Multivariate Regression: Part 3

EC420 MSU

Justin Kirkpatrick
Last updated February 14, 2021

This lecture

Goal:

1. Review MLR.1 - MLR.6
2. Testing Hypotheses about $\hat{\beta}_j$.
3. Testing Hypotheses about $\hat{\beta}_j$ and $\hat{\beta}_k$
4. F-test
5. Testing for Heteroskedasticity

Gauss-Markov Regression Assumptions:

MLR.1	The population, y is a linear function of the parameters x and u: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$
MLR.2	The sample $(y_i, x_i) : i = 1, 2, \cdots, n$ follows the population model and are independent
MLR.3	No multicolinearity / "full rank": x_j is not a linear transformation of x_k for all j, k.
MLR.4	Zero conditional mean: $E[u\|x_1, x_2, \cdots, x_k] = 0$ for all x.
MLR.5	$Var[u\|x_1, \cdots, x_k] = \sigma_u^2$ for all x.
MLR.6	u is normally distributed ($u \sim N$)

Distribution of β

We now know that:

1. $E[\hat{\beta}_j] = \beta_j$ (MLR.1-MLR.4)
2. $Var(\hat{\beta}_j) = \frac{\sigma_u^2}{SST_{x_j}(1-R_j^2)}$ (MLR.5 or HC robust)
3. $\hat{\beta}_j \sim N$ (MLR.6)

$\Rightarrow \hat{\beta}_j \sim N(\beta, \frac{\sigma_u^2}{SST_{x_j}(1-R_j^2)})$

Therefore

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{Var}(\hat{\beta}_j)}} \sim t_{N-K-1}$$

Distribution of β

Most of the time, we are only interested in one hypothesis test:

$$H_0 : \beta_j = 0 \qquad H_a : \beta_j \neq 0$$

Which is a two-tailed test.

This makes the test statistic:

$$\frac{\hat{\beta}_j - 0}{\sqrt{\widehat{Var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var}(\hat{\beta}_j)}} \sim t_{N-K-1}$$

And Wooldridge uses $\boldsymbol{se}(\hat{\beta}_j)$ as the notation for the denominator.

This indicates a difference with $\boldsymbol{sd}(\hat{\beta}_j)$ which is the term when we *know* $\hat{\sigma}_u^2$.

The t-distribution is similar to the Normal

In fact, when there are around 30 degrees of freedom, it is identical (and we don't have to worry about that estimated $\hat{\sigma}_u^2$

Wooldridge calls this $t_{\hat{\beta}_j}$ "the" t-statistic

It is what we will use to test the null hypothesis that $\beta_j = 0$

Hypothesis testing with t

Let's think about that "standardization" when $H_0 : \beta_j = 0$:

$$t_{\beta_j} = \frac{\hat{\beta}_j - 0}{\widehat{se}(\hat{\beta}_j)}$$

What we're doing in the numerator, $\beta_j - 0$ is looking at the difference between *what we observed*, $\hat{\beta}_j$, and what we hypothesize is the true β_j.

Of course, our hypothesis is just *what we want to test* for β_j

If we could ignore the randomness of $\hat{\beta}_j$, we'd just compare $\hat{\beta}_j$ with $H_0 : \beta_j = 0$.

- That is, we'd say "is $\hat{\beta}_j$ zero?"

But $\hat{\beta}_j$ is random (thus, the ^)

Hypothesis testing - single parameter

Let's say we have a coefficient $\hat{\beta}_j = 5$ and $se(\hat{\beta}_j) = 2$

- We *know* how to calculate these from data

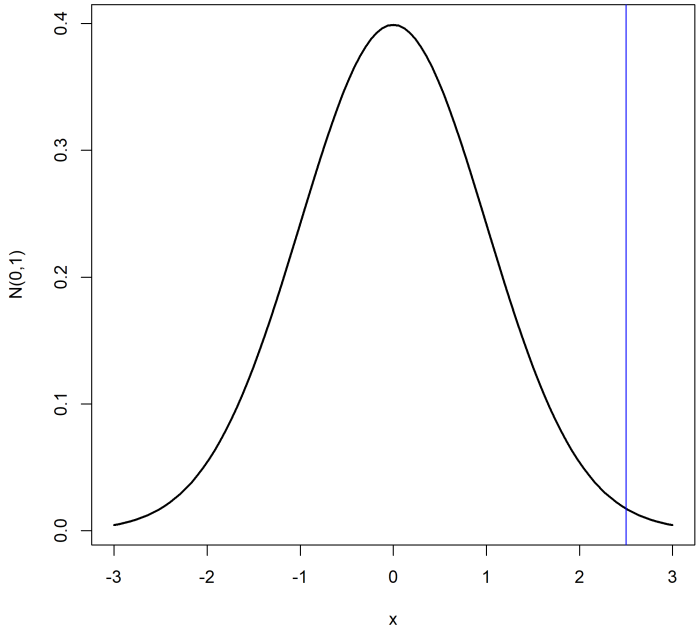
Let's say we have a $H_0 : \beta_j = 0$

Let's look at the (unstandardized) distribution of $\hat{\beta}_j$

Hypothesis testing - single parameter

And just for completeness, let's standardize our $\hat{\beta}_j$

$$\frac{\hat{\beta}_j - H_0}{se(\hat{\beta}_j)} = 2.5$$

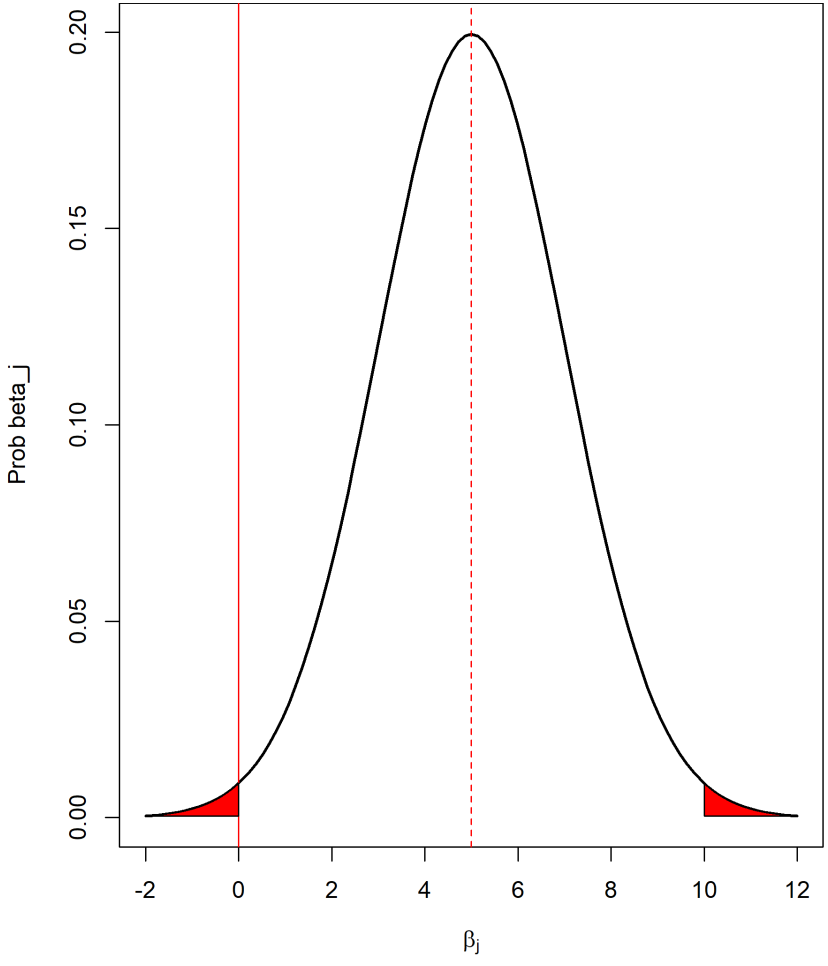


We could look at how likely it is to observe such an extreme value

- How? Well, the area under the curve tells us how likely we are to randomly draw that value from that distribution.
- This is a **two-tailed** test (remember the null hypothesis), so we need to look at the probability of getting something that extreme on the right side as well.
- Let's calculate just using the left side first.

Here I'm using N-K-1 = 300, which is very big

Since this is two-tailed, we want to see the probability of getting something *as extreme* (in either direction):



The red area is the p-value. It is small because it is very unusual to observe something as extreme as H_0 = 0 when $\hat{\beta}_j = 5$ and $\widehat{se}(\hat{\beta}_j) = 2$

We use the commonly-accepted threshold of 95%. That is, our hypothesis has to have a 5% chance *or less* of being observed in order for us to reject it.

- This means we reject any H_0 if the p-value is <.05

Rejection region and 95% Confidence Interval

We can also back-calculate the 95% range and look at what values would be rejected.

These are all equivalent:

- p-value < .05
- The hypothesized value of 0 lies in the rejection region
- absolute value of the t-statistic is greater than $\sim$ 1.96

They all happen when the hypothesized value is highly unlikely to occur under the $\hat{\beta}_j$ and $se(\hat{\beta}_j)$

When "highly unlikely" things occur, we reject the H_0 (null hypothesis)

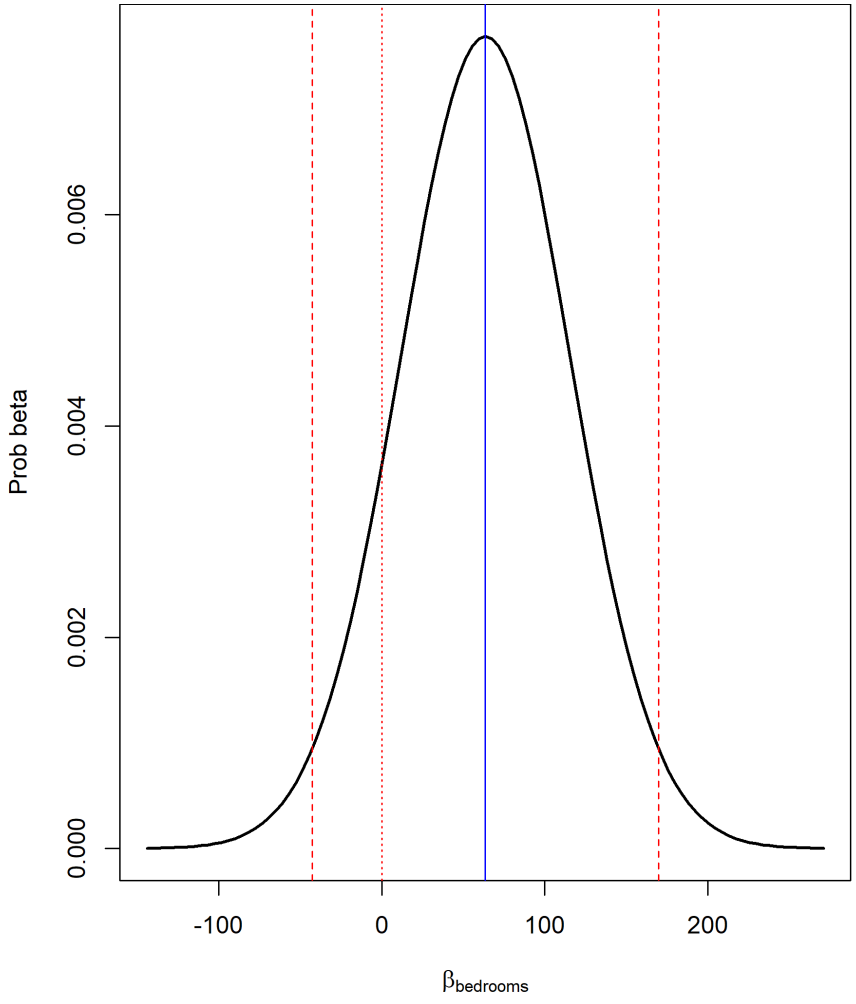
Let's look at an example: home price and bedrooms.

Hypothesis testing

$$HomePrice = \beta_0 + \beta_1 bedrooms + u$$

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  211.748    201.276   1.052   0.3018
## bedrooms      63.586     51.907   1.225   0.2308
```

Hypothesis testing

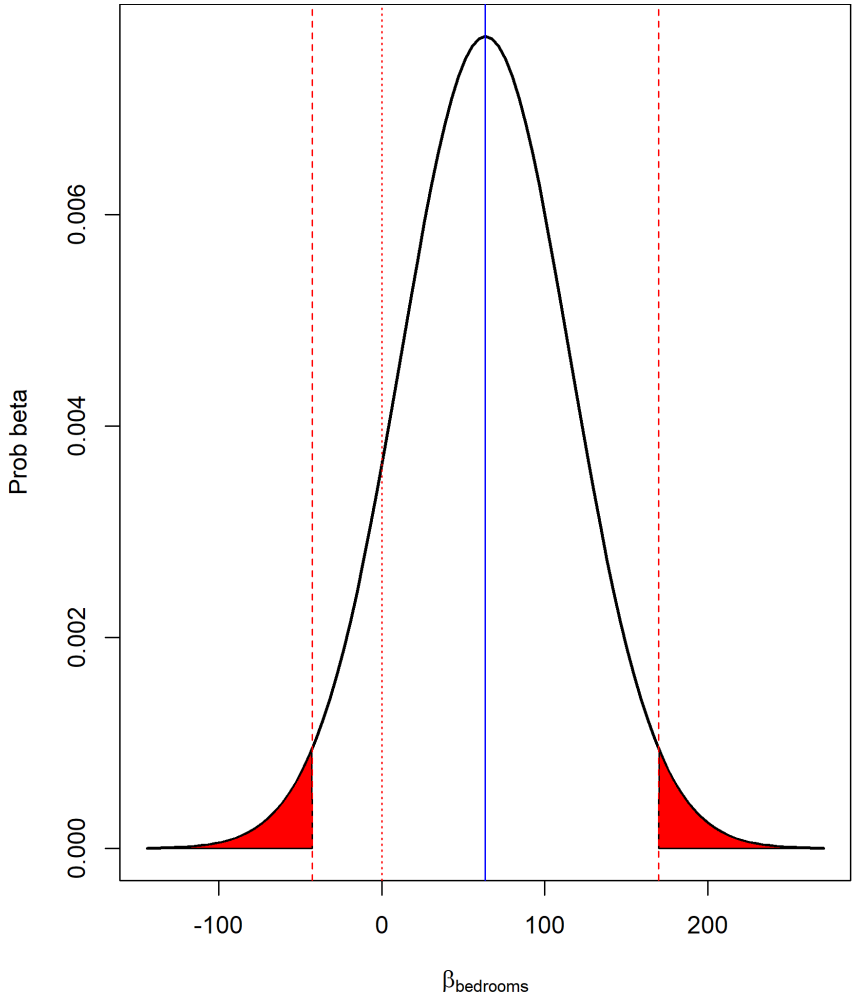


Here, $t_{crit,28} = 2.05$ taken from a table in Wooldridge for (.025, .975)

- The 95% CI is $63.59 \pm t_{crit,28} \times 51.91$
- 0 is within the 95% CI, so we fail to reject $H_0 : \beta_{bedrooms} = 0$
- The "rejection region" is outside of the red dotted lines

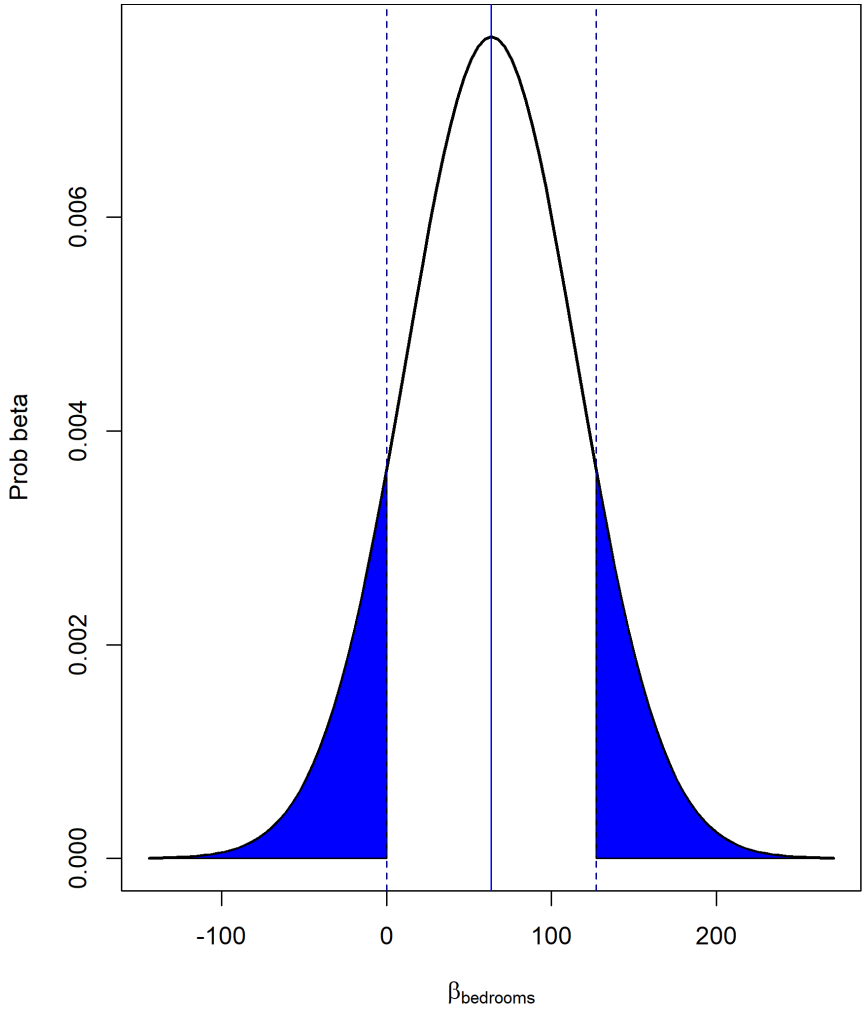
$t_{crit,28} = 2.05$ will be useful. Remember it.

Hypothesis testing



Red is the *rejection region* associated with our estimates $\hat{\beta}_{bedrooms} = 63.59$ and $se(\hat{\beta}_{bedrooms}) = 51.91$

Hypothesis testing



The area in blue is our p-value

- It is the area to the left of 0, and to the right of the just-as-extreme equivalent to zero.
- It is equal to $2 \times .115 = .231$.
- This is the (HC "robust") p-value Stata gave us.
- We *fail to reject* H_0 because the p-value is >.05

And finally, let's calculate

$$t = \frac{\hat{\beta}_{bedrooms} - 0}{se(\hat{\beta}_{bedrooms})} = \frac{63.59 - 0}{51.91} \sim t_{N-K-1}$$

The t statistic is 1.22

- Our $t_{critical, N-K-1} = 2.05$ from before.
- Our t-statistic is **not** larger in magnitude. We fail to reject

So we have failed to reject H_0:

- Because 0 does not fall outside of the *rejection region*
- Because 0 is within the 95% Confidence Interval
- Because the p-value is >.05
- Because $|t| < t_{critical, N-K-1}$

These are all equivalent, all lead to the same result.

Testing restrictions and multiple parameters

In economics, we sometimes want to test

- $H_0 : \beta_1 = \beta_2$
- $H_A : \beta_1 > \beta_2$
  - A one-tailed test

Let's look at how we can set this up using the hypothesis testing tools we have:

- $H_0 : \beta_1 = \beta_2$ is the same as $H_0 : \beta_1 - \beta_2 = 0$
- $H_A : \beta_1 - \beta_2 > 0$ follows.

Super - if we have $\hat{\beta}_1$ and $\hat{\beta}_2$, all we need is the $se(\hat{\beta}_1 - \hat{\beta}_2)$.

- This is **not, I repeat NOT** $se(\hat{\beta}_1) - se(\hat{\beta}_2)$.

The formula for the variance of the sum of two random variables

$$Var(\hat{\beta}_1 - \hat{\beta}_2) = Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1, \hat{\beta}_2)$$

And the se is the square root of the variance.

We have not been given the $Cov(\hat{\beta}_1, \hat{\beta}_2)$

- It is calculated by R in the variance-covariance matrix
- We won't worry about how. Just know that it is possible

R can be asked to generate that t-statistic and test it

```
vcov(myOLS)
```

Hypothesis testing

The example in Wooldridge, Section 4-4:

$$log(wage) = \beta_0 + \beta_{jc} jc + \beta_{univ} univ + \beta_{exper} exper$$

We'd like to test to see if *years in junior college, jc* have the same effect on log-wages as *years in university*, controlling for experience (ceteris paribus!)

We'd like to test $\beta_{jc} = \beta_{univ}$, which is:

- $H_0 : \beta_{jc} - \beta_{univ} = 0$
- $H_A : \beta_{jc} - \beta_{univ} \neq 0$

Hypothesis testing

Sure, we **could** rewrite the equation so that we get a coefficient that is equivalent to $\beta_{jc} - \beta_{univ}$:

$$log(wage) = \beta_0 + \beta_{jc}jc + \beta_{totcollege}(univ + jc) + \beta_{exper}exper$$

Since $\beta_{totcollege}$ captures both JC and University, β_{jc} captures the difference between the two - exactly what we want to test! The t-stat for that coefficient is our test.

But we could do it with a **linear hypothesis test**

Hypothesis testing

Of course, R does it for us as well without re-writing the equation:

```
library(car)
myOLS = lm(wage ~ jc + univ + exper, df)
linearHypothesis(model = myOLS, hypothesis.matrix = "jc - univ = 0" )
```

This tests if the coefficient β_{jc} is the same as β_{univ}.

- That is, it tests if the effect of jc is the same as the effect of $univ$
- It does so by calculating the $\hat{se}(\hat{\beta}_{jc} - \hat{\beta}_{univ})$

Hypothesis testing


```
NN = 400
df1 = data.frame(jc = rpois(2, NN),
                 univ = rpois(4, NN),
                 u = rnorm(NN, mean=0, sd = 5)) %>% dplyr::mutate(wage = 10 + 2.5*jc + 2

myOLS<-lm(wage ~ jc + univ, df1)
linearHypothesis(model = myOLS, hypothesis.matrix = 'jc - univ = 0')
```

```
## Linear hypothesis test
##
## Hypothesis:
## jc - univ = 0
##
## Model 1: restricted model
## Model 2: wage ~ jc + univ
##
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    398 10132
## 2    397 10113  1    19.118 0.7505 0.3868
```

Here, I create fake data where I know there are equal effects, and I test $H_0 : \beta_{jc} = \beta_{univ}$. We fail to reject ($p > .05$)

Hypothesis testing multiple linear restrictions

But what if we want to know if more than one coefficient is zero?

- Let's say we have run `lm(khomeprice ~ bedrooms + bathrooms + sqft, df)`
- And we want to know if $\beta_{bedrooms} = \beta_{bathrooms} = 0$.
  - Are all of these coefficients *jointly* zero?
  - Once we account for $sqft$, which is not being tested
  - Is the effect of $bedrooms$ **and** $bathrooms$ zero **all together**.

This differs from asking about each of the separately

- It might be that each one has no statistically significant effect, but taken together, they might jointly have some effect.
- It is also asking if these coefficients, together, explain much of y.

This is a *multiple linear restriction* test (W. 4.5)

We can do this in R with `linearHypothesis` as well.

Hypothesis testing


```
myOLS3 = lm(khomeprice ~ bedrooms + bathrooms + sqft, df)
coeftest(myOLS3, vcov=vcovHC(myOLS3, 'HC1'))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  78.272328 237.133362  0.3301   0.7440
## bedrooms      2.265803  50.977959  0.0444   0.9649
## bathrooms     0.926042   1.699717  0.5448   0.5905
## sqft          0.084879   0.105242  0.8065   0.4273
```

Hypothesis testing


```
linearHypothesis(myOLS3, c('bedrooms=0','bathrooms=0'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## bedrooms = 0
## bathrooms = 0
##
## Model 1: restricted model
## Model 2: khomeprice ~ bedrooms + bathrooms + sqft
##
##   Res.Df     RSS Df Sum of Sq     F Pr(>F)
## 1     28 2373617
## 2     26 2348144  2     25473 0.141 0.8691
```

We *fail to reject* the null hypothesis that both $\beta_{bedrooms}$ and $\beta_{bathrooms} = 0$ jointly.

Hypothesis testing

$H_0 : \beta_{bedrooms} = \beta_{bathrooms} = 0$

$H_A : H_0$ is not true

Rejecting the null hypothesis doesn't tell us which "part" of the hypothesis rejects.

- It doesn't say that bedrooms isn't zero,
- Or that bathrooms isn't zero.

We can think of these joint tests as *restrictions* - we are asking "do these coefficients, jointly, explain any of y?"

Testing if a group of coefficients all *jointly* equal zero is a special situation

Saying that β_1, β_2 are jointly zero is the same as saying that β_1, β_2 do not explain any variation in y.

- If they don't explain any variation in y (jointly), then *they can be left out of the model*.
- Testing if they are jointly zero is the same as testing if you can leave them out of the regression entirely.

So how would we test this?

Testing these "restrictions"

First, we run the *unrestricted* model:

```
lm(khomeprice ~ bedrooms + bathrooms + sqft, df)
```

Then, we take the SSR, the Sum of Squared Residuals

- $\sum \hat{u}^2$. Call it SSR_{UR}
- The unrestricted (UR) model also has degrees of freedom $N - K - 1 = N - 4 - 1$

Then, we run the *restricted* model:

```
lm(khomeprice ~ sqft, df)
```

- This is "restricted" because we are making a restrictive statement about $\beta_{bedrooms}, \beta_{bathrooms}$
- Specifically, we are saying **that they are equal to zero** in this model!
  - Doesn't get much more restrictive than that, does it?
- The restricted (R) model also has degrees of freedom $N - K - 1 = N - 1 - 1$.

We can compare the SSR of each model.

"Do the increased number of parameters (2) explain enough variance (reduce the variance of u) sufficiently to include them?"

What should our test do?

- If $SSR_R - SSR_{UR}$ is very big, then the unrestricted model (more β's) is more "explanatory", and that **set** of β's are not, jointly, zero.
  - Our test should reject the null that $\beta_{bedrooms} = \beta_{bathrooms} = 0$
- It should account for the difference in degrees of freedom

Here's our test statistic, F:

$$F = \frac{\frac{(SSR_R - SSR_{UR})}{q}}{\frac{SSR_{UR}}{N - K_{UR} - 1}}$$

Where q is the **number of restrictions we are testing**. Here, it is 2.

- Because we are testing $\beta_{bedrooms} = \beta_{bathrooms} = 0$
- q = $df_R - df_{UR}$

Here's our test statistic, F:

$$F = \frac{\frac{(SSR_R - SSR_{UR})}{q}}{\frac{SSR_{UR}}{N - K_{UR} - 1}}$$

F has a known distribution:

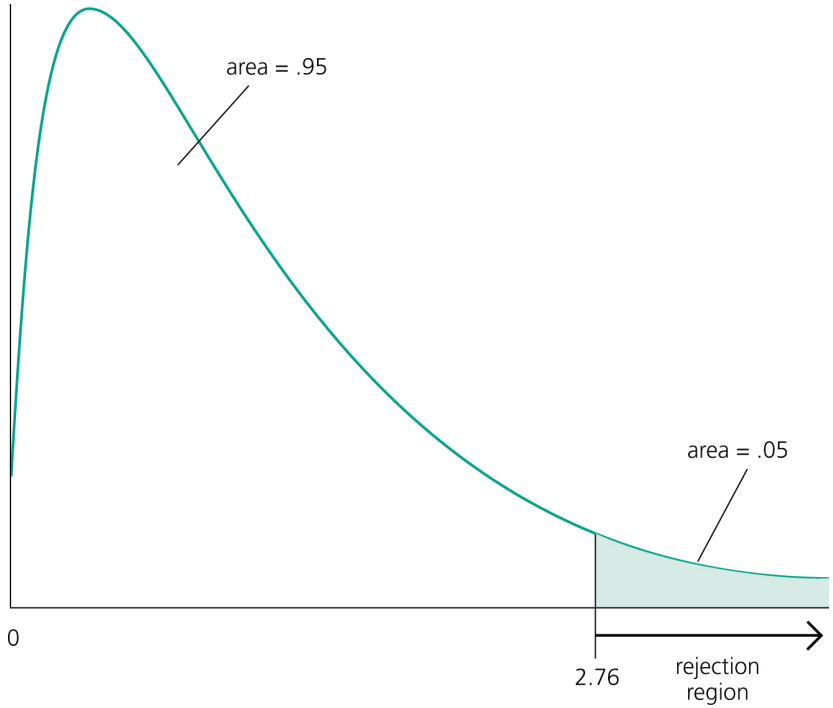
If you recall (you probably don't), we introduced an F-distribution back in stats review.

F is like t - it is defined only by its degrees of freedom.

F, unlike t, takes *two* degrees of freedom: the numerator $(q)$and the denominator $(N - K_{UR} - 1)$.

- $F \sim F_{q, N - K_{UR} - 1}$

Hypothesis testing



This is the $F_{3,60}$ distribution from Wooldridge Fig 4-7.

- The rejection region is always only on the right ($SSR_R - SSR_{UR} > 0$ always)
- When SSR_R is big relative to SSR_{UR} (the β's being tested explain a lot, making $\hat{u}^2$ smaller), the F-stat is larger
  - Which means it is further out to the right, closer or in the rejection region

Hypothesis testing

If F is big, it is more likely to be in the rejection region

When F is in the rejection region, we reject the $H_0 : \beta_{x_1} = \beta_{x_2} = \beta_{x_3} = 0$

When we reject H_0, all of these are true:

- Jointly, the coefficients are not all zero
- **The unrestricted model (the one with all coefficients in it) is a better model**
- It has sufficiently better explanatory power to justify the extra coefficients.

R automatically gives us an F-stat

It is the F test where the restriction is that all β's except the constant are 0

- Which is saying "does this model, with all it's coefficients and RHS variable x's, explain y any better than just using β_0.
- A model with only β_0 is equivalent to just guessing $\bar{y}$, and not using *any* of the x's.
- So the F test that R outputs is a test for whether or not all the coefficients (except the intercept, β_0) are zero.

```
##
## Call:
## lm(formula = khomeprice ~ bedrooms + bathrooms + sqft, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -532.68 -157.45  -35.56  205.13  615.72
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  78.27233  252.34541   0.310    0.759
## bedrooms      2.26580   57.98750   0.039    0.969
## bathrooms     0.92604    1.81426   0.510    0.614
## sqft          0.08488    0.13514   0.628    0.535
##
## Residual standard error: 300.5 on 26 degrees of freedom
## Multiple R-squared:  0.1535,    Adjusted R-squared:  0.05582
## F-statistic: 1.572 on 3 and 26 DF,  p-value: 0.2201
```

Hypothesis testing

So we know

- How to test a hypothesis about a single coefficient
- How to test a joint hypothesis about multiple coefficients
- How to test if many coefficients are jointly zero
- What the F test R gives us is testing:
  - Whether or not all the coefficients, jointly, are zero
  - Which is the same as saying whether or not all the coefficients, jointly, explain y any better than just using β_0

How do we know for sure

that we should be concerned about heteroskedasticity?

Like the F test for whether or not some coefficients are jointly zero, we have a test for heteroskedasticity.

We can use it to see if we need to apply our Heteroskedasticity-consistent errors (HC)

The test follows from the notion that σ_u^2 might increase **with one of our** X's.

Breusch-Pagan Test for Heteroskedasticity

1. Estimate our model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
2. Calculate $\hat{u}$ and $\hat{u}^2$.
3. Regress $\hat{u}^2 = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + v$
4. Check the p-value of the F statistic from this test.
    - If it is small, we reject the H_0
    - The H_0 is homoskedasticity
    - So rejecting $\rightarrow$ we have heteroskedasticity and should use HC (robust) errors

This works because we are doing a joint test for whether or not $x_1, \cdots, x_2$ explain (jointly) the magnitude (variance) of $\hat{u}$.