Unit 4: OLS Inference

Reading: Wooldridge Chapter 4-5

From OLS Estimation to Inference

OLS Estimation

- Last week, we saw how the first 4 Gauss-Markov Assumptions let us estimate the parameters of our linear population model with OLS
- Here's a very high-level summary of these assumptions

MLR.1 y is a linear function of the x's
MLR.2 Sampling is random
MLR.3 No perfect multicollinearity
MLR.4 Errors have zero mean conditional on x's.

- These four assumptions guarantee that OLS estimates are unbiased.
- We also talked about replacing MLR.4 with MLR.4'

MLR.4' All x's are uncorrelated with the error.

- For large samples, this is the more critical assumption since it guarantees that coefficients are consistent.

OLS Efficiency

- Consistency is important, but it only tells us that we're right in expectation
  - Every time we take a sample and compute coefficients, they'll still be wrong by some amount.
- So far, we've said nothing about how close our coefficients are to the true values
  - Without this, there's no sense of scale, or of how meaningful our estimates are
  - We compute a slope of 2, but could it be 3 if we repeated the experiment? What about 300? Are we convinced that the real value is not 0?
  - Our coefficients are random variables, because our x_i's and y_i's are all random variables.

- Our next step is to say something about how much our estimates will vary from one sample to the next.
- This is captured by the variance of each coefficient, $var(\hat{\beta}_j)$
- Computing this will be important for a variety of statistical inference tasks.
- To do this, we're going to add one more important assumption.

Artist: I sometimes write a ^ after a variable, but it's supposed to go right over the variable.

Homoskedasticity

Our new assumption is called homoskedasticity.

- It's the fifth and final Gauss-Markov assumption.
- **Assumption MLR.5 (Homoskedasticity)**
  - The variance of the error term is constant

    $\text{var}(u_i|x_1,x_2,...,x_k) = \sigma^2$

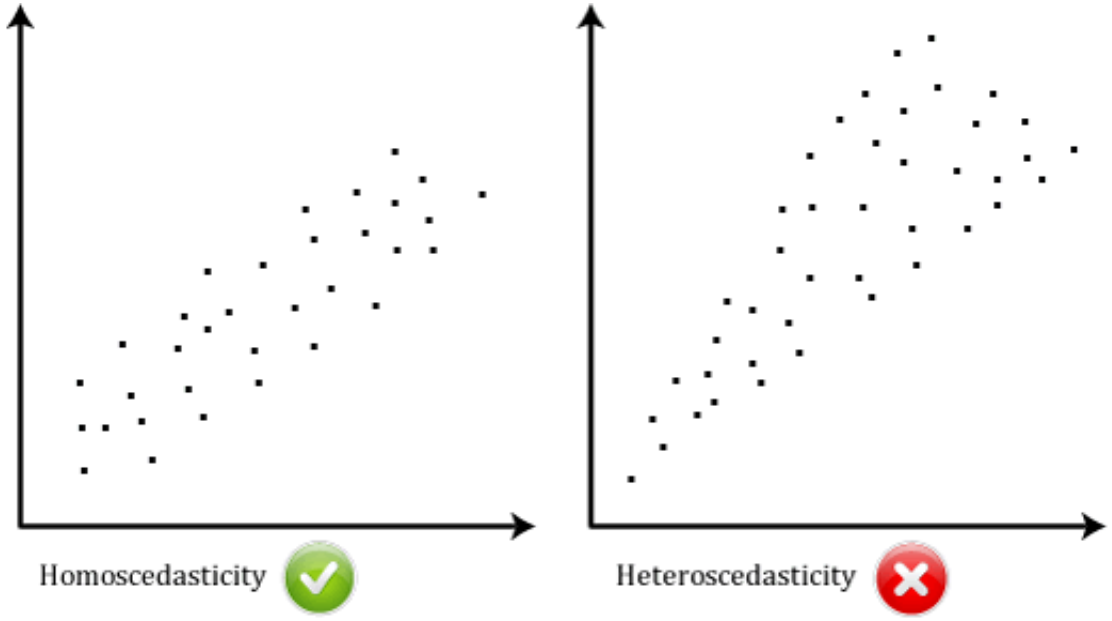
- In other words, the error term can't vary more for some values of our x's than others.
  - If you think of the error as including all unobserved factors, we mean that these factors vary an equal amount for all values of our x's.
- In other words, the values of the explanatory variables must contain no information about the variability of the error.
- This is a strong assumption, you'll find many real data sets for which this is unrealistic

Homoskedasticity

Graphically, homoskedasticity is easy to spot on a residual vs. predictor or residual vs. fitted value plot.

Remember that the residuals are our estimates of the error, so we look to see if these have constant variance.

For a fixed, x, imagine taking a vertical slice through these plots. The thickness of the band indicates the variance and this should be the same for all x's.



The left graph looks homoskedastic. On the right, variance seems to increase from left to right. It could also decrease, or increase and then decrease, etc.

Sampling Variance of OLS Esti

Artist, please highlight the 3 parts of the equation to draw students' attention

- **Theorem 3.2 (Sampling variances of OLS slope estimators)**
- Under assumptions MLR.1 – MLR.5, we can compute an exact formula for variance of our slope coefficient
- There are three terms that affect variance:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, \quad j = 1, \ldots, k$$

Variance of the error term. The more the error varies, the more noise there is to throw off our estimates, so variance increases

SST_j is the total sample variation in explanatory variable x_j.

$$\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

The more variation in x_j we have to work with, the more precise our estimate.

This last factor includes the R-squared from a regression of x_j on all other independent variables
This is the fraction of variation in x_j that cannot be explained by the other variables
So it's only the unique variation in x_j that's left in the denominator.
If we have multicolinearity, the unique variation will be small, and we'll lose precision.

Multicollinearity

- An example to illustrate how multicollinearity affects OLS variance
- We're modeling test scores as a function of expenditure on teachers, expenditure on instructional materials, and other expenditures.

$$avgscore = \beta_0 + \beta_1 teachexp + \beta_2 matexp + \beta_3 othexp + \ldots$$

- The different expenditure categories will be strongly correlated because if a school has a lot of resources it will spend a lot on everything.
  - In most schools, all expenditures will tend to be high, or all will tend to be low.
- To estimate the effect of, say, expenditure on teachers, we need information about situations in which this category changes differently from the other categories.

- As a result, sampling variance of the estimated effects will be large.
- We might decide to lump the categories together, or collect more data, depending on our goals

The Gauss-Markov Theorem

- Here's an interesting question: we have a formula for the variance of OLS coefficients. Could we do any better and get less variance?
- The famous Gauss-Markov theorem answers that:
- **Theorem 3.4 (Gauss-Markov Theorem)**
  - Under assumptions MLR.1 - MLR.5, the OLS estimators are the best linear unbiased estimators (BLUEs) of the regression coefficients.
- We already know what linear unbiased estimators are.
- Best means that OLS coefficients have the smallest possible variance.
- For any other linear unbiased estimator with coefficients $\tilde{\beta}_j$

$$Var(\hat{\beta}_j) \leq Var(\tilde{\beta}_j) \quad j = 0, 1, \ldots, k$$

- This theorem gives us a nice theoretical reason to use OLS.
  - It's the most famous benchmark for the performance of OLS.

More about BLUE

- Every letter in BLUE is necessary.
- OLS will only be best in the class of Linear Unbiased Estimators.
- If we look to estimators that aren't linear, we can get to even lower variance.
  - For example, when the variance of y changes with x, we can use a technique called weighted least squares to gain efficiency.
- If we look to estimators that are biased, we can also get to lower variance.
  - As a trivial example, we could use the estimators $\hat{\beta}_j = 0$
  - These are terrible estimators, but they do have zero variance
  - More seriously, there are sometimes biased estimators that are still consistent that can outperform OLS.
    - This is sometimes the case for maximum likelihood estimators

Troubleshooting Homoskedasticity

Testing for Heteroskedasticity

- Homoskedasticity is a very strong assumption, and it's rarely realistic for datasets we work with in this course.
  - Fortunately, this isn't a big problem for applied work because we've developed techniques that are robust to heteroskedasticity
  - In particular, our formula for the variance of OLS coefficients is wrong under heteroskedasticity, but you can use heteroskedasticity-robust standard errors instead.
    - Other names for these: Huber–White standard errors, Eicker–White, or Eicker–Huber–White
    - These researchers developed an expression for variance estimates that are unbiased without homoskedasticity.

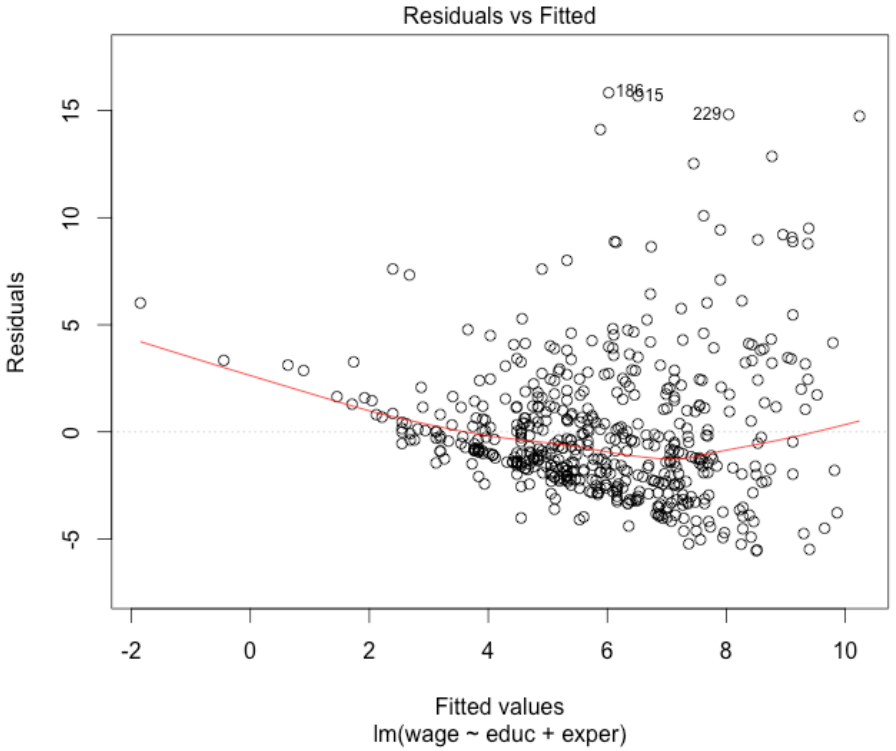
- Even though we can work around heteroskedasticity, it's worth talking about how to test for it.
- Usually, the best thing to do is look at your diagnostic plots.
- R provides two plots that can help identify heteroskedasticity

Testing Homoskedasticity

- First, look at the residuals vs. fitted values plot.
- Here's an example for a fitted wage model, using data from 1976.

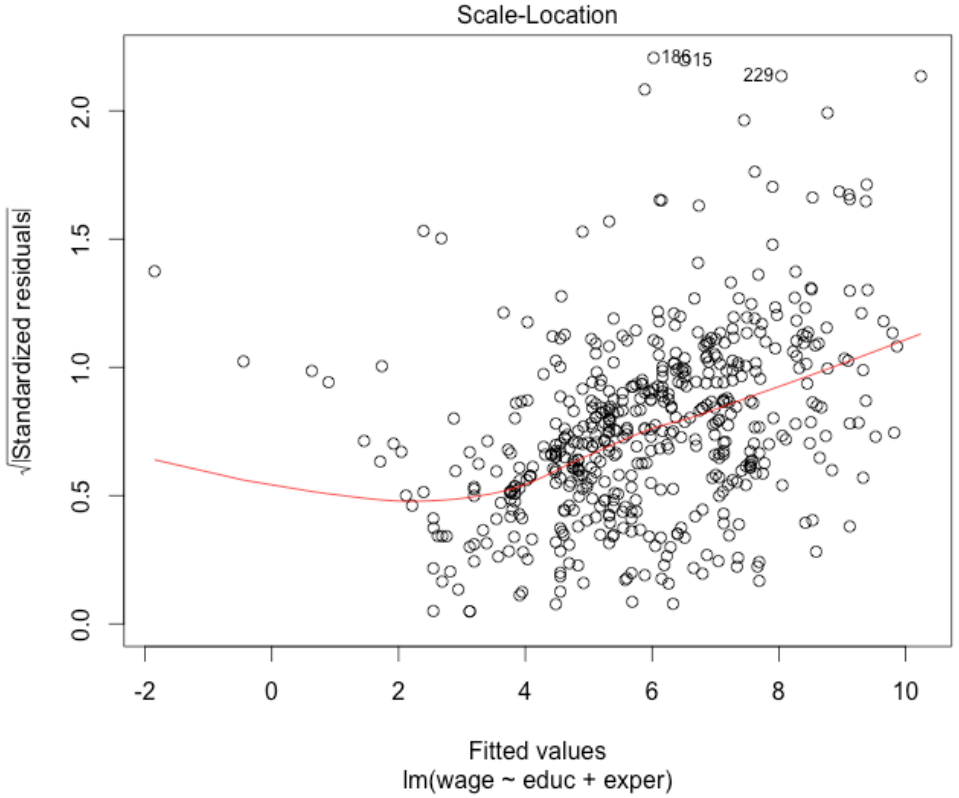
wage = -3.39 + 0.644educ + 0.070exper + u

- If errors are homoskedastic, we'd expect the band to have the same thickness from left to right.
  - Here, it seems to get thicker to the right.
  - It can be hard to tell if there are more data points on one side, making it look thicker.
- Remember that the red line is a smoother that approximates the mean of the residuals.
  - Here it seems to show a violation of zero-conditional mean
  - But the spline doesn't help with heteroskedasticity, since that's about the deviations from the mean.



Testing Homoskedasticity

- Second, R provides something called a scale-location plot
- This is related to the residual vs. fitted value plot, but we transform our residuals in two ways.
  - First, we take the absolute value.
    - Now more variance shows up as points higher on the y-axis
    - But at this point, there's a lot of skew – most points close to zero
  - Next we the square root.
    - This reduces skew, moving points away from the x-axis
- After these transformations, homoskedastic errors should show up as a horizontal band of points.
- Here's the scale-location plot for the same wage regression
  - Notice how the main band seems to travel upwards from left to right
  - R provides a red smoothing curve to help you see if it's truly flat.



Testing Homoskedasticity

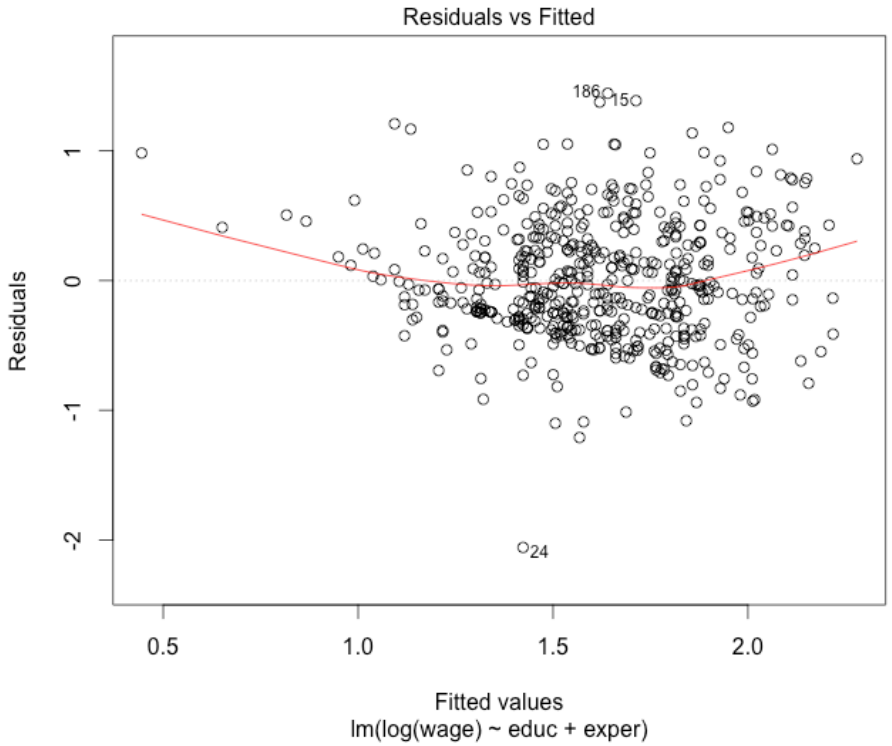
- There's also a statistical test that you can use to check for heteroskedasticity: the Breusch–Pagan test
- The null hypothesis is that there is homoskedasticity.
- If you get a significant result, you have evidence supporting heteroskedasticity.
- Warning: As with most tests of this nature, sample size matters a lot.
  - For a few hundred datapoints or more, almost any amount of heteroskedasticity will show up as a significant test, whether it is enough to worry about or not.
  - For small datasets of only 10 or 20 observations, the test will rarely be significant.
  - It's best to use this test keeping sample size in mind and in conjunction with the diagnostic plots.

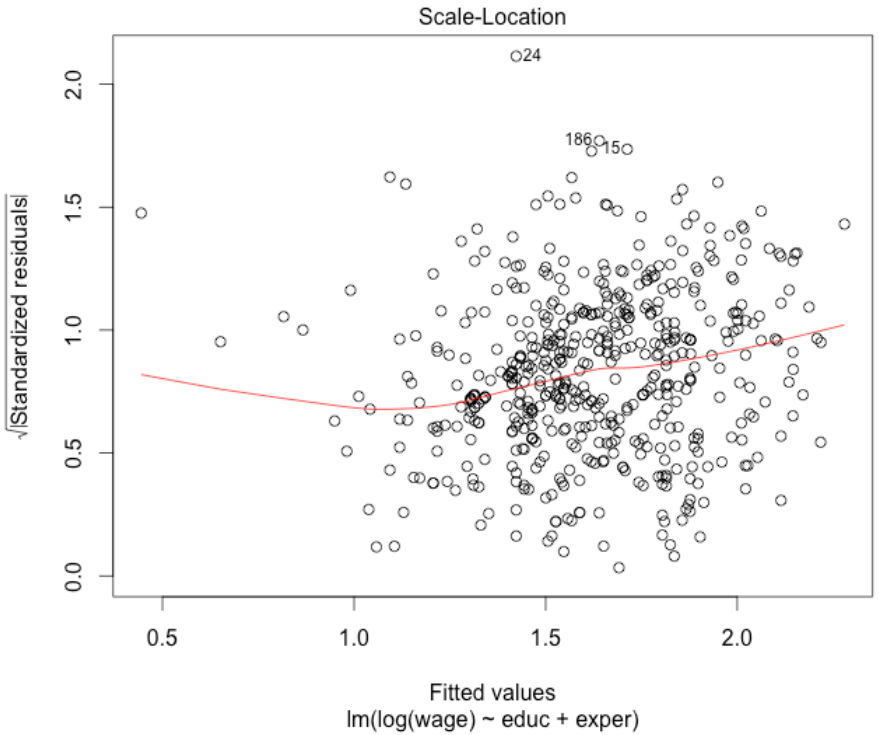
Responding to Heteroskedasticity

- If you detect heteroskedasticity, what should you do?
  - The simplest solution is to switch to heteroskedasticity-robust tools.
  - In particular, White standard errors are robust to heteroskedasticity.
    - Some researchers recommend using these by default, which is a good policy.
    - If you have homoskedasticity, White standard errors will be larger (more conservative) than regular standard errors, so it's slightly in your favor to use regular errors.

- At times, heteroskedasticity accompanies a violation of zero-conditional mean.
  - There may be an exponential relationship in the data.
    - Especially if var(y) seems correlated with E(y)
- This seemed to happen for our previous wage regression.
  - We predicted wage at the nominal level
    - We know there are a lot of outliers in the positive direction
    - There's also more variance on this side
  - More commonly, researchers will model the log of wage

let





- Here we've fitted the same model, replacing wage with the natural log of wage
  - Log(wage) = 0.21 + 0.09educ + 0.010exper + u.
  - Notice that the residual vs. fitted value plot looks more even.
  - The scale-location plot seems flatter too
    - (the left side of these graphs shows an uptick, but there are few datapoints there)
- Warning: in this case, it makes intuitive sense to take the log(wage)
  - We all have a sense for what a 10% increase in wage means
  - A $100 increase could be very different for different people
- We should not take the log of a variable just to fix heteroskedasticity if it doesn't make theoretical sense
  - We want our specification to be guided by our theory and tests that we wish to perform, not by what choices result in homoskedasticity.
  - That's why the best practice is simply to use heteroskedasticity-robust standard errors.

OLS Sampling Distributions

More Inference Tasks

- There's more to inference than just computing standard errors.
- Often, we want to test hypotheses.
  - Most commonly, we want to know if each parameters is significantly different from zero. Or could it just be noise?
  - We may also want to draw confidence intervals around our estimates
- Less commonly, we may want to estimate the likelihood function for our parameters.
    - We could use this for Bayesian updating, to measure evidence for one theory over another, or to draw likelihood intervals.
- For these tasks, it's not enough to know the variance of our estimates.
  - We need to know the actual shape of the sampling distribution.
  - Remember that the OLS estimators are continuous random variables, so they vary according to some probability density function
    - This is what we call the sampling distribution.
  - To have a sense of how meaningful our value is, we need to know the shape of the sampling distribution.

Sampling Distribution Shape

- In classical statistics, there are two major paths that we can take to establish the shape of the sampling distribution.
  - In both cases, our conclusion will be that the sampling distribution is normal.
  1. We add an assumption that errors are normally distributed in the population model.
     - Of course, this is a very strong assumption, and we'll have to check how realistic it is.
  2. If we have a large sample, we rely on the asymptotic properties of OLS, including the Central Limit Theorem (CLT), which tells us that our sampling distributions will be normal for large sample sizes.
- Beyond classical statistics, there are techniques like bootstrapping that can estimate the shape of the sampling distribution, even when it's not normal, but we'll focus on the classical methods this week.
- We'll begin with the assumption of normal errors first, then look at large sample sizes.

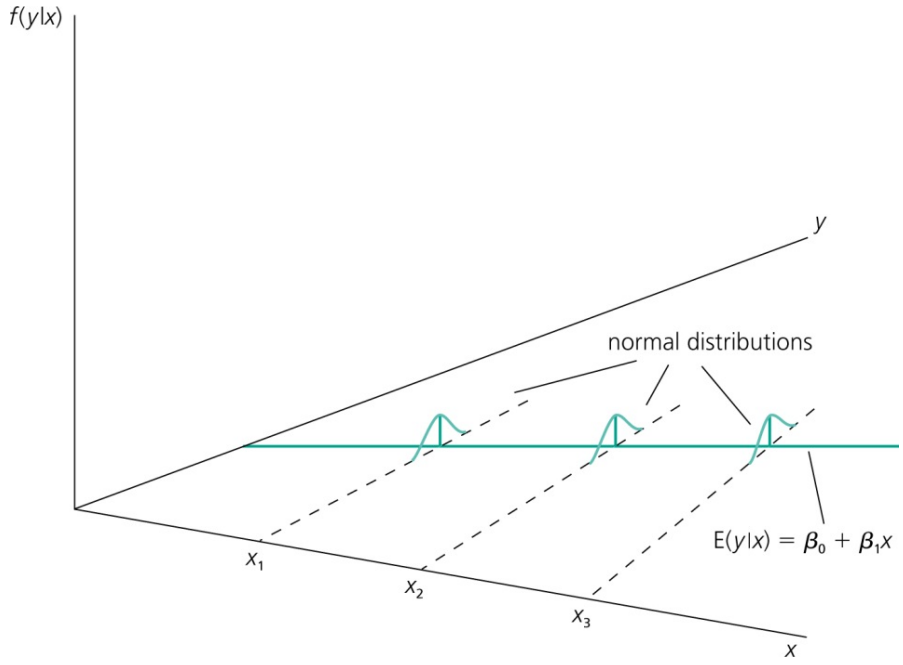
The Normality Assumption

Normal Error Term Assumption

- Our first technique to infer the sampling distribution of our OLS coefficients is to add an assumption about the shape of the error distribution –
- **Assumption MLR.6 (Normality of error terms)**

$$u_i \sim N(0, \sigma^2) \quad \text{independently of} \quad x_{i1}, x_{i2}, \ldots, x_{ik}$$

- We assume that the errors are drawn from a normal distribution with mean zero.
- We also assume that the errors are independent of our x's, so the distribution looks the same conditional on any values of the x's.
- This figure shows a graphical depiction of the normality assumption.
  - You can see the regression line, with the distribution of errors drawn for three values of x.
  - Each distribution has the same normal shape.



Normal Error Term Assumption

- How realistic is the normality assumption?
- There's a purely theoretical argument that says that the error term is the sum of many different unobserved factors
  - Sums of many independent factors are normally distributed by a version of the central limit theorem.
- Problems:
  - How many different factors? Number large enough?
  - What if a few factors are much more influential than others?
  - How independent are the different factors?
- In practice, we often look at the residuals and they're not normal at all.
    - If we have a highly skewed y variable, the errors are often skewed as well.
- We'll soon talk in detail about how to test for normality.
- For now, bear in mind that this is a rather strong assumption.

Classical Linear Model Assumptions

- When we add in normality to the 5 Gauss Markov assumptions, we have a collection of assumptions known as the classical linear model (CLM).

$\underbrace{MLR.1 - MLR.5}$

Gauss-Markov assumptions

$\underbrace{MLR.1 - MLR.6}$

Classical linear model (CLM) assumptions

- **Theorem 4.1 (Normal sampling distributions) states:**
- Under assumptions MLR.1-MLR6, the OLS coefficients are normally distributed.

$$\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j))$$

- Each β_j is normally distributed around the true parameter.
  - We already calculated the variance of β_j earlier, so we know the exact distribution

Standardizing the Distribution

- How can we use normality to test hypotheses?
- As a first step, we can normalize our estimator by subtracting its mean and dividing by its standard deviation.
  - This gives us the standard normal distribution.
  - Note that no matter what the true parameters are, we always get the exact same distribution.

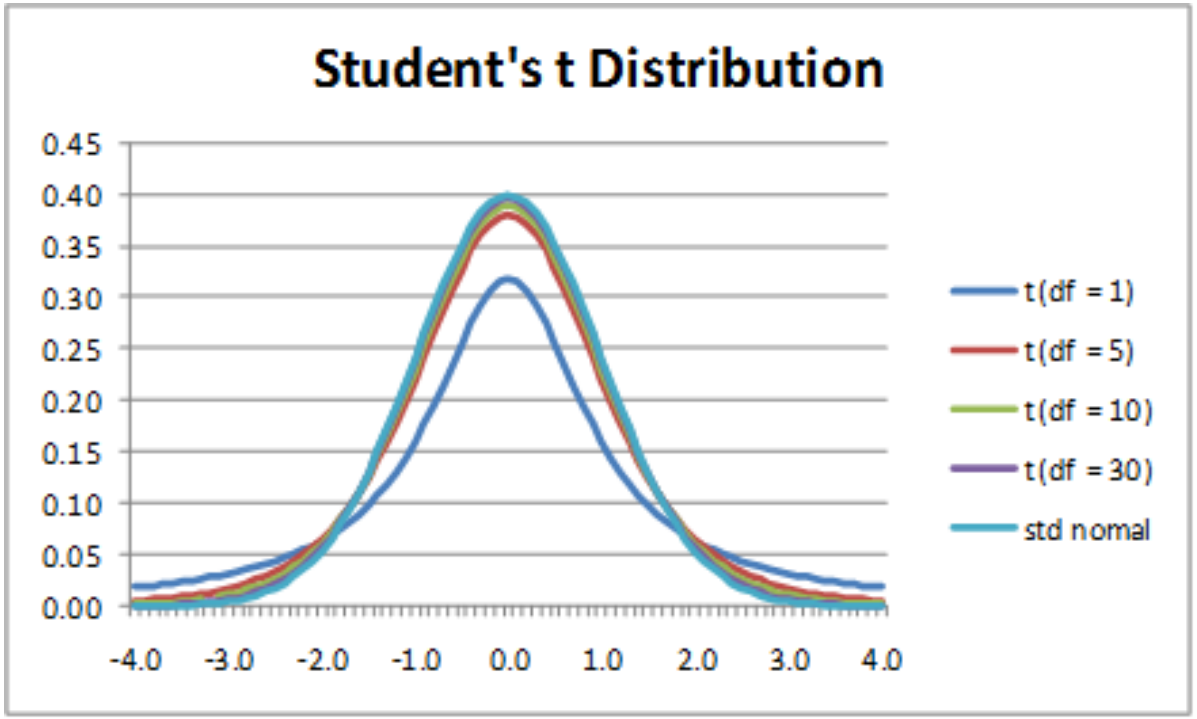
$$\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim N(0,1)$$

- In practice, however, we don't know the standard deviation to put in the denomenator,
  - We have to estimate it using the standard error of our sample.
- When we do this, it changes the normal distribution to a t-distribution.

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

  - The t-distribution is similar to a normal distribution, but the process of estimation introduces more variance and makes the tails of the distribution heavier.

t-Distribution



- In fact, there is an entire family of t-distributions.
  - The exact distribution depends on how many degrees of freedom we have.
  - Degrees of freedom appear quite frequently in statistics. They are equal to the number of data points we have minus the number of parameters we are estimating.
    - In this case, n-k-1, since we have k slope coefficients plus 1 intercept.
- Note that the t-distribution is close to the standard normal distribution if n-k-1 is large.
  - For reasonably large datasets (n > 30), we'd get essentially the same results with a normal distribution and a t-distribution
    - But we generally use a t-distribution to be fully correct and because it's just as easy to do in R.

Formulating a Null Hypothesis

- At this point, we have enough machinery to formulate a null hypothesis.
- Most of the time, our null will be that the population parameter is equal to zero.
  - After controlling for all other independent variables, there is no effect of x_j on y.
- H_0: $\beta_j = 0$.

- We're working in the frequentist framework. Our null hypothesis is specific enough (given our population model) to identify the distribution of our standardized coefficient.

$$\frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

- It's distributed according to a t-distribution and so we call it a t-statistic.

The t-Statistic

- Let's summarize how this all fits together.
- We collect our data and compute our OLS estimate, $\hat{\beta}_j$
- The further our estimate is from 0, the more evidence we have against our null hypothesis. But we have to normalize by a measure of variability.
- We therefore divide by an estimate of standard deviation to get our t-statistic,

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

- The t-statistic measures how many estimated standard deviations the estimated coefficient is away from zero.
- It's distributed according to a t-distribution, so we can talk about how significant the difference is.
  - We'll look at that process next.

One- and Two-Sided Tests

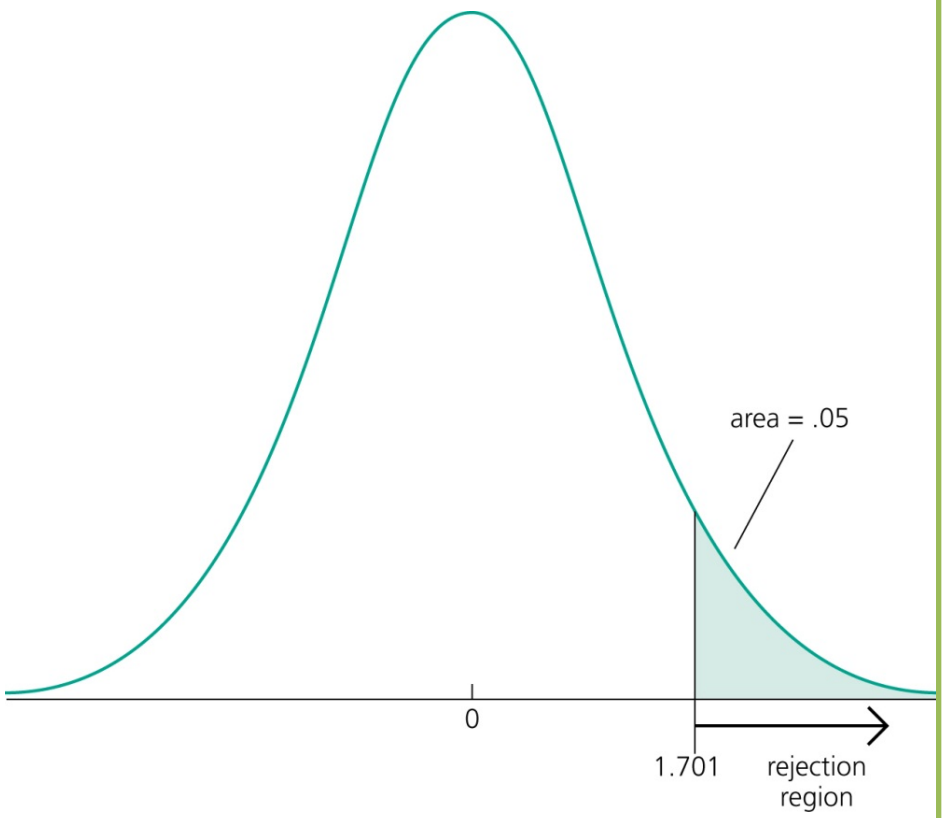
Testing a Null Hypothesis

- Remember that in the frequentist framework, we assume that the null hypothesis is true.
  - This tells us what distribution to expect for our statistic.
- We then select a decision rule: if the test statistic falls into a certain rejection region, we reject the null.

- Note: we have to choose our rejection regions before we see the data.
  - The rejection regions represent extreme values of our statistic
  - The chance of our statistic falling into a rejection region, assuming the null is true, is our type 1 error rate, α.
    - This is generally set to 0.05.
- In the context of t-tests, there are two main choices of rejection regions.

T-Statistics: One-Sided Test

- There are situations in which we are only interested in testing to see if our coefficient is positive (negative works the same way).
- We would need a strong theoretical reason to say why only positive coefficients are of interest to us.



In this case, we say that we are testing H_0 against the alternate hypothesis, H_1: $\beta_j > 0$.

This is a one-sided test. The entire rejection region is on the right tail. It stretches from a critical value to infinity.

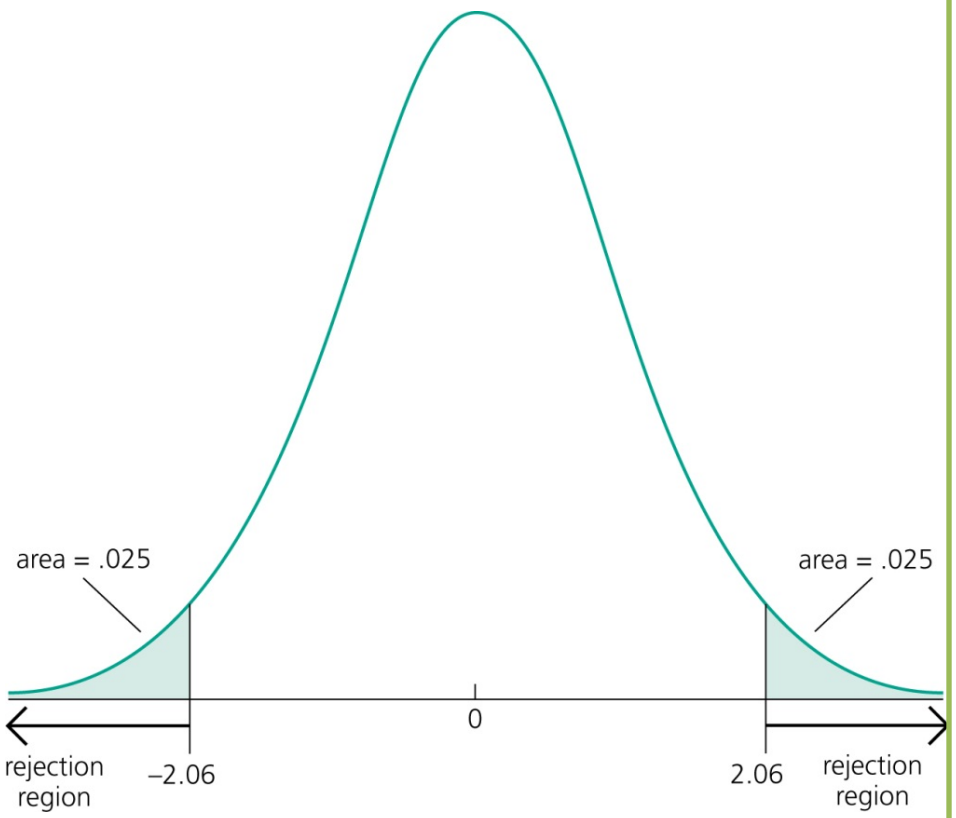
The critical value is chosen so that, if the null hypothesis is true, the statistic falls in the rejection region 5% of the time.

In the given example, this is the point of the t-distribution with 28 degrees of freedom that is exceeded in 5% of the cases, 1.701

Reject if t-statistic greater than 1.701

T-Statistics: Two-Sided Test

- More commonly, we don't have a strong enough justification to favor one direction over another. In this case, we set up a two-sided test.
- We have two rejection regions, which are symmetric on either side of zero.



Here, we say that we are testing H_0 against the alternate hypothesis, H_1: $\beta_j \neq 0$.

Construct the critical values so that, if the null hypothesis is true, the statistic falls in the right rejection region 2.5% of the time and the left 2.5% of the time.

In the given example, these are the points of the t-distribution that are exceeded in 2.5% of cases.

Our rule is reject if the t-statistic is less than -2.06 or greater than 2.06

As a rule of thumb, for large n, two-tailed critical values will be about 1.96 (or just remember 2)

One vs. Two-sided testing

- Generally speaking one-sided tests are rare.
  - We usually stick to two-sided tests
- Why?
  - A one-sided test means you really don't care if the statistic is on the other tail
    - What if you tested against $H_1: \beta_j > 0$, and got a t-statistic of -20. Would you really say that you failed to find evidence against the null?
  - The one-sided test may also indicate that you have a lot of confidence in your theory, but that confidence may not be shared by impartial researchers.
  - Finally, rejecting the null is easier with a two-sided test.
    - The critical value is closer to zero for a one-sided test, since it puts the entire rejection region on one tail
    - There's a temptation, if a two-tailed test fails, to switch to a one tailed test, which may then be significant.
      - Other researchers will therefore be less willing to trust one-tailed tests.
- For those reasons, we'll usually stick to two-tailed tests in this class.

t-Test Example

- In this example, we have a regression of college GPA on 3 explanatory variables: high school GPA, ACT score, and number of lectures skipped per week.
- In parentheses, we have the estimated standard error for each coefficient.
- Dividing the coefficient by its standard error, we get the t-statistic.
  - Remembering our rule of thumb, if the coefficient is over twice the error, we're likely to have significance.

$$\widehat{collGPA} = \underset{(.33)}{1.39} + \underset{(.094)}{.412}\, hsGPA + \underset{(.011)}{.015}\, ACT - \underset{(.026)}{.083}\, skipped$$

$$n = 141,\ R^2 = .234$$

In this case, we have a large same size, so our critical values are approximately those of a standard normal distribution

$$t_{hsGPA} = 4.38 > c_{0.01} = 2.61$$

$$t_{ACT} = 1.36 < c_{0.05} = 1.98$$

$$\left|t_{skipped}\right| = \left|-3.19\right| > c_{0.01} = 2.61$$

- Here, we're using $c_{0.01}$ to represent the critical value at the .01 level.
- The effects of hsGPA and skipped are significantly different from zero, even at the 1% high significance level.
- The effect of ACT is not significantly different from zero at the 5% level.

What does Significance Mean?

- Remember that in the frequentist framework, we never know the probability that our null hypothesis is true or false.
- If a variable is statistically significant then we believe it would be unlikely to occur under the null hypothesis.
  - Specifically, the level of significance (5%, 1%, etc.) indicates the "frequency of false positives" against which we may assure ourselves
- Also note that statistical significance does not imply that our variable is practically significant
  - We have to take a separate look at the effect size and consider it in context
- Not all variables that are practically significant have statistical significance.
  - Sometimes, especially in small samples, the noise is simply too great
  - But for large n, we expect that practically important relationships will eventually emerge as significant.
- Not all variables that are statistically significant are practically significant.
  - With enough data, any relationship will eventually be significant, no matter now unimportant it is.

p-Values for t-Tests

4 minute oyster

p-Values for t-Tests

- Remember that a p-value represents the probability, assuming that the null hypothesis is true, of getting a statistic as large as one we actually observe.
- This is easy to see in a t-distribution
  - We'll assume we're doing a two-sided test.

We compute our test statistic, in this case 1.85.

Then we compute the area above 1.85 and below -1.85. this is the p-value.

You can see that the statistic is larger than the critical value whenever p is less than alpha. We're really measuring the same thing in another way.

So we could rewrite our decision rule as reject H_0 if $p < \alpha$.

But the p-value gives us more information than just rejecting or not.

You can think about the p-value is the significance level at which one is indifferent between rejecting and not rejecting the null hypothesis.

Small p-values tell us how small a type 1 error rate we could have set, and still rejected rejected H_0.

Confidence Intervals

Confidence Intervals

- We've seen two ways to express significance of our OLS coefficients: whether we can reject H_0, and with a p-value.
- There's another way to do this that can help our understanding: with an interval estimator.
- Remember that our t-statistic is distributed as a t-distribution.
- Let's write $c_{0.05}$ for the critical value at a .05 confidence level.
- Then there's a 95% probability that

$$-c_{0.05} < \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} < c_{0.05}$$

- We can rearrange this to get the following expression:

$$\hat{\beta}_j - c_{0.05} se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + c_{0.05} se(\hat{\beta}_j)$$

- So there's a 95% chance that the following interval contains β_j.

$$\left[\hat{\beta}_j - c_{0.05} se(\hat{\beta}_j),\ \hat{\beta}_j + c_{0.05} se(\hat{\beta}_j) \right]$$

- This is called the 95% confidence interval.

Understanding Confidence Intervals

We have to be careful when interpreting a confidence interval. Some things to keep in mind:

- The confidence interval is a random variable.
  - It doesn't just move up and down, it will have different widths from sample to sample.
- A 95% confidence internal does NOT mean there's a 95% chance that β_j is inside.
  - That would be a Bayesian statement, and we're working in a frequentist framework.
  - B_j is fixed, it's the interval that's varying, and the probability is over possible values of the interval.
  - For all the same reasons that we don't know the probability that our null hypothesis is true, we can't know the probability that our real parameter is in any one confidence interval.
    - To do this, we'd need to introduce a prior subjective belief about where our coefficient was.
- The correct way to understand the 95%: If we take repeated samples and construct a 95% confidence interval for each one, 95% of them will cover the population parameter.
- Another useful way to think about confidence intervals: the 95% confidence interval is the set of all parameter values that we could not reject at the 95% level.
  - This means that a coefficient is significant exactly when our confidence interval does not include zero.
- Even though confidence intervals have shortcomings, a lot of researchers advocate for them, especially compared to p-values alone. They give us sense of scale that can help us understand how meaningful our results are.

Troubleshooting the Normality Assumption

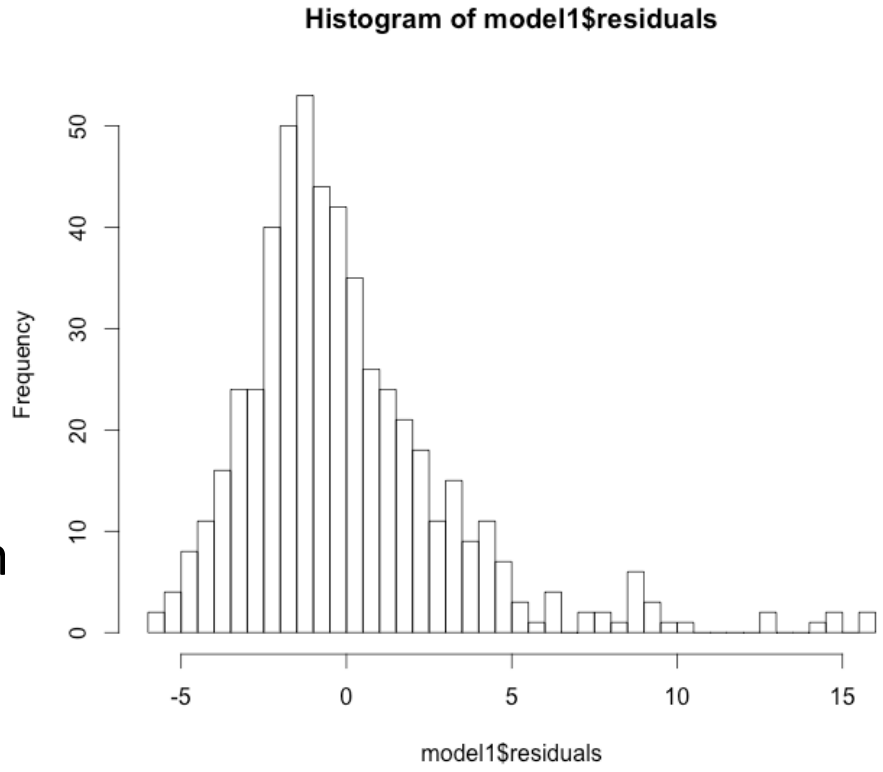
Testing Normality

- The normality assumption is important for statistical inference.
- To test it, after you fit a linear regression, examine your regression diagnostics.
  - Remember that residuals are our estimates of the error, so we're looking to see if the residuals look normal.

- Example: our fitted wage model from earlier

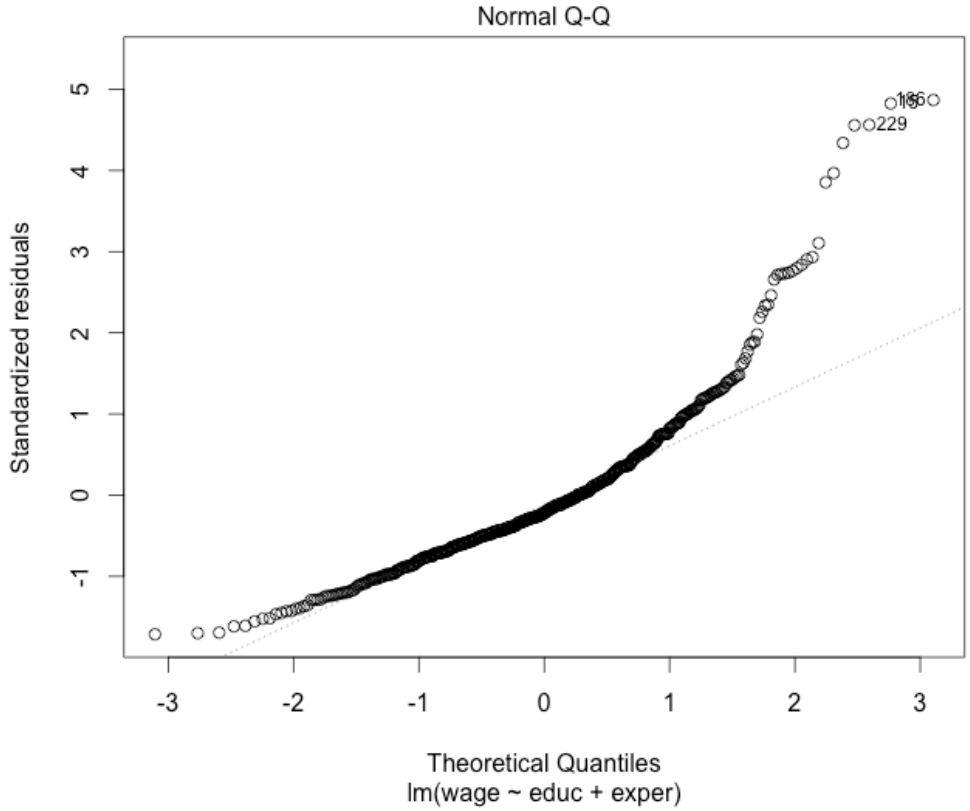
wage = -3.39 + 0.644educ + 0.070exper + u

  - Here's a histogram of the residuals.
    - Note the positive skew – many respondents with unusually high wages.
    - This is evidence against normality



Testing Normality

- R also provides a qq plot of the residuals.
  - For each data point, the y-coordinate is its standardized residual.
    - 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, you expect to see a perfect diagonal line.
  - The more the plot deviates from the diagonal, the less normal your residuals.
- Here, you can see the qq plot for the wage model. Notice that the positive skew shows up as a high slope on the right side.
  - This isn't a terrible qq-plot, but it definitely shows evidence of non-normality.



Testing Normality

- 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.
  - You have to 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 from normality will make the Shariro-Wilk test significant. In most scenarios, this is inevitable and it doesn't mean that the deviations from normality are large enough to worry about.
    - If you have a small dataset, say less than 30 observations, it's very hard 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

What should you do if you violate normality of errors?

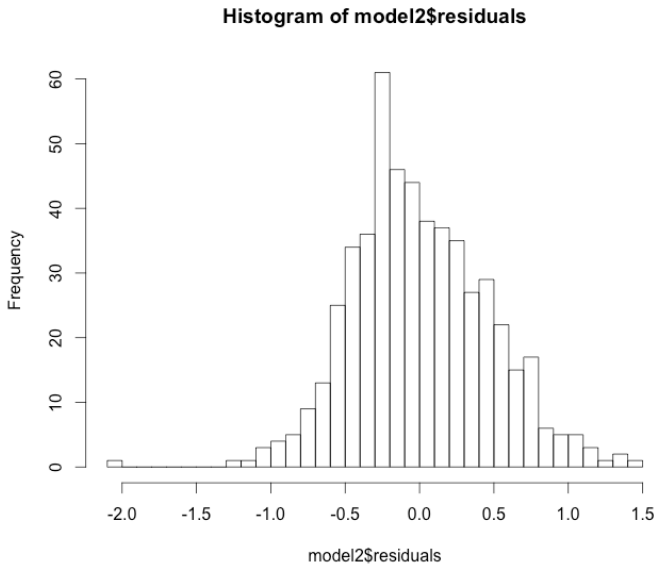
Here are some ideas, beginning with those we think you should try first.

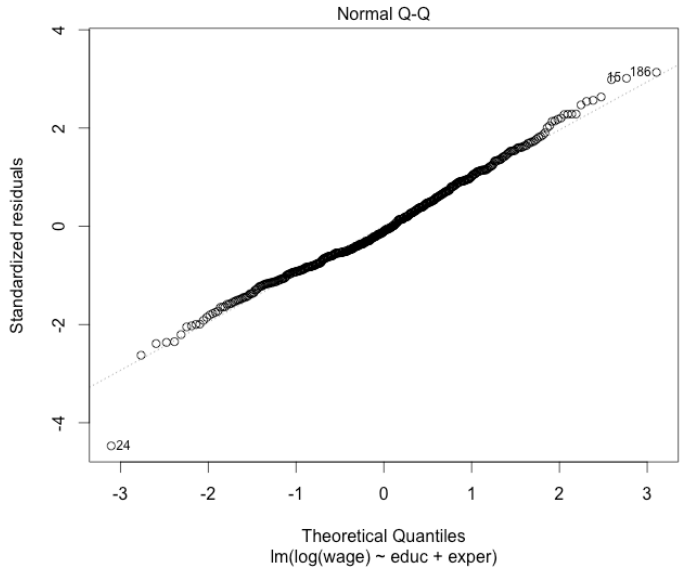
- First, if you have a large dataset, you can simply rely on the asymptotic properties of OLS.
  - There's a version of the central limit theorem that says that OLS estimators are normally distributed for large sample sizes.
  - This means that for big datasets, we usually don't require the normality assumption – the Gauss Markov assumptions are enough.
    - Unless we're aggregating the data into a small number of points
  - What constitutes a large dataset?
    - This is tricky – the CLT doesn't tell us what n we need, it's about what happens as n -> ∞.
    - The common rule of thumb is n = 30 is enough to invoke the central limit theorem.
  - If you have 30 observations, you're generally ok, but if you have less than 100 or so, it's a good idea to examine your qq plot anyway.
    - You should also examine your qq plot if you have reason to suspect an unusually skewed distribution – the CLT takes longer to work in these circumstances
  - In the wage example we just saw, the error didn't look normal, but we had over 30 observations, so we don't need normality to establish our sampling distributions.

Responding to Normality Violations

- 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 your y variable is skewed (e.g. GDP or household income) the residuals will often be skewed too. 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.
    - The same transformation helps with normality as well, as you can see in these plots.
    - This is a nice example of a normal looking qq plot.





Responding to Normality Violations

- If the residual versus fitted value plot 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.
  - For example, add a quadratic term on the right, so you're fitting a parabola instead of the line.
- Sometimes, you may be able to improve things by adding an appropriate predictor variable.
  - But this can really change the interpretation of the regression
  - Often requires domain-expertise

- Why do we place this strategy after relying on asymptotics?
  - In many cases, we 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. We want our model to reflect any guiding theory, and maximize our understanding of the results.
  - Changing the specification to achieve normality forces us to compromise these goals, so we only recommend doing this when necessary.

Responding to Normality Violations

Finally, another possible option is to estimate the sampling distribution of our coefficients through bootstrapping.

- The basic idea is we resample from our dataset in order to estimate the sampling distribution of our coefficients.
- If you had infinite resources and time and really 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.
- To bootstrap, we simulate repeated samples from the population by resampling from our one existing sample.
  - Each of our simulated samples has n datapoints, but we replace them as we draw, so some are drawn multiple times, and each resample looks different.
- Bootstrapping is a very general method that can help us estimate sampling distributions in all kinds of statistical procedures.
- At the same time, we normally wouldn't use it for OLS regression.
  - 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 central limit theorem, our Bootstrap results may be questionable as well.
- So in most cases, we recommend using OLS asymptotics, or altering the modeling specification to achieve normality.

OLS Asymptotics

Large-Sample Properties

- Let's take a few minutes to focus on the asymptotic properties of OLS
  - Often, you'll hear these called large-sample properties.
  - Since a lot of you will be working with huge datasets, it's good to summarize how these work.
- We've listed a lot of assumptions and a lot of these look rather daunting.
- It turns out that, as long as we have a large sample size, we don't need many of the stronger assumptions In the classical model
  - As long as we have a large sample and use heteroskedasticity-robust standard errors, we generally focus on MLR.1-3 and MLR.4'

Crucial Assumptions for Large Samples

- **Assumption MLR.1 (Linear in parameters)**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

- **Assumption MLR.2 (Random sampling)**

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : \; i = 1, \ldots n\}$$

  Data points are independent draws from population

- **Assumption MLR.3 (No perfect collinearity)**

- **Assumption MLR.4‘ (Exogeneity)**

  $Cov(x_j, u)=0$ for all j.

OLS Consistency

- We already know that under MLR.1-3 and MLR.4', OLS estimators are consistent.

$$\text{plim}_{n\to\infty}(\hat{\beta}_j) = \beta_j$$

- This means that we can always get the right answer if we collect an infinite number of datapoints.

Asymptotic Normality

- What about the shape of the distribution? The Central Limit Theorem tells us that our coefficients have an asymptotically normal sampling distribution.
  - The proof is tough, so we skip it here, but there's a sketch in an appendix of Wooldridge.
- The theorem in the book is stated under MLR.1-MLR.5,

Theorem 5.2 (Asymptotic normality of OLS)

Under MLR.1- MLR.5,

$$\frac{(\hat{\beta}_j - \beta_j)}{se(\hat{\beta}_j)} \overset{a}{\sim} N(0,1) \quad \text{also} \quad plim\ \hat{\sigma}^2 = \sigma^2$$

Note that if you use heteroskedasticity-robust standard errors, you can drop MLR.5.

Asymptotic Normality

- What does this mean from a practical standpoint?
  - As n increases, sampling distributions become normal
    - We can never see this happen, because we only get one sample, but the math tells us it's happening.
  - Since the t-distribution is asymptotically normal, it doesn't matter if we use a normal or t-distribution in stating our theorem.
  - This means that t-tests are valid for large samples.
  - The same is true for confidence intervals and F-tests.
- For large samples, we're left with two key assumptions that we need to focus on:
  - Random sampling. Are the observations correlated in some way? Is there clustering or a time dimension?
  - Exogeneity. Is any x correlated with the error? Is there some unmeasured factor that ends up in the error that's related to an x.
- Most of the rest of this course is about what to do when we can't meet these assumptions.