Linearity

- Our linear model assumption expresses y as a linear function of our xs.
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$
- At this point there's nothing to test because we haven't constrained our error in any way.
  - This formula is always true for some definition of u.
  - Given any set of coefficients, we can define $u = y - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \ldots - \beta_k x_k$.

Random Sampling

- This assumption says that all data points are independent random draws from our population distribution.
- Use knowledge of where the data came from to assess the assumption.
  - What was the procedure for collecting data points?

There are two common ways that the assumption can fail.

1. **Clustering:** when individuals are collected into groups, and researchers can only access a limited number of these groups, known as clusters
   - Even with clustering, OLS coefficients are unbiased.
   - Estimates are much less precise under clustering.
   - Use clustered standard errors or other techniques to account for this.

Random Sampling (cont.)

2. **Autocorrelation** or **serial correlation**
    - This is common for time series data.
    - This occurs when the error for one data point is correlated with the error for the next data point.
    - The Durbin-Watson statistic compares the differences between successive data points to the magnitude of the data points.

$$d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

- R computes the Durbin-Watson statistic under the null hypothesis of no serial correlation; if significant, evidence of correlation.
- There is no simple fix for serial correlation.

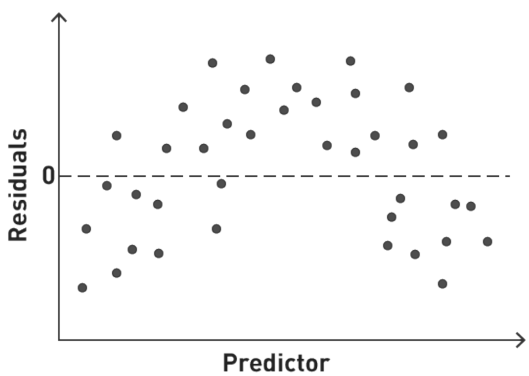
Multicollinearity

- Multicollinearity assumption only rules out perfect multicollinearity.
- The response is simple: drop redundant variables.
- When variables are highly correlated but not perfectly collinear, OLS will still work but estimates will be much less precise.
    - This means we sometimes have to make tough choices.
    - E.g., do we put in a variable and suffer a lot of precision, or leave it out even though we think it has an important effect on the outcome?

Zero-Conditional Mean

- For any possible value of our predictors, our error is zero in expectation.
- To examine this assumption when there's just one predictor, we could create a residuals versus predictor plot.
  - With our *x* on the *x*-axis and our residuals on the *y*-axis
  - Residuals are our estimates of error, so we can see how they change for different values of *x*.

Zero-Conditional Mean Plot

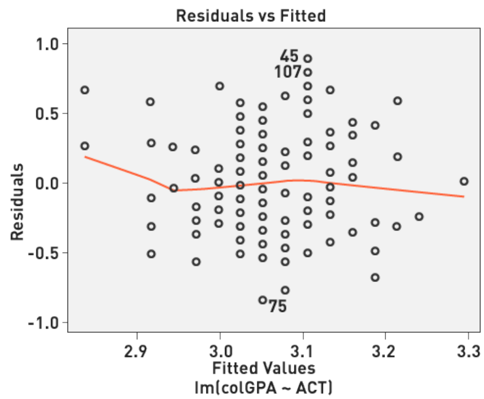


- On this plot, we can eyeball where the mean of the residual changes from left to right.
- You can see that the mean of the residuals seems to go up and then down.
- For zero-conditional mean, we'd want this to be a flat band.

Residuals vs. Fitted Values

- For multiple regression, we can't plot all possible *x* values in two dimensions.
- We could create a separate residual versus predictor plot for every *x*.
  - We would have a lot of graphs; might not reveal all violations of zero-conditional mean.
- More commonly, we would create a residual vs. fitted values plot.
  - The *y*-axis shows residuals, and *x*-axis has predicted values of *y*.
    - These are a linear function of *x*, so if there's a nonzero mean for some values of some *x*, it's likely to show up in this plot.
    - If there's just one *x*, the fitted value of *y* is just a linear scaling of *x*, so the plot is essentially the same as the residual vs. predictor plot.
  - We're looking to see if the plot looks like a flat band.
- Most software, including R, will easily create a residual vs. fitted value plot.

Residuals vs. Fitted Values Plot



- This one is better than the last—there's more of a flat band from left to right.
- R helps us tell if the conditional mean is zero by including a red spline curve.
- Ideally, this curve is completely flat.
- Here, there's a bit of curvature, but it's minor.
    - Might be that there are too few data points on the left of the graph, so the mean could be high randomly.

Responding to Violations of Zero-Conditional Mean

- If the conditional mean of the error is not constant, we may be able to change functional form.
    - Curvature in the residual vs. fitted plot may indicate a linear relationship between x and the log of y (or the log of x and the log of y, etc.).
    - We may allow a more flexible functional form by regressing y on x and x^2; this fits a parabola to the data and may correct violations.
    - These methods have trade-offs.
- Adding new variables may fix the zero-conditional mean assumption.
- If these options fail, we may not be able to meet zero-conditional mean.
    - However, we may be able to meet a weaker assumption: exogeneity.

Exogeneity Defined

- Explanatory variables that are correlated with the error term are called **endogenous**.
  - The term means "originates within the system."
  - Endogeneity is not a direct statement about causality —it's about correlation, which could be present for many reasons.
- Endogeneity is a violation of zero-conditional mean, and its presence implies that OLS coefficients are biased and inconsistent.
- Explanatory variables that are uncorrelated with the error term are called **exogenous**.
  - If x_j is exogenous, $Cov(x_j, u) = 0$.

Exogeneity

- Assumption MLR.4' (Exogeneity): $Cov(x_j, u) = 0$ for all j
- Theorem: Under MLR.1–3 and MLR.4', the OLS estimators are consistent.
  $$\underset{n \to \infty}{\text{plim}} \left(\hat{\beta}_j \right) = \beta_j$$