First Assumptions

- We'll begin with a fairly weak set of assumptions about our population model.
    - These are often realistic and safe.
- These are the first four Gauss-Markov assumptions.
    - But these assumptions are not enough for the Gauss-Markov theorem.
- With just the first four assumptions, we'll show that OLS estimators are unbiased.
    - This relates to the **U** in BLUE.

Linearity and Random Sampling

- Assumption MLR.1 (linear in parameters): the basic population model — *y* is a linear function of the *x*'s.
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \dots \ + \beta_k x_k + u$
    - At this point, we don't have to worry about this assumption—we haven't said anything about *u*, so it's not really a restriction.
    - Any population distribution could be represented as a linear model plus some error (error might be poorly behaved).
- Assumption MLR.2 (random sampling): The data is a random sample drawn from the population.
  $\{(x_{i1}, x_{i2}, \ \dots \ + x_{ik}, y_i): \ i = 1, \ \dots n\}$
  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ \dots \ + \beta_k x_{ik} + u_i$
    - All data points follow the population distribution.
    - Data points must be **iid**—independently and identically distributed.

Multicollinearity

- Assumption MLR.3 (no perfect collinearity): In the sample (and population), none of the independent variables are constant and there are no exact relationships among the independent variables.
    - Rules out only perfect collinearity/correlation between explanatory variables—imperfect correlation is allowed.
        - In practice, high correlation can greatly increase errors.
    - If an explanatory variable is a perfect linear combination of other explanatory variables, it is superfluous and may be eliminated.
    - Constant variables are also ruled out (collinear with the intercept term).

Multicollinearity Example

$$\text{VoteA} = \beta_0 + \beta_1 \text{ expendA} + \beta_2 \text{ expendB} + \beta_3 \text{ totexpend}$$

- Here is a model that predicts the share of the vote earned by Candidate A as a function of how much A spends, B spends, and total campaign spending.
- Here, *totexpend* is a linear combination of the other variables, so it has no unique variation for OLS to work with.
    - Whatever coefficients we choose, we could subtract one from β_1 and β_2 and add one to β_3 and the model stays exactly the same—there is no unique set of coefficients to estimate.
- To solve this problem, one variable has to be dropped from the model.

Zero-Conditional Mean

- Assumption MLR.4 (zero-conditional mean): The value of the explanatory variables must contain no information about the mean of the unobserved factors.
  $E(u_i \mid x_{i1}, x_{i2}, \ldots, x_{ik}) = 0$
- This assumption is the strongest so far.
- This assumption enforces linearity.
- MLR.1 establishes a linear population model, but MLR.4 ensures that the population actually follows that linear model.

Four Assumptions

1. Linearity
2. Random sampling
3. Multicollinearity
4. Zero-conditional mean

Unbiased Coefficients:
Theorem 3.1 (Unbiasedness of OLS)

- Under MLR.1-4, OLS estimates are unbiased.
  $E(\hat{\beta}_j) = \beta_j$
- Remember, unbiasedness is an average property in repeated samples; in a given sample, the estimates may still be far away from the true values.
- But at least we know that in expectation, we're measuring the right thing.