Assumptions of OLS (Linear Regression)

Ordinary Least Squares (OLS), or linear regression, relies on a few key assumptions that allow us to interpret our results accurately and confidently. If our models violate these assumptions, the math will (often) still work, but our results will be called into question.

Here are the assumptions of OLS regression, in brief:

1. Linearity
2. Representative sampling
3. No multi-collinearity (or perfect collinearity)
4. No heteroskedasticity
5. Independence (no autocorrelation)
6. Normally distributed errors (at mean of 0).

1. Linearity

The entire foundation of ordinary least squares (OLS) regression is the creation of a linear model (a line of best fit that minimizes "errors").If you have reason to suspect that your model behaves differently (like, maybe it's more of a U-shaped curve or an asymptotic curve) maybe don't do a regression. The key here is to have a continuous/interval DV and ensure your theory supports a linear relationship between the IVs and that DV.

2. Representative sampling

Your sample should be representative of the population you wish to generalize to. This is just basic research methods, not something special to OLS.

3. No multicollinearity

OLS assumes that there is no multicollinearity among the independent variables.

Multicollinearity is a situation where one of our independent variables can be predicted from the other independent variables, too!

- Multicollinearity can cause problems with large variances and inverted signs of the coefficients (so interpretation is harder).

Perfect collinearity is also a problem: it means including multiple variables that completely/perfectly determine one another.

- E.g.: if we have a variable for POLS major, we can't also have a *separate* variable for non-POLS major (because it perfectly predicts POLS major and prevents the math from running).

Diagnosis: check correlations among the IVs. In SPSS, tick the "collinearity diagnostics" box in the "Statistics" options and inspect the VIF values for large numbers (>5.0).

Solutions: remove massively collinear variables from the model entirely, find alternate measures/indicators that solve collinearity issues, exclude some “base category'“ from the model when we create multiple “dummy'“ variables.

4. No heteroskedasticity

OLS assumes that the errors are *homoskedastic*.

Heteroskedasticity is a big scary word that just describes inconsistency in the variance among the residuals (i.e., errors). Homoskedasticity is the opposite: ‘same-ness’ or consistency in the variance among the residuals.

- If the variance is not consistent, it will make the standard errors and confidence intervals too narrow or too wide, which undermines trust in our results.

Diagnosis: In SPSS, select the “Plots” option, then request a plot of *ZRESID (Y) against *ZPRED (X) to inspect the tightness of the clusters (tighter is better).

Solution: redefine the variable, weight the variable, transform the variable to make homoskedastic errors.

5. Independence (No autocorrelation)

OLS assumes that observations and variables are independent of one another, which means that they shouldn't *cause* each other. Think like this: in running the test, we want to find out whether the evidence supports the notion that our IVs help cause the DV, but we don’t want some of our IVs to cause our other IVs, too.

Autocorrelation means correlation with itself.

- Autocorrelation happens a lot in data over time, when the outcome last year explains the outcome this year, for example.
- Autocorrelation also happens when we measure the same thing in multiple ways and include them all in the model.

Diagnosis: In SPSS, tick the “Durbin-Watson” box in the “Statistics” options and inspect the Durbin-Watson statistic.

- D-W = 2 (no autocorrelation),
- D-W > 2 (negative autocorrelation),
- D-W < 2 (positive autocorrelation)
- When D-W is between, like, 1.5 and 2.5, you're probably fine.

Solution: Remove obviously autocorrelated things (like, if you have an independent variable, *gdp04*, in a model that could be predicting another independent variable, *gdp08*, in the same model). Really think through your theory and examine your measures to ensure that you haven’t included multiple measures or indicators of the same thing. You can also add or subtract covariates from the model to mitigate some of the autocorrelation.

6. Normally distributed errors

OLS assumes that the mean of the error terms is zero, and that the errors are normally distributed around that mean of zero. If they're not normally distributed around a mean of zero, then we can't trust our predictions.

Diagnosis: In SPSS, select the “Plots” option, then request a “normal probability plot” and inspect how much the plot deviates from the diagonal zero line (less deviation is better).

Solution: Cry. Just take my word for it; it helps. Then, explain the issue in your analysis so you and readers can take the results of the model with a grain of salt.