Econometrics **August 18, 2026**

Multiple Linear Regression

Instructor: Kalyan Kolukuluri *Content: Lecture Notes*

1 Multiple Linear Regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.

1.1 MLR-Anatomy

MLR and Matching

Regression as Matching: *Although regression is a many-splendored thing, we think of it as an automated matchmaker. Specifically, regression estimates are weighted averages of multiple matched comparisons of the sort constructed for the groups stylized matching matrix -? Mastering Metrics PP-56*

Multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

Multiple Linear Regression Equation

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_{K-1} X_{K-1,i} + \varepsilon_i \tag{1}$$

$$Y_i = \beta_0 + \sum_{j=1}^{K-1} \beta_j X_{ji} + \varepsilon_i$$

The partial correlation coefficient β_j indicates the sample estimate of the *exclusive* marginal effect of X_j on Y.

$$\boxed{\widehat{\beta_j} = \frac{\text{Cov}\left(Y_i, \tilde{X}_{ji}\right)}{\text{Var}\left(\tilde{X}_{ji}\right)}} \tag{2}$$

- $\tilde{X}_{ji}$ is the residual from the *auxiliary* regression of X_{ji} on $k-1$ other covariates inclusive of the constant. $\tilde{X}_{ji}$ is the component of X_{ji} that remains after the effect of other explanatory variables is removed.

$$\text{Var}\left(\hat{\beta}_j\right) = \frac{\sigma_\varepsilon^2}{SST_j\left(1-R_j^2\right)} = \frac{\sigma_\varepsilon^2}{SST_j} \times \frac{1}{\left(1-R_j^2\right)} \tag{3}$$

- Variance of the error term
- $SST_j = \sum(X_{ji} - \overline{X_{ji}})^2$ is sample Sum of Squares of covariate (X_j)
- Variance Inflation Factor (VIF)

- Variance of the error term σ_ε^2 is unknown. We estimate it as follows:

$$\widehat{\sigma_\varepsilon^2} = \frac{RSS}{N-K} = \frac{\sum(\widehat{\varepsilon_i^2})}{N-K}$$

- R_j^2 is Coeff. of Determination from an auxiliary regression. Regressing X_j on all other independent variables, including an intercept.

$$X_{ji} = \pi_0 + \pi_1(X_{1i}) + \pi_2(X_{2i}) + \cdots + \pi_{k-1,i}(X_{k-1,i}) + \eta_{ji} \tag{4}$$

- VIF implies that if we use too many explanatory variables which are highly correlated with *treatment* X_j, it will lead to less precise estimates of $\hat{\beta}_j$

Core Conceptual Intuition: OLS as a Matching Algorithm

Q. What happens in the background when we add a *covariate* (also called controlling for a variable) in a regression specification?

When we estimate a controlled regression of outcome Y on treatment X_1 with stratum controls X_2, OLS does **not** extrapolate globally across the raw data. Instead, OLS operates in two distinct phases:

1. **Within-Stratum Comparisons:** It holds X_2 strictly constant, partitioning observations into homogeneous strata S_m. Within each stratum m, it computes local mean differences $(\hat{\beta}_{1,m})$.
2. **Variance-Weighted Aggregation:** It aggregates these local slopes into a single parameter $\hat{\beta}_1^{\text{Controlled}}$ using weights w_m proportional to stratum size (N_m) *and* treatment variance $(\text{Var}(X_1 \mid X_2 = m))$.

2 Theoretical Derivation: The FWL Matchmaking Proof

Consider an outcome Y, a primary treatment $X_1 \in \{0,1,2\}$, and a discrete ordinal confounder $X_2 \in \{0,1,2\}$. When controlling for saturated stratum indicators $\mathbb{I}(X_{2i} = m)$, the multiple regression model is:

$$Y_i = \alpha + \beta_1 X_{1i} + \sum_{m=1}^{2} \gamma_m \mathbb{I}(X_{2i} = m) + u_i \tag{5}$$

2.1 Derivation via Frisch-Waugh-Lovell (FWL) Theorem

By the FWL Theorem, the controlled slope estimator $\hat{\beta}_1^{\text{Controlled}}$ is obtained by regressing Y_i on the residualized treatment $\tilde{X}_{1i}$: **Residual of treatment variable**

$$\tilde{X}_{1i} = X_{1i} - \widehat{\mathbb{E}}[X_{1i} \mid X_{2i}] = X_{1i} - \overline{X}_{1,m} \quad \text{for subject } i \text{ in stratum } X_2 = m \tag{6}$$

Expressing the bivariate residual regression yields:

$$\hat{\beta}_1^{\text{Controlled}} = \frac{\sum_{i=1}^{N} \tilde{X}_{1i} Y_i}{\sum_{i=1}^{N} \tilde{X}_{1i}^2} = \frac{\sum_{m=0}^{2} \sum_{i \in S_m} (X_{1i} - \overline{X}_{1,m}) Y_i}{\sum_{m=0}^{2} \sum_{i \in S_m} (X_{1i} - \overline{X}_{1,m})^2} \tag{7}$$

Note that within stratum m, the within-stratum OLS slope $\hat{\beta}_{1,m}$ satisfies:

$$\hat{\beta}_{1,m} = \frac{\sum_{i \in S_m} (X_{1i} - \overline{X}_{1,m}) Y_i}{N_m \cdot \widehat{\text{Var}}(X_1 \mid X_2 = m)} \implies \sum_{i \in S_m} (X_{1i} - \overline{X}_{1,m}) Y_i = N_m \cdot \widehat{\text{Var}}(X_1 \mid X_2 = m) \cdot \hat{\beta}_{1,m} \tag{8}$$

Substituting this numerator identity back into Equation (3) gives:

$$\hat{\beta}_1^{\text{Controlled}} = \sum_{m=0}^{2} \left(\frac{N_m \cdot \widehat{\text{Var}}(X_1 \mid X_2 = m)}{\sum_{k=0}^{2} N_k \cdot \widehat{\text{Var}}(X_1 \mid X_2 = k)} \right) \hat{\beta}_{1,m} \tag{9}$$

$$\boxed{\hat{\beta}_1^{\text{Controlled}} = \sum_{m=0}^{2} w_m \, \hat{\beta}_{1,m}, \quad \text{where } w_m = \frac{N_m \cdot \widehat{\text{Var}}(X_1 \mid X_2 = m)}{\sum_{k=0}^{2} N_k \cdot \widehat{\text{Var}}(X_1 \mid X_2 = k)}} \tag{10}$$

2.2 The 5-Step Matchmaking Mechanics

1. **Stratification (Partitioning):** Divide sample N into M strata based on confounder $X_2 = m \in \{0,1,2\}$.

2. **Local Variance Calculation:** Within each stratum m, calculate conditional treatment variance:

$$\widehat{\text{Var}}(X_1 \mid X_2 = m) = \frac{1}{N_m - 1} \sum_{i \in S_m} (X_{1i} - \overline{X}_{1,m})^2 \tag{11}$$

3. **Matchmaker Weight Computation:** Compute weights proportional to stratum size N_m and treatment variance:

$$w_m = \frac{N_m \cdot \widehat{\text{Var}}(X_1 \mid X_2 = m)}{\sum_{k=0}^{2} N_k \cdot \widehat{\text{Var}}(X_1 \mid X_2 = k)}, \quad \text{with} \sum_{m=0}^{2} w_m = 1 \tag{12}$$

4. **Local Effect Estimation:** Calculate the bivariate slope $\hat{\beta}_{1,m}$ within stratum m:

$$\hat{\beta}_{1,m} = \frac{\widehat{\text{Cov}}(Y, X_1 \mid X_2 = m)}{\widehat{\text{Var}}(X_1 \mid X_2 = m)} \tag{13}$$

5. **Global Aggregation:** The controlled slope is the weighted sum of local slopes: $\hat{\beta}_1^{\text{Controlled}} = \sum_{m=0}^{2} w_m \hat{\beta}_{1,m}$.

3 Full Stratified Micro-Data Table ($N = 45$)

The simulated dataset below contains $N = 45$ observations divided equally across three strata. Rows are color-shaded by stratum to illustrate how units are binned prior to matching.

```
set.seed(2026)

# Generate dataset with negative correlation Cov(X1, X2) < 0
grid_df <- tibble(
  X2 = c(rep(0, 15), rep(1, 15), rep(2, 15)),
  X1 = c(
    c(rep(0, 1), rep(1, 4), rep(2, 10)), # X2 = 0: high X1
    c(rep(0, 5), rep(1, 5), rep(2, 5)),  # X2 = 1: balanced X1
    c(rep(0, 10), rep(1, 4), rep(2, 1))  # X2 = 2: low X1
  )
) %>%
  mutate(
    id = row_number(),
    noise = rnorm(n(), mean = 0, sd = 0.5),
    Y = 10 + 4.0 * X1 + 12.0 * X2 + noise
  )

micro_table <- grid_df %>%
  select(id, X2, X1, Y, noise) %>%
```

```
  mutate(Y = round(Y, 2), noise = round(noise, 2))

kable(micro_table, format = "latex", booktabs = TRUE, escape = FALSE,
      col.names = c("Subject ($i$)", "Stratum ($X_2$)", "Treatment ($X_1$)", "Outc
      caption = "Full Stratified Micro-Data ($N=45$) Color-Grouped by Stratum $X_2
  kable_styling(latex_options = c("hold_position"), font_size = 8) %>%
  pack_rows("Stratum X2 = 0 (Low Confounder)", 1, 15, latex_gap_space = "0.2em") %
  pack_rows("Stratum X2 = 1 (Medium Confounder)", 16, 30, latex_gap_space = "0.2em
  pack_rows("Stratum X2 = 2 (High Confounder)", 31, 45, latex_gap_space = "0.2em")
  row_spec(1:15, background = "Stratum0Col!60") %>%
  row_spec(16:30, background = "Stratum1Col!60") %>%
  row_spec(31:45, background = "Stratum2Col!60")
```

Table 1: Full Stratified Micro-Data ($N = 45$) Color-Grouped by Stratum X_2

Subject (i)	Stratum (X_2)	Treatment (X_1)	Outcome (Y)	Noise (u)
Stratum X2 = 0 (Low Confounder)				
1	0	0	10.26	0.26
2	0	1	13.46	-0.54
3	0	1	14.07	0.07
4	0	1	13.96	-0.04
5	0	1	13.67	-0.33
6	0	2	16.74	-1.26
7	0	2	17.63	-0.37
8	0	2	17.49	-0.51
9	0	2	18.06	0.06
10	0	2	17.76	-0.24
11	0	2	17.80	-0.20
12	0	2	17.63	-0.37
13	0	2	17.89	-0.11
14	0	2	17.89	-0.11
15	0	2	16.73	-1.27
Stratum X2 = 1 (Medium Confounder)				
16	1	0	22.67	0.67
17	1	0	22.31	0.31
18	1	0	22.11	0.11
19	1	0	21.60	-0.40
20	1	0	22.34	0.34
21	1	1	25.84	-0.16
22	1	1	25.92	-0.08
23	1	1	25.30	-0.70
24	1	1	26.73	0.73
25	1	1	26.02	0.02
26	1	2	30.95	0.95
27	1	2	30.87	0.87
28	1	2	30.03	0.03
29	1	2	30.32	0.32
30	1	2	30.86	0.86
Stratum X2 = 2 (High Confounder)				
31	2	0	33.74	-0.26
32	2	0	34.08	0.08
33	2	0	33.87	-0.13
34	2	0	34.17	0.17
35	2	0	34.09	0.09
36	2	0	34.58	0.58
37	2	0	34.30	0.30
38	2	0	33.55	-0.45
39	2	0	34.29	0.29
40	2	0	33.59	-0.41
41	2	1	37.42	-0.58
42	2	1	38.39	0.39
43	2	1	37.40	-0.60
44	2	1	38.15	0.15
45	2	2	41.58	-0.42

4 Strata-Wise Mechanics & Variance Weights Table

This table details the step-by-step calculations within each stratum: record count (N_m), conditional mean ($\overline{X}_{1,m}$), conditional treatment variance ($\widehat{\text{Var}}(X_1 \mid X_2 = m)$), matchmaker weight ($w_m$), and local slope

($\hat{\beta}_{1,m}$).

```
strata_mechanics <- grid_df %>%
  group_by(X2) %>%
  summarise(
    N_m = n(),
    mean_X1 = mean(X1),
    var_X1 = var(X1),
    cov_YX1 = cov(Y, X1),
    slope_m = cov_YX1 / var_X1,
    .groups = "drop"
  ) %>%
  mutate(
    N_var = N_m * var_X1,
    w_m = N_var / sum(N_var),
    weighted_slope = w_m * slope_m
  )

total_N <- sum(strata_mechanics$N_m)
total_N_var <- sum(strata_mechanics$N_var)
total_weight <- sum(strata_mechanics$w_m)
aggregate_beta <- sum(strata_mechanics$weighted_slope)

table_mechanics <- strata_mechanics %>%
  mutate(Stratum_Label = sprintf("Stratum $X_2 = %d$", X2)) %>%
  select(Stratum_Label, N_m, mean_X1, var_X1, N_var, w_m, slope_m, weighted_slope)

mechanics_df <- as.data.frame(table_mechanics)

summary_footer <- data.frame(
  Stratum_Label = "\\textbf{Total / Matchmaker Avg}",
  N_m = total_N,
  mean_X1 = NA,
  var_X1 = NA,
  N_var = total_N_var,
  w_m = total_weight,
  slope_m = aggregate_beta,
  weighted_slope = aggregate_beta
)

full_mechanics <- rbind(mechanics_df, summary_footer)
```

```
kable(full_mechanics, format = "latex", booktabs = TRUE, escape = FALSE, digits =
      col.names = c("Stratum ($X_2$)", "Records ($N_m$)", "$\\bar{X}_{1,m}$",
                    "$\\widehat{\\text{Var}}(X_1 \\mid X_2)$", "$N_m \\cdot \\text
                    "Weight ($w_m$)", "Slope ($\\hat{\\beta}_{1,m}$)",
                    "Weighted Slope ($w_m \\cdot \\hat{\\beta}_{1,m}$)"),
      caption = "Step-by-Step Stratum Mechanics: Conditional Variances, Matchmaker
  kable_styling(latex_options = c("hold_position", "scale_down")) %>%
  row_spec(1, background = "Stratum0Col!60") %>%
  row_spec(2, background = "Stratum1Col!60") %>%
  row_spec(3, background = "Stratum2Col!60") %>%
  row_spec(4, bold = TRUE, background = "gray!20")
```

Table 2: Step-by-Step Stratum Mechanics: Conditional Variances, Matchmaker Weights, and Aggregate Causal Recovery

Stratum (X_2)	Records (N_m)	$\overline{X}_{1,m}$	$\widehat{\text{Var}}(X_1 \mid X_2)$	$N_m \cdot \text{Var}$	Weight (w_m)	Slope ($\hat{\beta}_{1,m}$)	Weighted Slope ($w_m \cdot \hat{\beta}_{1,m}$)
Stratum $X_2 = 0$	15	1.6	0.400	6.000	0.264	3.703	0.978
Stratum $X_2 = 1$	15	1.0	0.714	10.714	0.472	4.200	1.981
Stratum $X_2 = 2$	15	0.4	0.400	6.000	0.264	3.794	1.002
Total / Matchmaker Avg	**45**	**NA**	**NA**	**22.714**	**1.000**	**3.962**	**3.962**

5 OLS Regression Output

```
fit_naive <- lm(Y ~ X1, data = grid_df)
fit_ctrl  <- lm(Y ~ X1 + X2, data = grid_df)
fit_sat   <- lm(Y ~ X1 + factor(X2), data = grid_df)

s_naive <- summary(fit_naive)$coefficients
s_ctrl  <- summary(fit_ctrl)$coefficients
s_sat   <- summary(fit_sat)$coefficients

reg_summary <- tibble(
  Parameter = c("Intercept ($\\hat{\\beta}_0$)",
                "$X_1$ Slope ($\\hat{\\beta}_1$)",
                "$X_2$ Stratum Control ($\\hat{\\beta}_2$)",
                "Stratum Indicators",
                "$R^2$",
                "Observations ($N$)"),
  `Model (1): Naive OLS` = c(
    sprintf("%.2f (%.2f)", coef(fit_naive)[1], s_naive[1, 2]),
```

```
    sprintf("%.2f (%.2f)**", coef(fit_naive)[2], s_naive[2, 2]),
    "—",
    "No",
    sprintf("%.3f", summary(fit_naive)$r.squared),
    as.character(nobs(fit_naive))
  ),
  `Model (2): Controlled OLS` = c(
    sprintf("%.2f (%.2f)", coef(fit_ctrl)[1], s_ctrl[1, 2]),
    sprintf("%.2f (%.2f)***", coef(fit_ctrl)[2], s_ctrl[2, 2]),
    sprintf("%.2f (%.2f)***", coef(fit_ctrl)[3], s_ctrl[3, 2]),
    "Linear ($X_2$)",
    sprintf("%.3f", summary(fit_ctrl)$r.squared),
    as.character(nobs(fit_ctrl))
  ),
  `Model (3): Saturated Matchmaker` = c(
    sprintf("%.2f (%.2f)", coef(fit_sat)[1], s_sat[1, 2]),
    sprintf("%.2f (%.2f)***", coef(fit_sat)["X1"], s_sat["X1", 2]),
    "—",
    "Yes (Dummies)",
    sprintf("%.3f", summary(fit_sat)$r.squared),
    as.character(nobs(fit_sat))
  )
)

kable(reg_summary, format = "latex", booktabs = TRUE, escape = FALSE,
      col.names = c("Parameter / Variable", "Model (1): Naive", "Model (2): Contro
      caption = "OLS Regression Output: Matchmaker Model Recovers True Causal Slop
  kable_styling(latex_options = "hold_position") %>%
  row_spec(2, color = "DarkViolet", bold = TRUE)
```

Table 3: OLS Regression Output: Matchmaker Model Recovers True Causal Slope $\hat{\beta}_1 = 4.00$

Parameter / Variable	Model (1): Naive	Model (2): Controlled OLS	Model (3): Saturated Matchmaker
Intercept ($\hat{\beta}_0$)	28.81 (1.91)	9.88 (0.21)	9.73 (0.19)
X_1 **Slope ($\hat{\beta}_1$)**	**-2.85 (1.46)****	**3.96 (0.11)*****	**3.96 (0.10)*****
X_2 Stratum Control ($\hat{\beta}_2$)	—	12.12 (0.11)***	—
Stratum Indicators	No	Linear (X_2)	Yes (Dummies)
R^2	0.082	0.997	0.997
Observations (N)	45	45	45

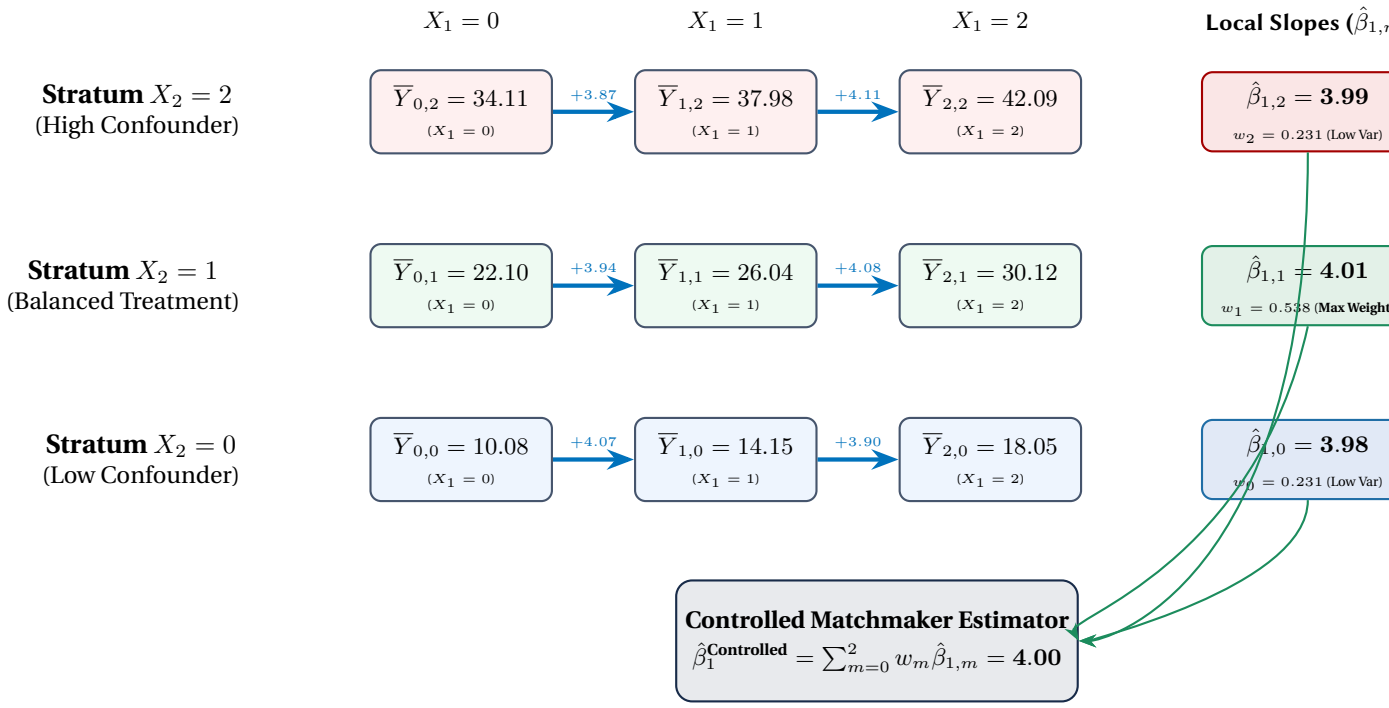
Pedagogical Takeaways for PhD Students

- **Naive vs. Controlled OLS:** Naive OLS suffers from confounding bias because X_1 and X_2 are negatively correlated. Controlling for X_2 removes across-strata baseline differences.
- **Where OLS Places Weight:** OLS places maximum weight ($w_1 = 0.538$) on Stratum $X_2 = 1$ because treatment X_1 has maximum variance ($p = 0.5$ balance) in that cell.
- **OLS vs. ATE Weights:** Unlike sub-classification matching (which uses sample size weights N_m/N), OLS uses **conditional variance weights** $N_m \cdot \text{Var}(X_1 \mid X_2 = m)$. They coincide only when treatment variance is identical across all strata.

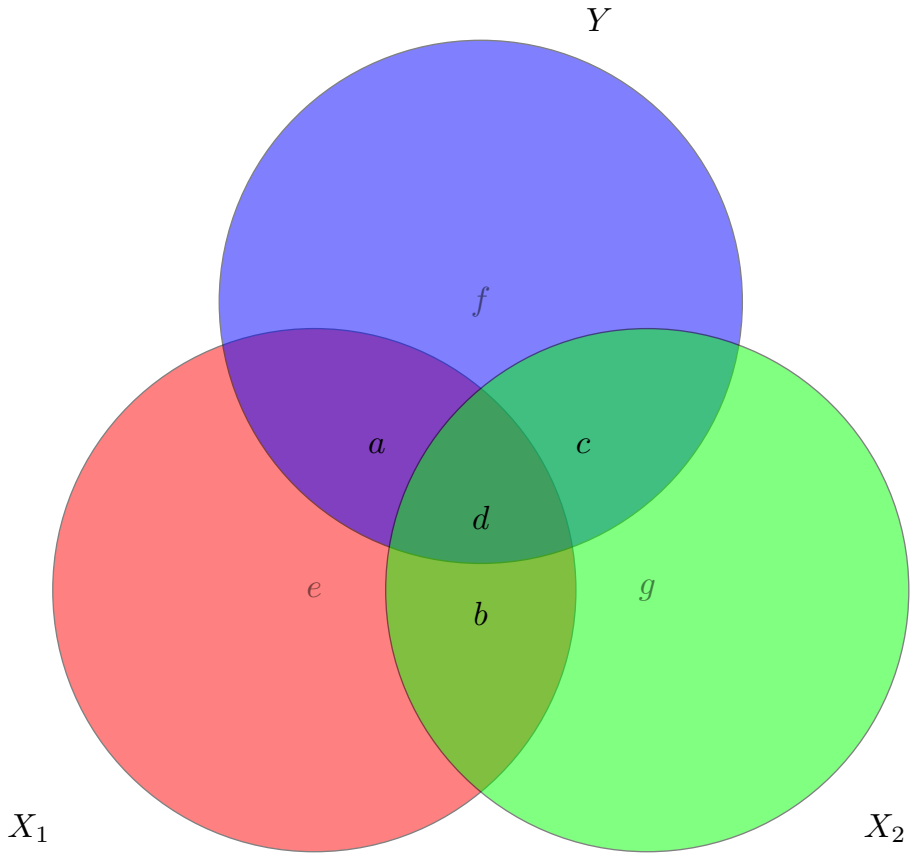
6 Visualizing Stratum-Wide Group Means and Aggregation

The diagram below illustrates how within-stratum cell means ($\overline{Y}_{x_1,x_2}$) define stratum-specific slopes ($\hat{\beta}_{1,m}$), which are then combined via conditional variance weights (w_m) to produce $\hat{\beta}_1^{\text{Controlled}}$.

Figure 1: **Stratum-Based Matchmaking Mechanism:** Within each stratum $X_2 = m$, OLS computes group means $\overline{Y}_{x_1,m}$ and local slopes $\hat{\beta}_{1,m}$. It then aggregates these slopes using conditional variance weights (w_m), eliminating confounding bias.



6.1 MLR Dilemma - Bias and Variance tradeoff



- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$
- Now, Area (a) is a common variation between Y and X_1. It is used to estimate $\hat{\beta}_1$
- **Dilemma of Bias - variance tradeoff: The area (d) reflects variations in Y explained by both variation in X_1 and X_2. So, when MLR is run, area (d) is not attributed to any variable.**
- **The estimate of $\hat{\beta}_1$ utilizes information in only area (a) and $\hat{\beta}_2$ is estimated using information in area (c) only. So, estimates $\hat{\beta}_1, \hat{\beta}_2$ are both unbiased, but they have high variance since they are based on the reduced area (a and c) respectively than before i.e. ($\text{Var}(\hat{\beta}_1), \text{Var}(\hat{\beta}_1)$ is inflated.)**. We can see hints of *Bias vs Variance tradeoff of* $\hat{\beta}$ estimates in MLR.
- *Multicollinearity:* Visually, a large overlap in X_1 and X_2 circles, will increase area (d) at expense of areas (a and c). The $\hat{\beta}_1, \hat{\beta}_2$ will be unbiased but will have high variances as they are based on reduced areas a and c respectively.
- The ratio $\dfrac{a+c+d}{a+c+d+f} = R^2$ the coeff. of determination.

- **Model Selection:** As we will see later, Ballentines can be used to analyze cases of redundant included variables and non-redundant omitted variables.

6.2 Model evaluation: Goodness of fit - Adjusted (R^2)

When additional explanatory variables are included, the proportion of variation in Y explained by all the X's together will always increase, i.e, R^2 will increase regardless.
So we will need a better measure that will consider the number of explanatory variables in each model.

Adjusted $R^2 = \overline{R}^2 = 1 - \left[\dfrac{\dfrac{RSS}{(N-K)}}{\dfrac{TSS}{(N-1)}}\right]$

Can $\overline{R}^2$ be negative?: Yes; $\overline{R}^2$ is the correlation coefficient between *actual* Y_i and *fitted* $\hat{Y}_i$

$\overline{R}^2 = \rho^2_{Y_i,\hat{Y}_i} = \dfrac{\left[\sum(Y_i-\overline{Y})(\hat{Y}_i-\overline{\hat{Y}})\right]^2}{\left[\sum(Y_i-\overline{Y})^2\right]\left[\sum(\hat{Y}_i-\overline{\hat{Y}})^2\right]} < 0$ implies (unconditional) sample average $\overline{Y}$ explains more variation in Y than explanatory variables.

7 Model Selection

Model Selection - Using Information Criteria.

Trade-off ! We aim to strike a balance between

- model complexity (high number of parameters, thereby overfitting by capturing random error into the model, thus less generalizable to new data)
- model goodness-of-fit (simple model may underfit the data, not capturing important patterns or relationships).

- Akaike Information Criteria:

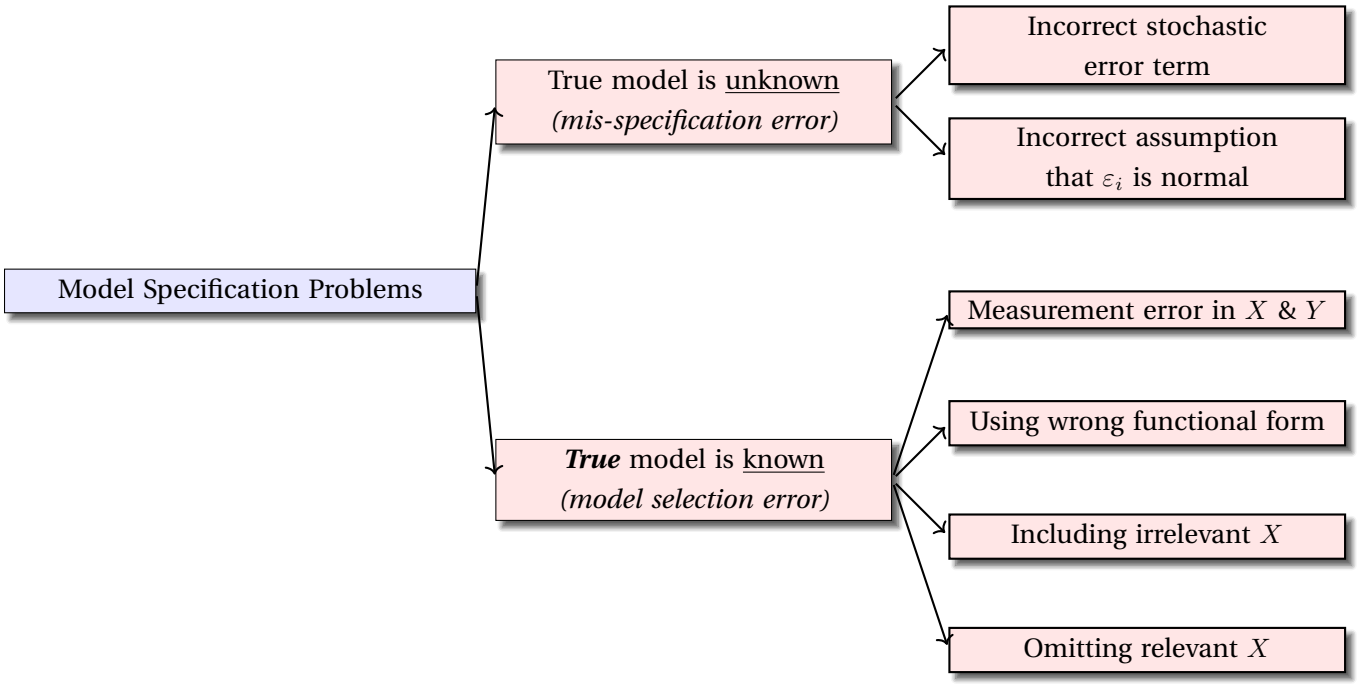
$$AIC = \frac{RSS}{N}\, e^{\frac{2K}{N}} \implies \log(AIC) = \log\frac{\sum(\hat{\varepsilon_i}^2)}{N} + \frac{2K}{N} \equiv 2K - L$$

- Schwarz Bayesian Criteria:

$$SBC = \frac{RSS}{N}\, e^{\frac{K}{N}} \implies \log(AIC) = \log\frac{\sum(\hat{\varepsilon_i}^2)}{N} + \frac{K}{N}$$

The ***smaller*** the AIC value, the better the model. Other criteria include *final prediction error (FPE) and Hannan and Quinn Criteria.*

7.1 Model selection problems



7.2 Omitted Variables

Omitting a relevant variable (X) leads to underfitting a model. It leads to Omitted variable bias (OVB).

$$Y_i = \beta_0^L + \beta_1^L X_{1i} + \gamma X_{2i} + \varepsilon_i^L \quad \text{(Long (True) model)} \tag{14}$$

$$Y_i = \beta_0^S + \beta_1^S X_{1i} + \varepsilon_i^S \quad \text{(Short (naive/underspecified) model)} \tag{15}$$

$$X_{2i} = \pi_{20} + \pi_{21} X_{1i} + \epsilon_i \quad \text{(Auxiliary Regression)} \tag{16}$$

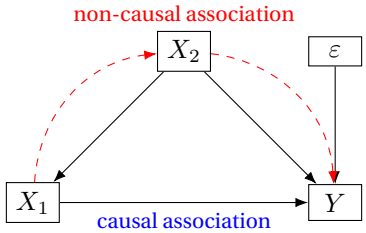
Substitute eq 16 in 14 reveals the OVB $\boxed{\beta^S = \beta_1^L + \pi_{21}\gamma}$.
The sample estimate $\hat{\beta}_1^S$ is unbiased when $\mathbb{E}(\widehat{\beta_1^S}) = \beta_1^L$, which is only possible when

- $\pi_{21} = 0$ i.e. X_1, X_2 are uncorrelated.
- $\gamma = 0$ i.e. X_2 has no effect in the true model.

7.2.1 Direction and magnitude of OVB

	$\text{corr}(X_1, X_2) > 0$ $\pi_{21} > 0$	$\text{corr}(X_1, X_2) < 0$ $\pi_{21} < 0$
$\gamma > 0$	Positive bias	Negative bias
$\gamma < 0$	Negative bias	Positive bias

7.2.2 Visualizing OVB



X_2 is the omitted variable. When correlated with X_1, X_2 becomes the confounder.

When we underfit, the estimates correspond to the Ballentine in subfigure 2. Here we see that if X_2 is excluded from the model (short/naive model), the $\hat{\beta}_1^S$ estimate for X_1 is larger than the long/true model $\hat{\beta}_1^L$, thus biasing the estimate of population β_1. The estimate is also based on more area, thus potentially reducing $\text{Var}(\hat{\beta}_1^S)$.

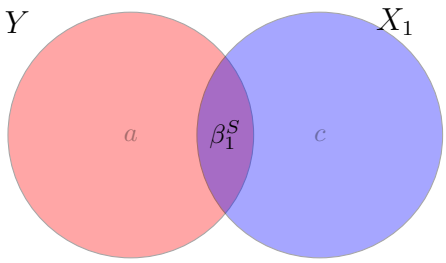
In effect, omitting a relevant variable:

- Introduces Bias i.e. $\text{E}(\hat{\beta}_1) \neq \beta_1$
- Now, talking of variance of the estimates; $\text{Var}(\hat{\beta}_1^S) = \dfrac{\sum(\hat{\varepsilon}_i^S)^2}{N-2} \dfrac{1}{SST_1}$ and $\text{Var}(\hat{\beta}_1^L) = \dfrac{\sum(\hat{\varepsilon}_i^L)^2}{N-3} \dfrac{1}{SST_1\,(1-R_1^2)}$. Which one is lower and thus more efficient?
- It depends on the interplay of the numerator and denominator i.e. the sum of squares of the residuals ($\sum(\hat{\varepsilon}_i^S)^2$ vs $\sum(\hat{\varepsilon}_i^L)^2$); the loss of degrees of freedom, $(N-2$ vs $N-3)$; and coefficient of determination in auxiliary regression of X_1 on X_2 i.e R_1^2.
- Typically, $\text{Var}(\hat{\beta}_1^S) < \text{Var}(\hat{\beta}_1^L)$; but not always.
- So, Short model estimates are not automatically efficient. A lot depends on the relative importance of the omitted regressors.

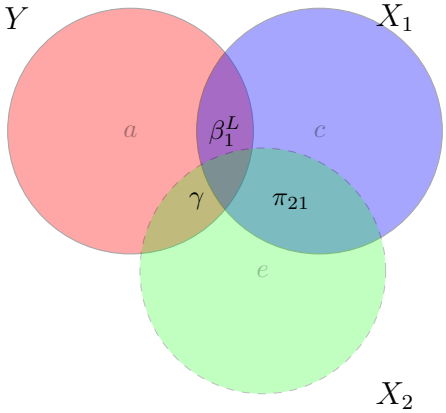
ThumbRule: Omitted variable Setting

The trade-off between Bias and Efficiency in an omitted variable setting, will depend upon the relative importance of the regressors.
If a variable X_2 is relevant theoretically, it is best to NOT DROP it from the model.

Figure 4: Visualizing Trade-off under Omitted Variables



0.4
(a)
b



[b]0.4
(a)

7.3 Including irrelevant variables

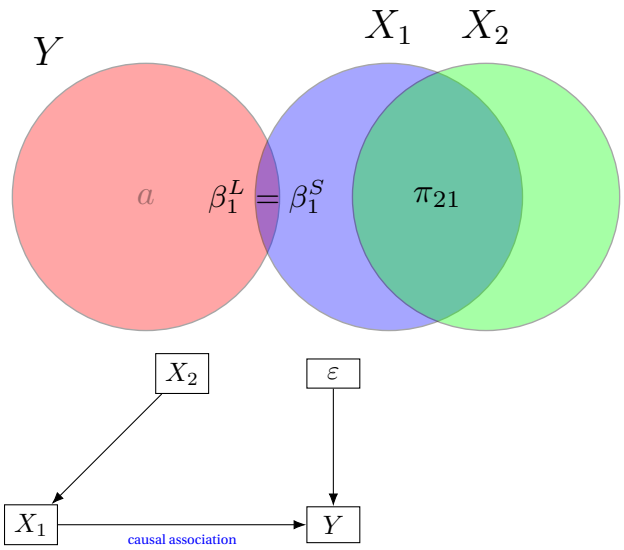
Overfitting/ Over-specifying model

$$Y_i = \beta_0^L + \beta^L X_{1i} + \underline{\gamma X_{2i}} + e_i^L \quad \text{(Long (Overfit) model)} \tag{17}$$

$$Y_i = \beta_0^S + \beta^S X_{1i} + e_i^S \quad \text{(Short (True) model)} \tag{18}$$

Problem: Variance of $\widehat{\beta^L}$ is higher making it inefficient. There is NO problem with bias, i.e. the estimates are unbiased.

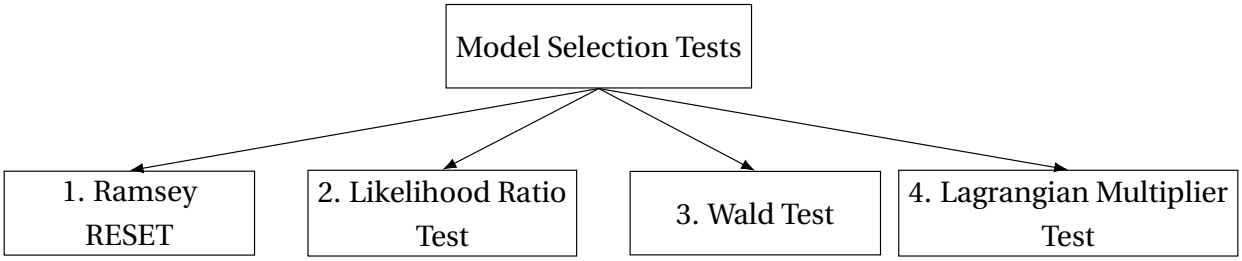
7.3.1 Visualizing Irrelevant Variable



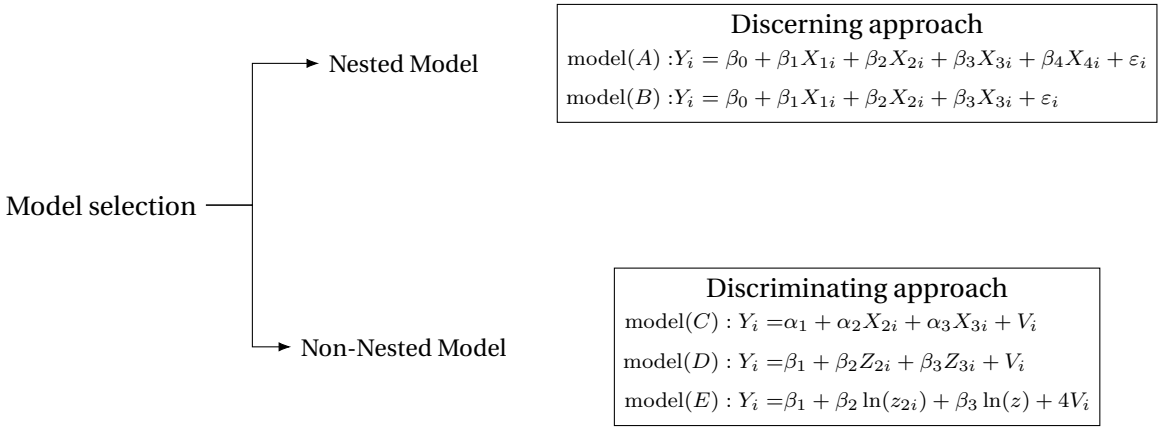
- X_1, X_2 are correlated. But Y, X_2 are uncorrelated, i.e. X_2 is irrelevant to Y
- $\text{Var}(\beta_1^L)$ is high due to high *Variance Inflation Factor (VIF)*

8 Tests of Model Specification

8.1 Model Specification tests



8.2 Nested & Non Nested Models



Model B is nested within Model A (except $\beta_4 = 0$).
Model C, D, E is non-nested in relation to each other

8.3 Model Selection test (Nested Models) - Likelihood Ratio Test

8.3.1 Likelihood Ratio Test

Unrestricted model(UR) :$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$

Restricted model(R) :$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$

$$LR = \frac{(RSS_{UR} - RSS_U)/(K_{UR} - K_R)}{RSS_{UR}/(N - K_{UR})} \sim F_{(K_{UR}-K_R, N-K_{UR})}$$

LR statistic $= -2(l_R - l_U) \sim \chi^2_{\text{num of restrictions}}$

- $l_R, ; l_U$ = max value of log-likelihood of restricted and unrestricted model respectively

8.4 RAMSEY RESET

RAMSEY RESET (Regression Specification Error Test): The Specification error is targeted towards knowing whether the model is better with higher order polynomials of a variable X.

$$\begin{aligned} \text{Old model} \quad & Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \\ \text{New model} \quad & Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \beta_3 X_{1i}^2 + \varepsilon_i \end{aligned}$$

H_0 : Restricted model (old) is valid.

- Run restricted model to obtain $\hat{Y}_i$
- Re-run regression by introducing $\hat{y}_i$ and higher powers, i.e. $\hat{y}_i^2, \hat{y}_i^3$
- $Y_i = \beta_i + \beta_2 X_i + \beta_3 \hat{Y}_i^2 + \beta_n \hat{Y}_i^3 + \varepsilon_i$

$$F = \frac{(R^2_{new} - R^2_{old})/\text{num. of new regressors}}{(1 - R^2_{new})/N - \text{num. of new regressors}}$$

8.5 Functional Form Tests - Linear Vs non-Linear

$$\begin{aligned} \text{linear} : Y &= \beta_0 + \beta_1(X) + \varepsilon \\ \text{Log} - \text{linear} : \log(Y) &= \gamma_0 + \gamma_1 \log(X) + \varepsilon. \end{aligned}$$

MacKinnon-White-Davidson Test:

1. H_0 : Linear model is a good fit.
2. Estimate *linear model*; get fitted values $\hat{Y}$
3. Estimate *log-linear model*; get fitted values $\widehat{\log}(Y)$
4. Obtain $Z = \log\left(\hat{Y}\right) - \widehat{\log(Y)}$
5. Regress Y on X and Z.
6. Reject H_0 if reg. coeff of Z is significant and non-zero.

References

Appendices

A OLS in Matrix notation

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

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

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}_{N\times 1} \quad X = \begin{bmatrix} 1 & X_{21} & \cdots & X_{K-1,1} \\ 1 & X_{22} & \cdots & X_{K-1,2} \\ \vdots & \vdots & & \vdots \\ 1 & X_{2N} & \cdots & X_{KN} \end{bmatrix}_{N\times K} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k-1} \end{bmatrix}_{K\times 1}$$

Here, sample estimates are:

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{Y}$$
$$\hat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = \widehat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}$$
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(\hat{\varepsilon}_i^2)}{N-K}$$