MLR Assumptions

- MLR. 1 - Linear in parameters $y = \beta_0 + \beta_1 x + u$
  Should look like, have relationship like, the linear model.
- MLR. 2 - Data drawn from a random sample (i.e., the errors are independent - no autocorrelation in the data)
- **MLR.3** - No perfect collinearity (no exact linear relationship among the independent variables)

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$$

- MLR.4 - Zero conditional mean for the error term (non-endogenous)

$$E(u_i|x_i) = 0$$

  A bias exists if a term hidden in the error term is correlated with an independent variable and has a statistically significant effect on the dependant variable.

- **MLR.5** - Homoskedasticity (i.e., the errors have equal variance)

$$Var(u_i|x_i) = \sigma^2$$

- MLR. 6 - Normality of the error term

  Large sample sizes should be relatively normal based on the Central Limit Theorem. Check normality of residuals around regression line. Does not affect bias of coefficients.

- Validity (i.e. data maps to research question)

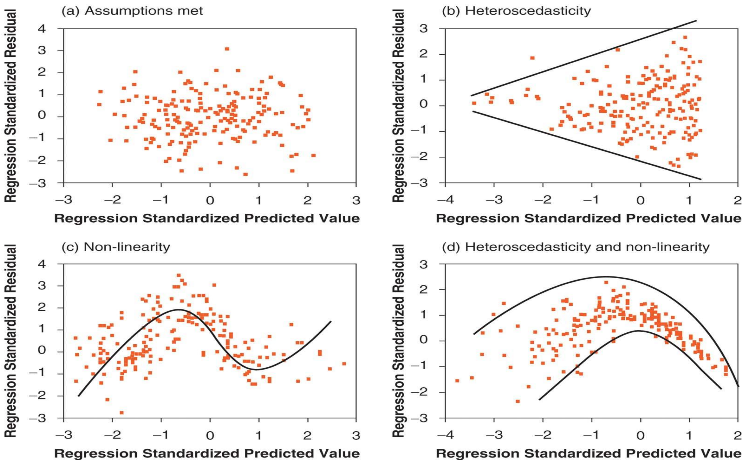
$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \quad SSR = \sum_{i=1}^{n}\hat{u}_i^2$$

<u>Total sum of squares,</u> represents total variation in dependent variable

<u>Model/Explained sum of squares,</u> represents variation explained by regression

<u>Residual sum of squares,</u> represents variation <u>not</u> explained by regression

$$var(\hat{\beta}_j) = \frac{\sigma^2}{N * var(X_j) * (1 - R_j^2)} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x} \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N-k}$$



- Coefficient of determination (R^2): How much of the variability in Y have we been able to explain with our model?
- Residual or regression standard error: How close are the actual observed Y-values to the model-based fitted values?
- Slope parameter (and p-value): How strong is the evidence of a linear association between Y and X?

Omitted variable bias occurs if:
The omitted variable is correlated with an independent variable.
The omitted variable has an effect on indep variable.

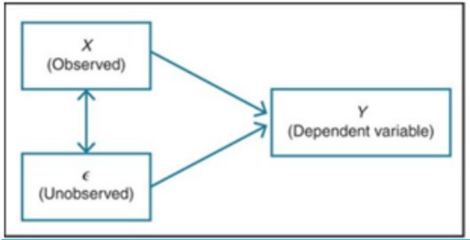


TABLE 3.2 Summary of Bias in $\tilde{\beta}_1$ When x_2 Is Omitted in Estimating Equation (3.40)

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

- Upward bias occurs when $\hat{\beta}_1 > \beta_1$
- Downward bias occurs when $\hat{\beta}_1 < \beta_1$
- Biased toward zero refers to instances when $\hat{\beta}_1$ is closer to zero than β_1

Testing equivalence of coefficients:
May want to do this test to test which strategy gives better returns

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

Test $H_0 : \beta_1 - \beta_2 = 0$ against $H_1 : \beta_1 - \beta_2 < 0$

Cannot base equality of coefficients based on normal regression, use approach like combined terms

Comparing restricted and unrestricted models:
Good to test exclusion restrictions if multiple coefficients are unexpectedly insignificant

$H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$
against $H_1 : H_0$ is not true

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \sim F_{q,n-k-1}$$

Can also use F-test if t-test for coefficients of simple main effects is not showing significance

Interactions:
- Be careful to avoid concluding that a variable involved in an interaction is not important based on its simple main effect.
- Generally best to center predictor variables if there is an interest in interpreting simple main effects (especially if the zero point is outside the relevant range).

$$\boxed{\bar{R}^2 = 1 - \frac{SSR/(n-K-1)}{SST/(n-1)}}$$

R-squared does not account for number of variables, while adjusted R-squared punishes extra unimportant variables

Quadratic coefficient interpretation:

Look at slope or derivative … "...the effect of experience on wage decreases as experience increases. Thus, there are diminishing marginal effects" … Slope decreases and variable increases.