac{\mathbf{e}'\mathbf{e}}{n}$$

Dichotomous Dependent Variables: LPM

The linear probability model is based on the assumption of linearity to model dichotomous dependent variables

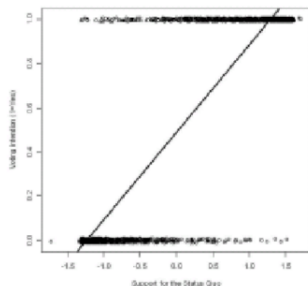
$$Y_i = \underbrace{\alpha + \beta x_i}_{E(Y|x_i)=\pi_i} + \varepsilon_i$$
$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

LPM is BLAT, but:

1. *Errors are not normally distributed*
2. *Heteroskedasticity*

$$\text{var}(\varepsilon_i) = \pi_i(1 - \pi_i)$$

3. *Linearity assumption*



Generalized Linear Models

So far, we have discussed the identity link, but many possible options for $g(\cdot)$

- Binary (dichotomous)
 - ▶ Logit, Probit, or c-log-log
- Unordered categorical (polytomous)
 - ▶ Multinomial logit
- Ordered categorical
 - ▶ Ordered logit or Probit
- Counts
 - ▶ Poisson
- Rare counts
 - ▶ Negative binomial
- Zero-inflated models for rare non-zero outcomes (e.g. war)