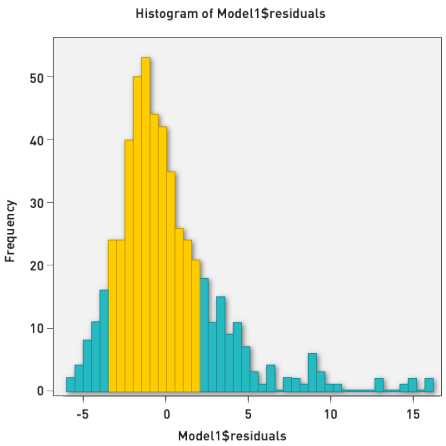
Testing Normality

The normality assumption is important for statistical inference.

- To test it, examine your regression diagnostics after you fit a linear regression.
- Remember: Residuals are estimates of error; we want to see if they look normal.
- Example: our fitted wage model from earlier
  wage = −3.39 + 0.644educ + 0.070exper + *u*

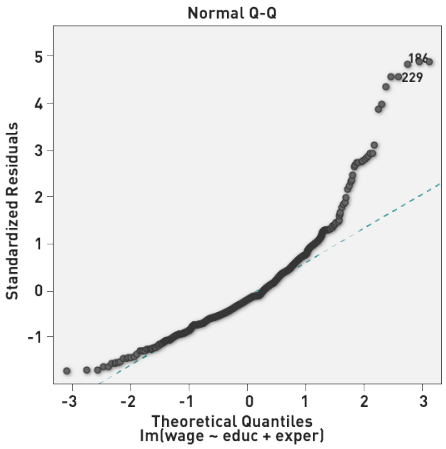


• Note the positive skew; many respondents with unusually high wages
• This is evidence against normality.

Testing Normality: Q-Q Plot

R also provides a q-q plot of the residuals.

- For each data point, the *y*-coordinate is its standardized residual (i.e., residual divided by the standard deviation of the residuals).
- The *x*-coordinate is what the standardized residual would be if the errors were perfectly normally distributed.
- For normally distributed errors, expect to see a perfect diagonal line.
- The more the plot deviates from the diagonal, the less normal the residuals.



• Note that the positive skew shows up as a high slope on the right.
• This isn't a terrible q-q plot, but it shows evidence of non-normality.

Testing Normality: Shapiro-Wilk

Finally, you could run a normality test like the Shapiro-Wilk test on your residuals.

- The null hypothesis in this test is that errors are normal.
- Be careful when you interpret the results.
    - They don't directly tell you how large the deviations from normality are; if you have a huge dataset, even tiny deviations will show significance.
    - In most scenarios, this is inevitable and it doesn't mean that the deviations are large enough to worry about.
    - A small dataset, say $n < 30$, makes it difficult to reject the assumption of normality, no matter what the distribution looks like.
- It's usually best to combine this test with a look at the diagnostic plots.

Responding to Normality Violations

First, if you have a large dataset, you can simply rely on the asymptotic properties of OLS.

- There's a version of the central limit theorem that says that OLS estimators are normally distributed for large sample sizes.
    - This means we need the Gauss-Markov assumptions for big datasets; we don't need the normality assumption unless we're aggregating the data into a small number of points.
- A large dataset is usually identified as $n \geq 30$, but the CLT doesn't tell us what n we need; it's about what happens as $n \to \infty$.
- With 30 observations, you're generally okay, but if you have less than 100 or so, it's good to examine the q-q plot anyway.
    - You should also examine your q-q plot if you suspect an unusually skewed distribution; the CLT takes longer to work in these circumstances.

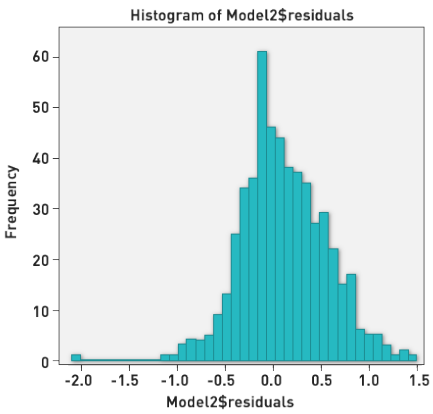
In the previous wage example, the error didn't look normal, but we had more than 30 observations, so we don't need normality to establish our sampling distributions.

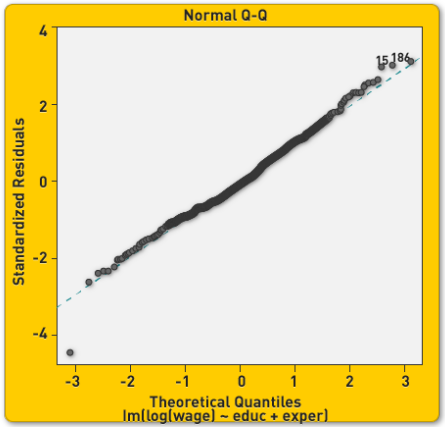
Responding to Normality Violations: Small Datasets

Next, if you do **not** have a large dataset and your residuals don't look normal, a good next step is to look for an alternate specification that meets normality.

- It often helps to transform your *y* variable: if *y* is skewed, the residuals will often be skewed too, so we might use the log of *y* as the outcome instead; if the transformed variable is more normal, the residuals will often be more normal as well.
- Earlier, we suggested using log of wage in our wage model; even though we have a large sample, it makes theoretical sense.
    - Log(wage) = 0.21 + 0.09educ + 0.010exper + *u*

Responding to Normality Violations: Small Datasets (cont.)





- The same transformation helps with normality as well, as you can see in these plots.
- This is a nice example of a normal-looking q-q plot.

Responding to Normality Violations: Normal Errors and Zero-Conditional Mean

- If the residual versus fitted value plot also shows curvature, you have a violation of both normal errors and zero-conditional mean. You might be able to correct both using a more flexible functional form.
  - E.g., add a quadratic term on the right, fitting a parabola instead of the line.
- You may be able to improve things by adding an appropriate predictor variable. (This can change the interpretation of the regression and often requires domain expertise.)
- Why place this strategy after relying on asymptotics?
  - We often want to draw some understanding from our fitted model.
  - Our top priority when choosing variables is matching our intuition and exposing the effects we want to measure.
  - Changing the specification to achieve normality forces us to compromise these goals, so we only recommend doing this when necessary.

Responding to Normality Violations: Bootstrapping

Another option is to estimate sampling distribution of our coefficients via bootstrapping.

- The basic idea: resample from our dataset in order to estimate the sampling distribution of our coefficients.
- If you had infinite resources and time and wanted to know what your sampling distribution looked like, you could collect a huge number of samples and plot the estimate you get for each one on a histogram.
    - This would approach the true sampling distribution.

Responding to Normality Violations: Bootstrapping (cont.)

- To bootstrap, we simulate repeated samples from the population by resampling from our one existing sample.
    - Each simulated sample has *n* data points, but we replace them as we draw, so some are drawn multiple times and each resample looks different.
    - This is a very general method that can help us estimate sampling distributions in all kinds of statistical procedures.
- We normally wouldn't use it for OLS regression, though.
    - Bootstrapping also relies on asymptotic properties in order to work; in order to know that, our bootstrap samples approximate real population samples—if we can't use the CLT, our bootstrap results may be questionable as well.

In most cases, we recommend using OLS asymptotics or altering the modeling specification to achieve normality.