Econometrics **August 18, 2026**

OLS Mechanics

Instructor: Kalyan Kolukuluri *Content: Lecture Notes*

Abstract

These lecture notes examine how Ordinary Least Squares (OLS) regression parameters relate to conditional group means when regressors take discrete values. We first analyze the binary regressor case $X_1 \in \{0,1\}$, demonstrating that linear OLS perfectly fits conditional group means. We then analyze a multi-valued discrete regressor $X_1 \in \{0,1,2\}$, deriving how linear OLS imposes a constant marginal effect constraint across groups. Finally, we formally introduce the concept of a **saturated model**, proving why dummy indicator specifications recover the exact Nonparametric Conditional Expectation Function (CEF) without functional form bias (Angrist & Pischke, 2009; Wooldridge, 2010). Complete Monte Carlo R code, color-coded sample data tables, formatted regression output tables, and step-by-step arithmetic breakdowns are provided throughout.

1 Classical Linear Regression Model

Population regression function (PRF) = $\mathbb{E}(Y_i \mid X_i) = \beta_0 + \beta_1(X_i)$
It is also called the *CEF (Conditional Expectation Function)*.

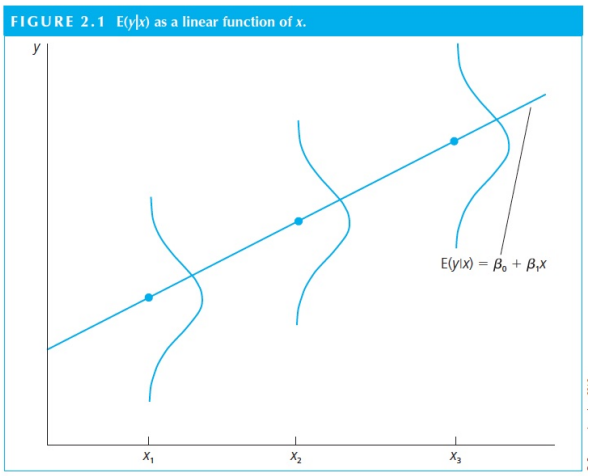
1.1 What is random in a linear regression?

Regression can be performed on a population. However, we always have a sample to work with. So, we have a sample equivalent called the *Sample Regression Function (SRF)*.

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1(X_i) + \hat{\varepsilon}_i$$

Evidently, the estimates $\hat{\beta}_0, \hat{\beta}_1$ are unique to a sample; with redraws of sample, there will be different $\hat{\beta}_0, \hat{\beta}_1$ estimates. In this light, what is random? It is the sampling variation that causes different values of estimates for the SRF parameters. Thus, SRF parameters, $\hat{\beta}_0, \hat{\beta}_1$ are random.

Figure 1: Population regression function



Population data is in grey. Dashed grey line indicates population regression function (β_0, β_1). Red and purple data is from two random samples from the population. Red and Purple lines have different intercepts and slopes (sample estimates are say, $\hat{\beta_1^{red}}, \hat{\beta_1^{purple}}$).

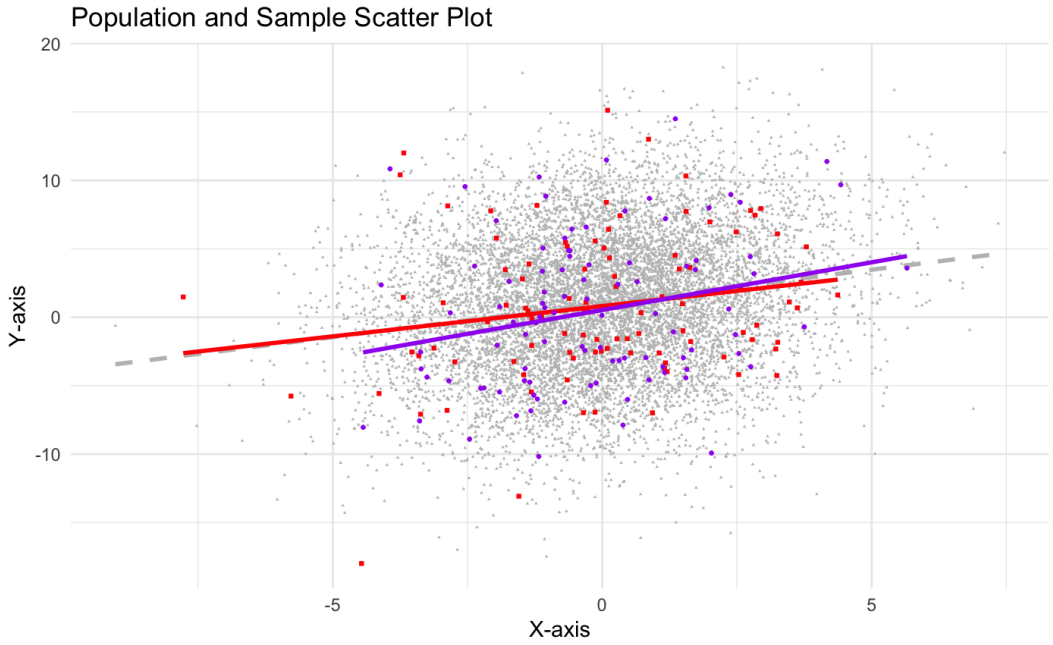
Linearity in Parameters

Linear regression refers to a specification that is linear in the parameters.

$$\left.\begin{array}{l} \mathbb{E}\left[Y_i \mid X_i\right] = \alpha + \beta X_i^2 \\ \mathbb{E}\left[\log Y_i \mid X_i\right] = \alpha + \beta\left(Y_i\right) \end{array}\right\} \text{Both specify linear regressions.}$$

But, $\mathrm{E}[Y_i \mid X_i] = \alpha + \beta^2(X_i)$ is not a linear model.

Figure 2: Population β and Sample $\hat{\beta}$



Claim: Regression is a useful tool regardless of whether the underlying CEF is linear
To understand this claim, we need two theoretical properties:

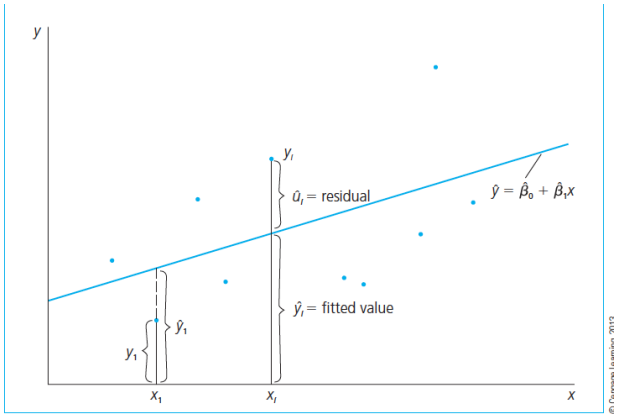
1. If $\mathbb{E}[Y_i \mid X_{1i}, X_{2i}, \cdots X_{Ki}] = \beta_0 + \sum_{k=1}^{K} \beta_{1k} X_{ki}$ then the regression of Y_i on $X_{1i}, X_{2i}, \cdots X_{Ki}$ has intercepts β_0 , and slopes $\beta_{11} \cdots \beta_{1k}$. In other words, if the underlying CEF is linear then the regression of Y_i on $X_{1i}, X_{2i}, \cdots X_{Ki}$ is it.

2. If $\mathbb{E}[Y_i \mid X_{1i}, X_{2i}, \cdots X_{Ki}]$ is non-linear, then the regression of Y_i on $X_{1i}, X_{2i}, \cdots X_{Ki}$ gives the best linear approximation to this non-linear CEF. It minimizes the expected squared deviation between the fitted values from a linear model and the true CEF. The second property tells us that we can expect regression estimates of a treatment effect to be close to those we would get by matching on covariates and then averaging within-cell treatment and control differen despite the true CEF not being linear.

1.2 Parameter Estimates and Variances - ($\hat{\beta}_0, \hat{\beta}_1$)

1. $\hat{\beta}_1 = \dfrac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2} = \dfrac{\text{Cov}\left(X_i, Y_i\right)}{\text{Var}\left(X_i\right)}$
2. $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1(\overline{X})$

Figure 3: Population regression function



3. $\text{Var}(\hat{\beta}_1) = \dfrac{\sigma_\varepsilon^2}{\sum(X_i - \overline{X})^2}$. Given σ_ε^2, larger variation in X $\implies$ smaller $\text{Var}(\beta_1)$. As Sample size (N) increases, $\text{Var}(\hat{\beta}_1)$ is smaller.
4. $\text{Var}(\hat{\beta}_0) = \dfrac{\sum(X_i^2)}{N\sum(X_i - \overline{X})^2}\sigma_\varepsilon^2$
5. $\text{Cov}(\beta_1, \beta_0) = \dfrac{-\overline{X}\sigma_\varepsilon^2}{\sum(X_i - \overline{X})^2}$.

Sample data has all information to calculate sample based parameter estimates; except population variance σ_ε^2.

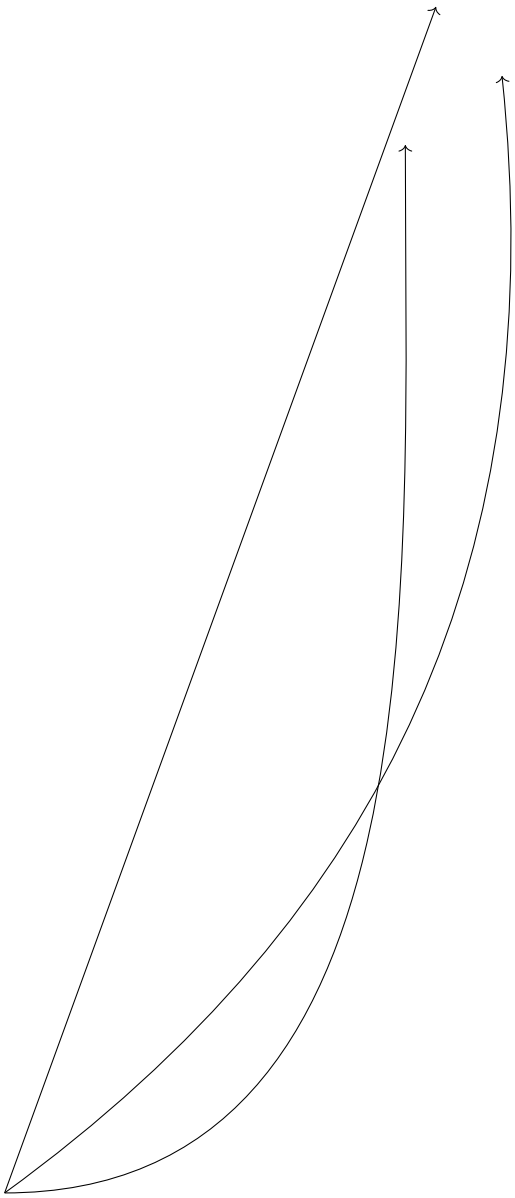
So, to calculate the above values we need to estimate σ_ε^2. $\mathbb{E}(\sigma_\varepsilon^2) = \dfrac{\sum \hat{\varepsilon_i^2}}{N-2}$ or $\dfrac{\sum \hat{\varepsilon_i^2}}{N-K}$ in general case.

2 Partitioning variation- OLS

- Total Sum of Squares (TSS) = Estimated SS (ESS) + Residual SS (RSS) (Gujarati et. al)
- Some other books indicate this as: Total Sum of Squares (SST) = Explained Sum of Squares (SSE) + Residual Sum of Squares (SSR)
- $\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{N}\left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2$

$$\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{N}\left(\hat{Y}_i-\overline{Y}\right)^2+\sum_{i=1}^{N}\left(Y_i-\hat{Y}_i\right)^2$$

- Total Sum of Squares (TSS)
- Estimated Sum of Squares (ESS)
- Residual Sum of Squares



2.1 Linear Regression Table

Simple Linear Regression: $Y_i = \beta_0 + \beta_1(X_i) + \varepsilon_i$
In linear regression, β_0 is the same as grand mean $\overline{\overline{Y}}$ in the ANOVA case.

Source of variation	D.O.F	Sum of Squares	Mean Square	F-statistic
Regression	1	ESS = $\frac{S_{xy}^2}{S_{xx}}$	Mean ESS = ESS/1	F = Mean ES-S/Mean RSS
Residual	N-2[1]	RSS = $\sum_{i=1} \hat{\epsilon_i}^2$	Mean RSS = RSS /(N-2)	
Total pooled variation	N-1	TSS = $\sum_{i=1} \left(Y_i - \overline{Y}\right)^2$		

2.2 CLRM - Assumptions

CLRM Assumptions of the Population Linear Model

1. **MLR A1:** Linear regression model. The regression model is linear in the parameters
2. **MLR A2:** X values are fixed in repeated sampling. i.e. X is *non-stochastic.*
3. **MLR A3: No perfect multicollinearity:** NO perfect linear relationship between X's.
4. **MLR A4:** Zero mean value of disturbance $\mathbb{E}(\varepsilon_i \mid X_i) = 0 \implies \text{Cov}(X_i, \varepsilon_i) = 0$. No specification bias or specification error in the model.
5. **MLR A5:** *Homoskedasticity* $\text{Var}(\varepsilon_i|X_i) = \text{Var}(\varepsilon_i) = \sigma_\varepsilon^2$ Not all Y's are equally *reliable*. So, we woould prefer to sample X's close to $E(Y_i|X_i)$ and this will induce higher $\text{var}(\hat{\beta}_1)$
6. **MLR A6: Normality of** ε_i. Since distribution of β_0, β_1 is dependent on distribution of ε_i, any hypothesis tests are only valid as long as $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$
7. **MLR A7: No autocorrelation** between the disturbances. $\text{Cov}(\varepsilon_i, u_j|X_i, X_j) = 0$ The observations are sampled independently [2]

[2] Challenging for Time Series data

2.3 Gauss Markov Theorem

Theorem 1 (Gauss Markov Theorem) *Given the assumptions of the CLRM, the least squares estimators in the class of unbiased linear estimators have minimum variance, i.e., they are **BLUE**.*

2.4 Prediction - CI

Mean Prediction

At some instance of the regressor (X_0)

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta} X_0$$

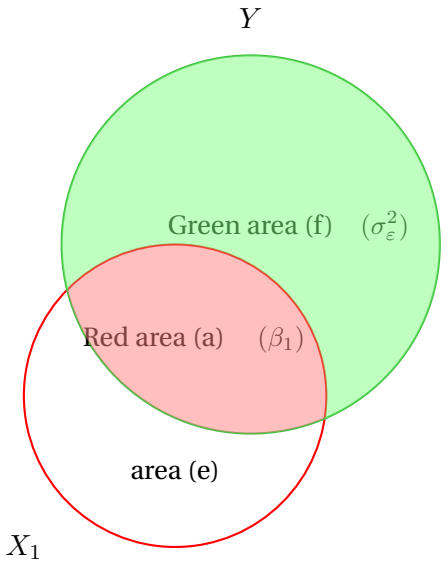
$$\text{Var}\left(\hat{Y}_0\right) = \sigma^2 \left[\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum X_i^2}\right]$$

So, the variance of the mean prediction $\text{Var}(\hat{Y}_0)$ increases as one moves away from sample mean $\overline{X}$.

Individual prediction

$$\text{Var}\left(Y_0 - \hat{Y}_0\right) = \sigma^2 \left[1 + \frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum X_i^2}\right]$$

3 Visualizing Regression (Venn Diagrams/ Ballentines)



3.1 Interpreting Ballentines

Econometrics Concept: Ballentine Diagrams

A **Ballentine Diagram** uses overlapping circles to visually represent variation:

- **Each Individual Circle:** Represents the total variance of a single variable (e.g., Y, X_1, X_2).
- **Overlapping Areas:** Represent the covariance (shared variance) between those variables.
- **OLS Regression Mechanics:**
  - The portion of the Y circle covered by X_1 and X_2 combined represents the R^2 (explained variance).
  - The portion of Y left uncovered represents the **residual variance** (error term, u).
- **Multicollinearity:** If the circles for X_1 and X_2 overlap heavily with each other inside Y, that shared zone represents variation that cannot be uniquely assigned to either regressor. This causes larger standard errors.

Popularized by (**??**).

1. $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$
2. Circular area of a Ballentine: Variation in a variable (Example: The green outlined circle= Red + Green area represents variation in Y)
3. Common overlap area: How much does variation in one variable is explained by variation in the other? (Example: Red shaded area represents variation in Y explained exclusively by X_1). *Area of overlap (a), is used to estimate the magnitude of* β_1. *The larger this overlap area, the more information is used to estimate* β_1, *implying smaller* $\text{Var}(\beta_1)$
4. Area with no overlap in Y: Variation in Y that cannot be explained by any other variable. *Magnitude of this area (area (f)) represents magnitude of the estimate of* σ^2_ε .
5. Coefficient of Determination (R^2): The Red area over the (Red+ Green) areas is the Explained Sum of Squares to Total Sum of Squares $R^2 = \dfrac{ESS}{TSS}$. The ratio $\dfrac{a}{a+f} = R^2$ the coeff. of determination.

4 OLS Mechanics

In empirical economics, regressors frequently take discrete values representing treatment categories, education levels, or policy tiers. Understanding how OLS aggregates group-specific outcomes ($\overline{Y}_k$) into intercept ($\hat{\beta}_0$) and slope ($\hat{\beta}_1$) estimates is fundamental to structural interpretation and causal inference (Greene, 2018; Stock & Watson, 2020).

5 Case 1: Binary Regressor ($X_1 \in \{0, 1\}$)

5.1 Mathematical Derivation

Consider the univariate OLS model where the regressor X_1 takes binary values:

$$Y_i = \beta_0 + \beta_1 X_{1i} + e_i, \quad X_{1i} \in \{0, 1\} \tag{1}$$

Let n_0 be the sample size of Group 0 ($X_1 = 0$) and n_1 be the sample size of Group 1 ($X_1 = 1$), with sample means $\overline{Y}_0 = \dfrac{1}{n_0} \sum_{i:X_{1i}=0} Y_i$ and $\overline{Y}_1 = \dfrac{1}{n_1} \sum_{i:X_{1i}=1} Y_i$.

Taking conditional sample expectations:

$$\hat{E}[Y_i \mid X_{1i} = 0] = \hat{\beta}_0 = \overline{Y}_0 \tag{2}$$

$$\hat{E}[Y_i \mid X_{1i} = 1] = \hat{\beta}_0 + \hat{\beta}_1 = \overline{Y}_1 \tag{3}$$

Solving for the slope coefficient $\hat{\beta}_1$ yields:

$$\hat{\beta}_1 = \overline{Y}_1 - \overline{Y}_0 \tag{4}$$

Thus, when X_1 is binary, OLS fits a straight line that connects $(0, \overline{Y}_0)$ and $(1, \overline{Y}_1)$ (Stock & Watson, 2020).

Key Result: Binary OLS

- **Intercept** ($\hat{\beta}_0$): The mean outcome of the control/reference group ($\overline{Y}_0$).
- **Slope** ($\hat{\beta}_1$): The exact difference in sample averages between treatment and control groups ($\overline{Y}_1 - \overline{Y}_0$).
- **Residuals** (e_i): Average to zero within **each** group separately ($\sum_{i:X_1=0} e_i = 0$ and $\sum_{i:X_1=1} e_i = 0$).

5.2 R Simulation & Data Generation

```r
# Set seed for reproducibility
set.seed(12345)

n0 <- 10
n1 <- 10

# Simulating Y from N(mu, sigma^2)
Y_group0 <- rnorm(n0, mean = 3.0, sd = 0.8)
Y_group1 <- rnorm(n1, mean = 8.0, sd = 0.8)

df_binary <- data.frame(
  X1 = c(rep(0, n0), rep(1, n1)),
  Y  = c(Y_group0, Y_group1)
)

# Fit OLS
ols_binary <- lm(Y ~ X1, data = df_binary)
summary(ols_binary)
```

Listing 1: R Simulation: Binary Regressor ($X_1 \in \{0, 1\}$)

Color-Coded Data Table and Regression Output Table 1 presents the individual observations generated by Listing 1. Rows in **Blue** correspond to $X_1 = 0$ (Control) and rows in **Green** correspond to $X_1 = 1$ (Treatment).

Step-by-Step Arithmetic Breakdown (Binary Case)

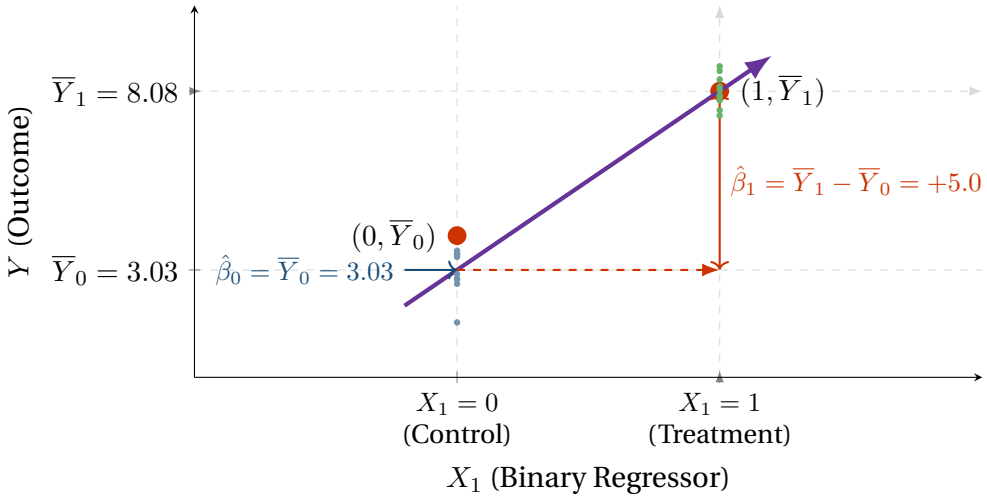
From the simulated output generated by Listing 1 and Table 1:

1. **Group 0 Mean**: $\overline{Y}_0 = \dfrac{30.312}{10} = \mathbf{3.0312}$
2. **Group 1 Mean**: $\overline{Y}_1 = \dfrac{80.784}{10} = \mathbf{8.0784}$
3. **Intercept Estimation**: $\hat{\beta}_0 = \overline{Y}_0 = \mathbf{3.0312}$
4. **Slope Estimation**: $\hat{\beta}_1 = \overline{Y}_1 - \overline{Y}_0 = 8.0784 - 3.0312 = \mathbf{5.0472}$
5. **Verification**: Table 2 yields $\hat{\beta}_0 = 3.0312$ and $\hat{\beta}_1 = 5.0472$. Within-group residual sums are identically 0.0000.

5.3 Visual Illustration

Figure 5 illustrates how the linear OLS line passes through the group centroids $(0, \overline{Y}_0)$ and $(1, \overline{Y}_1)$.

Figure 5: OLS Regression with a Binary Regressor ($X_1 \in \{0, 1\}$). The regression line passes directly through both group sample means.



6 Case 2: Multi-valued Discrete Regressor ($X_1 \in \{0, 1, 2\}$)

6.1 Mathematical Derivation under Linear OLS

Now consider a discrete regressor that takes three ordered values: $X_1 \in \{0, 1, 2\}$. If we estimate a single bivariate linear model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + e_i \tag{5}$$

For balanced sample sizes ($n_0 = n_1 = n_2 = n$), the sample means are $\overline{X} = 1$ and $\overline{Y} = \dfrac{\overline{Y}_0 + \overline{Y}_1 + \overline{Y}_2}{3}$.

Applying the standard OLS formula for $\hat{\beta}_1$:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{k=0}^{2} n_k (X_k - \overline{X})(\overline{Y}_k - \overline{Y})}{\sum_{k=0}^{2} n_k (X_k - \overline{X})^2} \\
&= \frac{n(-1)(\overline{Y}_0 - \overline{Y}) + n(0)(\overline{Y}_1 - \overline{Y}) + n(1)(\overline{Y}_2 - \overline{Y})}{n(-1)^2 + n(0)^2 + n(1)^2} \\
&= \frac{n(\overline{Y}_2 - \overline{Y}_0)}{2n} = \frac{\mathbf{\overline{Y}_2 - \overline{Y}_0}}{\mathbf{2}}
\end{aligned} \tag{6}$$

Key Result: Multi-valued Linear OLS

- **Middle Group Neutrality**: When X_1 is symmetric around 1 and sample sizes are equal, the middle group mean $\overline{Y}_1$ **drops out completely** from the slope estimation $\hat{\beta}_1 = \dfrac{\overline{Y}_2 - \overline{Y}_0}{2}$.
- **Constant Marginal Effect Constraint**: Linear OLS forces $E[Y \mid X_1 = 1] - E[Y \mid X_1 = 0] = E[Y \mid X_1 = 2] - E[Y \mid X_1 = 1] = \hat{\beta}_1$.
- **Potential Structural Bias**: If the true Nonparametric Conditional Expectation Function (CEF) is non-linear (e.g., concave or convex), linear OLS cannot fit all three centroids, leading to non-zero average residuals within specific groups.

6.2 R Simulation: Linear vs. Saturated Specifications

```r
# Set seed
set.seed(12345)

n <- 10  # 10 subjects per group

```

```r
# Simulating Non-linear Data: Large jump at X1=1, flattening at X1=2
Y_g0 <- rnorm(n, mean = 2.0, sd = 0.7)
Y_g1 <- rnorm(n, mean = 9.0, sd = 0.7)  # Non-linear hump
Y_g2 <- rnorm(n, mean = 10.0, sd = 0.7)

df_multi <- data.frame(
  X1 = c(rep(0, n), rep(1, n), rep(2, n)),
  Y  = c(Y_g0, Y_g1, Y_g2)
)

# Linear OLS Model
ols_linear <- lm(Y ~ X1, data = df_multi)

# Saturated Model (Dummy Variable Specification)
df_multi$X1_factor <- factor(df_multi$X1)
ols_saturated <- lm(Y ~ X1_factor, data = df_multi)
```

Listing 2: R Simulation: Multi-Valued Regressor ($X_1 \in \{0, 1, 2\}$)

Color-Coded Comparative Observations Table Table 3 presents the full simulated dataset ($N = 30$) color-coded by group: Group 0 (Blue), Group 1 (Green), and Group 2 (Red). Notice how Linear OLS creates systemic residuals in Group 1 ($e^{\text{lin}} = +2.0140$), whereas the Saturated Model eliminates group-level residuals entirely ($e^{\text{sat}} = 0.0000$).

6.3 Comparative Regression Output Table

Table 4 highlights the contrast between the linear OLS specification and the saturated dummy model for $X_1 \in \{0, 1, 2\}$.

Step-by-Step Arithmetic Breakdown (Multi-Valued Case)

From Listing 2, Table 3, and Table 4:

1. **Group Sample Means**:

$$\overline{Y}_0 = \mathbf{2.0273}, \quad \overline{Y}_1 = \mathbf{9.0686}, \quad \overline{Y}_2 = \mathbf{10.0686} \tag{7}$$

2. **Linear OLS Slope Calculation**:

$$\hat{\beta}_1 = \frac{\overline{Y}_2 - \overline{Y}_0}{2} = \frac{10.0686 - 2.0273}{2} = \mathbf{4.0207} \tag{8}$$

   Notice that $\overline{Y}_1 = 9.0686$ is completely ignored in determining the linear slope!

3. **Linear OLS Intercept Calculation**:

$$\hat{\beta}_0 = \frac{5\overline{Y}_0 + 2\overline{Y}_1 - \overline{Y}_2}{6} = \frac{5(2.0273) + 2(9.0686) - 10.0686}{6} = \mathbf{3.0339} \tag{9}$$

4. **Systemic Non-linearity Residuals in Linear OLS**:

   - Group 0 Average Residual: $2.0273 - 3.0339 = \mathbf{-1.0066}$
   - Group 1 Average Residual: $9.0686 - 7.0546 = \mathbf{+2.0140}$ (Major misfit!)
   - Group 2 Average Residual: $10.0686 - 11.0753 = \mathbf{-1.0067}$

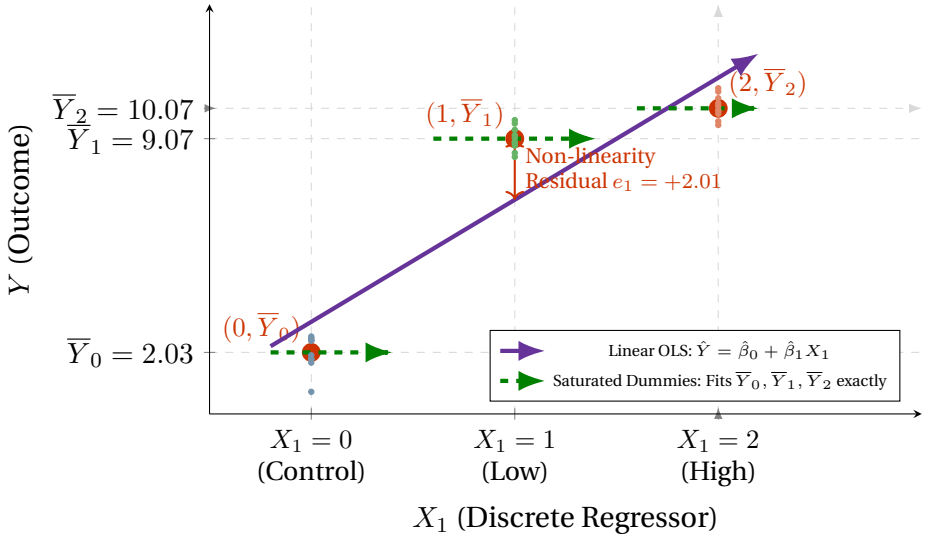
5. **Saturated Dummy Model Solution** (`lm(Y ~ factor(X1))`):

   - Intercept $\hat{\gamma}_0 = \overline{Y}_0 = \mathbf{2.0273}$
   - Dummy D_1 coefficient $\hat{\gamma}_1 = \overline{Y}_1 - \overline{Y}_0 = 9.0686 - 2.0273 = \mathbf{7.0413}$
   - Dummy D_2 coefficient $\hat{\gamma}_2 = \overline{Y}_2 - \overline{Y}_0 = 10.0686 - 2.0273 = \mathbf{8.0413}$
   - Fitted values match group means perfectly: $\hat{Y}_0 = 2.0273, \hat{Y}_1 = 9.0686, \hat{Y}_2 = 10.0686$, and R^2 increases from 0.7860 to 0.9855.

6.4 Visual Illustration

Figure 6 illustrates linear OLS fitted across three discrete groups where the middle group exhibits a non-linear shift.

Figure 6: Linear OLS fit vs. Saturated Dummy Specification for $X_1 \in \{0, 1, 2\}$. Linear OLS misses the non-linear jump at $X_1 = 1$.



7 Saturated Models

7.1 Formal Definition

A regression model is defined as **saturated** if it contains a separate parameter for every possible discrete combination or category of the explanatory variable(s) (Angrist & Pischke, 2009; Wooldridge, 2010).

Formally, if a discrete variable X_1 takes K distinct values $\{x_1, x_2, \ldots, x_K\}$, a saturated model includes K independent parameters (e.g., an intercept plus $K - 1$ indicator dummy variables):

$$Y_i = \gamma_0 + \sum_{k=1}^{K-1} \gamma_k \cdot \mathbb{I}(X_{1i} = x_k) + e_i \tag{10}$$

where $\mathbb{I}(\cdot)$ is an indicator function taking the value 1 if the condition holds and 0 otherwise.

7.2 Binary vs. Multi-Valued Saturation

1. **Binary Case** ($X_1 \in \{0, 1\}$): The model $Y_i = \beta_0 + \beta_1 X_{1i} + e_i$ has 2 parameters (β_0, β_1) for $K = 2$ discrete groups. Thus, **simple binary linear OLS is inherently saturated**.

2. **Multi-Valued Case ($X_1 \in \{0, 1, 2\}$)**: The linear model $Y_i = \beta_0 + \beta_1 X_{1i} + e_i$ has 2 parameters for $K = 3$ groups. It is **unsaturated**. To saturate the model, we specify dummy indicators:

$$Y_i = \gamma_0 + \gamma_1 D_{1i} + \gamma_2 D_{2i} + e_i \tag{11}$$

where $D_{1i} = \mathbb{I}(X_{1i} = 1)$ and $D_{2i} = \mathbb{I}(X_{1i} = 2)$.

7.3 Parameter Identification in Saturated Models

In the saturated specification for $X_1 \in \{0, 1, 2\}$:

$$\hat{\gamma}_0 = \overline{Y}_0 \quad \text{(Reference group mean)} \tag{12}$$

$$\hat{\gamma}_1 = \overline{Y}_1 - \overline{Y}_0 \quad \text{(Group 1 shift relative to reference)} \tag{13}$$

$$\hat{\gamma}_2 = \overline{Y}_2 - \overline{Y}_0 \quad \text{(Group 2 shift relative to reference)} \tag{14}$$

Core Theorem: Saturated OLS and the CEF

As established by Angrist & Pischke (2009), saturated OLS specifications fit the **Nonparametric Conditional Expectation Function (CEF)** perfectly:

$$\hat{Y}_i = \hat{E}[Y_i \mid X_{1i} = x_k] = \overline{Y}_k \quad \forall k \in \{1, \ldots, K\} \tag{15}$$

Because the saturated model places no functional form constraints on how group means vary across categories, it completely eliminates functional form specification error.

8 Summary Comparison Table

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Greene, W. H. (2018). *Econometric Analysis* (8th ed.). Pearson.

Stock, J. H., & Watson, M. W. (2020). *Introduction to Econometrics* (4th ed.). Pearson.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). MIT Press.

The End

Table 1: Color-Coded Observations and OLS Residuals: Binary Case ($N = 20$)

Subject ID (i)	**Treatment** (X_{1i})	**Outcome** (Y_i)	**Fitted Value** ($\hat{Y}_i$)	**Residual** ($e_i = Y_i - \hat{Y}_i$)
1	0	3.4684	3.0312	$+0.4372$
2	0	3.5676	3.0312	$+0.5364$
3	0	2.9126	3.0312	-0.1186
4	0	2.6372	3.0312	-0.3940
5	0	3.4847	3.0312	$+0.4535$
6	0	1.5456	3.0312	-1.4856
7	0	3.5041	3.0312	$+0.4729$
8	0	2.7791	3.0312	-0.2521
9	0	2.7727	3.0312	-0.2585
10	0	3.4000	3.0312	$+0.3688$
Control Summary	$\overline{X}_0 = 0.00$	$\mathbf{\overline{Y}_0 = 3.0312}$	$\hat{\beta}_0 = 3.0312$	$\mathbf{\sum e_{i\|X=0} = 0.0000}$
11	1	7.8208	8.0784	-0.2576
12	1	8.4063	8.0784	$+0.3279$
13	1	7.5385	8.0784	-0.5399
14	1	8.2100	8.0784	$+0.1316$
15	1	8.6538	8.0784	$+0.5754$
16	1	7.3916	8.0784	-0.6868
17	1	8.7818	8.0784	$+0.7034$
18	1	7.9175	8.0784	-0.1609
19	1	8.1065	8.0784	$+0.0281$
20	1	7.9572	8.0784	-0.1212
Treat Summary	$\overline{X}_1 = 1.00$	$\mathbf{\overline{Y}_1 = 8.0784}$	$\hat{\beta}_0 + \hat{\beta}_1 = 8.0784$	$\mathbf{\sum e_{i\|X=1} = 0.0000}$

Table 2: OLS Regression Table: Binary Regressor ($Y = \beta_0 + \beta_1 X_1 + e$)

Variable	**Coefficient** ($\hat{\beta}$)	**Std. Error**	t-**Statistic**	p-**Value**
Intercept ($\hat{\beta}_0$)	3.0312	0.1706	17.77	$< 0.0001^{***}$
Treatment X_1 ($\hat{\beta}_1$)	5.0472	0.2413	20.92	$< 0.0001^{***}$
Observations (N)	20			
Residual Std. Error	0.5395	(df = 18)		
Multiple R^2	0.9605	**Adjusted** R^2: 0.9583		
F-**Statistic**	437.5	(p-value: 1.21×10^{-14})		

Significance levels: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 3: Comparative Data Table: Linear OLS vs. Saturated Dummy Model ($N = 30$)

			Linear OLS Model		**Saturated Dummy Model**	
ID (i)	X_{1i}	**Outcome (Y_i)**	**Fitted ($\hat{Y}^{\text{lin}}$)**	**Residual (e^{lin})**	**Fitted ($\hat{Y}^{\text{sat}}$)**	**Residual (e^{sat})**
1	0	2.4124	3.0339	-0.6215	2.0273	$+0.3851$
2	0	2.4992	3.0339	-0.5347	2.0273	$+0.4719$
3	0	1.9260	3.0339	-1.1079	2.0273	-0.1013
4	0	1.6851	3.0339	-1.3488	2.0273	-0.3422
5	0	2.4266	3.0339	-0.6073	2.0273	$+0.3993$
6	0	0.7274	3.0339	-2.3065	2.0273	-1.2999
7	0	2.4436	3.0339	-0.5903	2.0273	$+0.4163$
8	0	1.8067	3.0339	-1.2272	2.0273	-0.2206
9	0	1.8011	3.0339	-1.2328	2.0273	-0.2262
10	0	2.5475	3.0339	-0.4864	2.0273	$+0.5202$
G0 Summary	$\overline{X}_0 = 0$	$\mathbf{\overline{Y}_0 = 2.0273}$	$\hat{Y}_0 = 3.0339$	$\mathbf{\sum e = -10.066}$	$\hat{\gamma}_0 = 2.0273$	$\mathbf{\sum e = 0.0000}$
11	1	8.8432	7.0546	$+1.7886$	9.0686	-0.2254
12	1	9.3555	7.0546	$+2.3009$	9.0686	$+0.2869$
13	1	8.5962	7.0546	$+1.5416$	9.0686	-0.4724
14	1	9.1838	7.0546	$+2.1292$	9.0686	$+0.1152$
15	1	9.5721	7.0546	$+2.5175$	9.0686	$+0.5035$
16	1	8.4677	7.0546	$+1.4131$	9.0686	-0.6009
17	1	9.6841	7.0546	$+2.6295$	9.0686	$+0.6155$
18	1	8.9278	7.0546	$+1.8732$	9.0686	-0.1408
19	1	9.0932	7.0546	$+2.0386$	9.0686	$+0.0246$
20	1	8.9626	7.0546	$+1.9080$	9.0686	-0.1060
G1 Summary	$\overline{X}_1 = 1$	$\mathbf{\overline{Y}_1 = 9.0686}$	$\hat{Y}_1 = 7.0546$	$\mathbf{\sum e = +20.140}$	$\hat{\gamma}_0 + \hat{\gamma}_1 = 9.0686$	$\mathbf{\sum e = 0.0000}$
21	2	9.8907	11.0753	-1.1846	10.0686	-0.1779
22	2	10.4030	11.0753	-0.6723	10.0686	$+0.3344$
23	2	9.6437	11.0753	-1.4316	10.0686	-0.4249
24	2	10.2313	11.0753	-0.8440	10.0686	$+0.1627$
25	2	10.6196	11.0753	-0.4557	10.0686	$+0.5510$
26	2	9.5152	11.0753	-1.5601	10.0686	-0.5534
27	2	10.7316	11.0753	-0.3437	10.0686	$+0.6630$
28	2	9.9753	11.0753	-1.1000	10.0686	-0.0933

Table 4: Regression Output Comparison: Linear OLS vs. Saturated Model

	Model (1): Linear OLS		**Model (2): Saturated Dummies**	
Variable	**Coef. ($\hat{\beta}$)**	**Std. Err.**	**Coef. ($\hat{\gamma}$)**	**Std. Err.**
Intercept	3.0339^{***}	(0.5280)	2.0273^{***}	(0.1508)
Linear Regressor (X_1)	4.0207^{***}	(0.3960)	—	—
Group 1 Dummy ($\mathbb{I}(X_1 = 1)$)	—	—	7.0413^{***}	(0.2132)
Group 2 Dummy ($\mathbb{I}(X_1 = 2)$)	—	—	8.0413^{***}	(0.2132)
Observations (N)	30		30	
Residual Std. Error	1.8320 (df = 28)		0.4768 (df = 27)	
Multiple R^2	0.7860		**0.9855**	
Adjusted R^2	0.7784		**0.9844**	
F-Statistic	102.9^{***} (p: 1.82×10^{-10})		917.8^{***} (p: $< 2.2 \times 10^{-16}$)	

Significance levels: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 5: Comparative Overview: Linear vs. Saturated Models across Supports of X_1

Specification	**Support of X_1**	**Intercept ($\hat{\beta}_0$)**	**Slope / Shifts**	**Fits Group Means?**
Linear OLS	$\{0, 1\}$	$\overline{Y}_0$	$\hat{\beta}_1 = \overline{Y}_1 - \overline{Y}_0$	**Yes** (Passes through both)
Linear OLS	$\{0, 1, 2\}$	$\frac{5\overline{Y}_0 + 2\overline{Y}_1 - \overline{Y}_2}{6}$ *	$\hat{\beta}_1 = \frac{\overline{Y}_2 - \overline{Y}_0}{2}$ *	**No** (Imposes linearity constraint)
Saturated Dummies	$\{0, 1, 2\}$	$\overline{Y}_0$	$\hat{\gamma}_1 = \overline{Y}_1 - \overline{Y}_0, \hat{\gamma}_2 = \overline{Y}_2 - \overline{Y}_0$	**Yes** (Fits all 3 means perfectly)

*Calculated assuming equal group sizes ($n_0 = n_1 = n_2$).