# Inference

William Brasic

The University of Arizona

# Inference

**Definition 1: Inference**

Inference is the process of drawing conclusions about a population's characteristics based on the sample data.

- We conduct inference by performing hypothesis tests and constructing confidence intervals.
- To make inferences in finite samples, we have to make assumptions regarding the sampling distribution of our estimator.

# Normality of Errors

**MLR Assumption 6: Normality of Errors**

For each $i = 1, \ldots, n$,

1. The population error $u_i$ is independent of $\boldsymbol{x}_i$
   - This implies $\mathbb{E}[u_i|\boldsymbol{x}_i] = \mathbb{E}[u_i]$
2. The population error $u_i \sim \mathbb{N}(0, \sigma^2)$
   - $\mathbb{E}[u_i] = 0$ and $\mathbb{V}[u_i] = \sigma^2$
   - Could also write as $\boldsymbol{u} \sim \mathbb{N}(0, \sigma^2 I_n)$

- This assumption implies MLR Assumptions 4 and 5.

# Conditional Normality of Outcome

**Theorem 1: Conditional Normality of Outcome**

Under MLR Assumptions 1-6, for each $i = 1, \ldots, n$,

$$y_i \mid \boldsymbol{x}_i \sim \mathbb{N}(\boldsymbol{x}_i' \boldsymbol{\beta}, \sigma^2).$$
$$\boldsymbol{y} \mid X \sim \mathbb{N}(X\boldsymbol{\beta}, \sigma^2 I_n).$$

- $y_i$ is a linear function of a normal random variable, $u_i$, so $y_i$ is also normal.
- $\mathbb{E}[y_i \mid \boldsymbol{x}_i] = \mathbb{E}[\boldsymbol{x}_i' \boldsymbol{\beta} + u_i \mid \boldsymbol{x}_i] = \boldsymbol{x}_i' \boldsymbol{\beta}$.
- $\mathbb{V}[y_i \mid \boldsymbol{x}_i] = \mathbb{V}[\boldsymbol{x}_i' \boldsymbol{\beta} + u_i \mid \boldsymbol{x}_i] = \sigma^2$ because we treat $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ as known and constant, respectively.

# Normality of Errors

**Question 1**

Is the normality assumption of the error term reasonable?

# Normality of Errors

**Question 1**

Is the normality assumption of the error term reasonable?

**Answer to Question 1**

Probably not. Some cases where it likely fails is when

1. $y_i > 0$ for all observations.
2. $y_i$ takes on discrete values (e.g., binary outcome) for all observations.

# Conditional Normality of OLS Estimator

**Theorem 2: Conditional Normality of OLS Estimator**

Under MLR Assumptions 1-6, for each parameter $j = 1, \ldots, k$

$$\widehat{\beta}_j \mid X \sim \mathbb{N}\left(\beta_j, \mathbb{V}\left[\widehat{\beta}_j\right]\right) = \mathbb{N}\left(\beta_j, \left[\sigma^2(X'X)^{-1}\right]_{jj}\right)$$
$$\widehat{\boldsymbol{\beta}} \mid X \sim \mathbb{N}\left(\boldsymbol{\beta}, \sigma^2(X'X)^{-1}\right).$$

# Conditional Normality of OLS Estimator

**Proof 1: Conditional Normality of OLS Estimator Part 1**

Recall we can write $\widehat{\boldsymbol{\beta}}$ as

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{u}.$$

# Conditional Normality of OLS Estimator

**Proof 1: Conditional Normality of OLS Estimator Part 1**

Recall we can write $\widehat{\boldsymbol{\beta}}$ as

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{u}.$$

Under MLR Assumption 6, $\boldsymbol{u} \sim \mathbb{N}(0, \sigma^2 I_n)$. Since $\widehat{\boldsymbol{\beta}}$ is a linear combination of $\boldsymbol{u}$, the former is also normal.

# Conditional Normality of OLS Estimator

**Proof 1: Conditional Normality of OLS Estimator Part 1**

Recall we can write $\widehat{\boldsymbol{\beta}}$ as

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{u}.$$

Under MLR Assumption 6, $\boldsymbol{u} \sim \mathbb{N}(0, \sigma^2 I_n)$. Since $\widehat{\boldsymbol{\beta}}$ is a linear combination of $\boldsymbol{u}$, the former is also normal.

MLR Assumptions 1-4 then imply OLS is unbiased so $\mathbb{E}\left[\widehat{\boldsymbol{\beta}} \mid X\right] = \boldsymbol{\beta}$ and MLR Assumptions 1-5 imply $\mathbb{V}\left[\widehat{\boldsymbol{\beta}} \mid X\right] = \sigma^2(X'X)^{-1}$. □

Hooray!

# Conditional Standard Normality of OLS Estimator

**Property 1: Conditional Standard Normality of OLS Estimator**

Under MLR Assumptions 1-6, since

$$\widehat{\beta}_j \mid X \sim \mathbb{N}\left(\beta_j, \left[\sigma^2 (X'X)^{-1}\right]_{jj}\right),$$

it follows that

$$\left.\frac{\widehat{\beta}_j - \beta_j}{\sigma\sqrt{\left[(X'X)^{-1}\right]_{jj}}}\right| X = \left.\frac{\widehat{\beta}_j - \beta_j}{\text{sd}\left[\widehat{\beta}_j\right]}\right| X \sim \mathbb{N}(0,1).$$

- Problem: We don't know $\sigma$ and, consequently, don't know the standard deviation of $\widehat{\beta}_j$.

# Student's t-Distribution

**Theorem 3: Student's t-Distribution**

The T-statistic

$$T = \frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}\sqrt{[(X'X)^{-1}]_{jj}}} = \frac{\widehat{\beta}_j - \beta_j}{\text{se}\left[\widehat{\beta}_j\right]} \sim t_{n-k}$$

follows the Student's t-Distribution with $n-k$ degrees of freedom where $k$ is the number of estimated parameters (intercept included).

- Replace $\sigma$ with $\widehat{\sigma}$ and we go from standard normal distribution to Student's t-Distribution.
- Symmetric around zero and converges to standard normal as degrees of freedom tends to infinity.

# Student's t-Distribution

**Question 2**

What is the point of all this? Why do we care about the $T$-statistic? We don't know $\beta_j$ so we still can't compute it. Aren't we in the same problem as with the standard normal distribution?

# Student's t-Distribution

**Question 2**

What is the point of all this? Why do we care about the $T$-statistic? We don't know $\beta_j$ so we still can't compute it. Aren't we in the same problem as with the standard normal distribution?

**Answer to Question 2**

Although we don't know $\beta_j$, we may interested in the *hypothesis* that it is some value $a$. Then, we can plug in $a$ for $\beta_j$ in the $T$-statistic and carry out this hypothesis test and compute confidence intervals!

# Two-Tailed T-Test

## Definition 2: Two-Tailed T-Test

A two-tailed T-Test is a way of testing the null hypothesis $H_0$ that our true parameter $\beta_j$ is *equal* some value $a$. To conduct a two-tailed T-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_j = a$ versus $H_A : \beta_j \neq a$.

# Two-Tailed T-Test

**Definition 2: Two-Tailed T-Test**

A two-tailed T-Test is a way of testing the null hypothesis $H_0$ that our true parameter $\beta_j$ is *equal* some value $a$. To conduct a two-tailed T-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_j = a$ versus $H_A : \beta_j \neq a$.
2. Select a significance level $\alpha$.

# Two-Tailed T-Test

## Definition 2: Two-Tailed T-Test

A two-tailed T-Test is a way of testing the null hypothesis $H_0$ that our true parameter $\beta_j$ is *equal* some value $a$. To conduct a two-tailed T-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_j = a$ versus $H_A : \beta_j \neq a$.
2. Select a significance level $\alpha$.
3. Construct the T-statistic $T = \frac{\widehat{\beta}_j - a}{\text{se}\left[\widehat{\beta}_j\right]}$.

# Two-Tailed T-Test

## Definition 2: Two-Tailed T-Test

A two-tailed T-Test is a way of testing the null hypothesis $H_0$ that our true parameter $\beta_j$ is *equal* some value $a$. To conduct a two-tailed T-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_j = a$ versus $H_A : \beta_j \neq a$.
2. Select a significance level $\alpha$.
3. Construct the T-statistic $T = \dfrac{\widehat{\beta}_j - a}{\text{se}\left[\widehat{\beta}_j\right]}$.
4. Find the t-critical value $t_{1-\frac{\alpha}{2},n-k}$ that corresponds to $n-k$ degrees of freedom at the chosen significance level $\alpha$.

# Two-Tailed T-Test

**Definition 2: Two-Tailed T-Test**

A two-tailed T-Test is a way of testing the null hypothesis $H_0$ that our true parameter $\beta_j$ is *equal* some value $a$. To conduct a two-tailed T-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_j = a$ versus $H_A : \beta_j \neq a$.
2. Select a significance level $\alpha$.
3. Construct the T-statistic $T = \frac{\widehat{\beta}_j - a}{\text{se}\left[\widehat{\beta}_j\right]}$.
4. Find the t-critical value $t_{1-\frac{\alpha}{2}, n-k}$ that corresponds to $n-k$ degrees of freedom at the chosen significance level $\alpha$.
5. Reject the null hypothesis if and only if $|T| > t_{1-\frac{\alpha}{2}, n-k}$.

# Right-Tailed T-Test

## Definition 3: Right-Tailed T-Test

A right-tailed T-Test is a way of testing the null hypothesis $H_0$ that our true parameter $\beta_j$ is *less than or equal* to some value $a$. To conduct a right-tailed T-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_j \leq a$ versus $H_A : \beta_j > a$.

# Right-Tailed T-Test

## Definition 3: Right-Tailed T-Test

A right-tailed T-Test is a way of testing the null hypothesis $H_0$ that our true parameter $\beta_j$ is *less than or equal* to some value $a$. To conduct a right-tailed T-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_j \leq a$ versus $H_A : \beta_j > a$.
2. Select a significance level $\alpha$.

# Right-Tailed T-Test

## Definition 3: Right-Tailed T-Test

A right-tailed T-Test is a way of testing the null hypothesis $H_0$ that our true parameter $\beta_j$ is *less than or equal* to some value $a$. To conduct a right-tailed T-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_j \leq a$ versus $H_A : \beta_j > a$.
2. Select a significance level $\alpha$.
3. Construct the T-statistic $T = \dfrac{\widehat{\beta}_j - a}{\text{se}\left[\widehat{\beta}_j\right]}$.

# Right-Tailed T-Test

## Definition 3: Right-Tailed T-Test

A right-tailed T-Test is a way of testing the null hypothesis $H_0$ that our true parameter $\beta_j$ is *less than or equal* to some value $a$. To conduct a right-tailed T-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_j \leq a$ versus $H_A : \beta_j > a$.
2. Select a significance level $\alpha$.
3. Construct the T-statistic $T = \frac{\widehat{\beta}_j - a}{\text{se}\left[\widehat{\beta}_j\right]}$.
4. Find the t-critical value $t_{1-\alpha,n-k}$ that corresponds to $n-k$ degrees of freedom at the chosen significance level $\alpha$.

# Right-Tailed T-Test

## Definition 3: Right-Tailed T-Test

A right-tailed T-Test is a way of testing the null hypothesis $H_0$ that our true parameter $\beta_j$ is *less than or equal* to some value $a$. To conduct a right-tailed T-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_j \leq a$ versus $H_A : \beta_j > a$.
2. Select a significance level $\alpha$.
3. Construct the T-statistic $T = \dfrac{\widehat{\beta}_j - a}{\text{se}\left[\widehat{\beta}_j\right]}$.
4. Find the t-critical value $t_{1-\alpha,n-k}$ that corresponds to $n-k$ degrees of freedom at the chosen significance level $\alpha$.
5. Reject the null hypothesis if and only if $T > t_{1-\alpha,n-k}$.

# Left-Tailed T-Test

**Definition 4: Left-Tailed T-Test**

A left-tailed T-Test is a way of testing the null hypothesis $H_0$ that our true parameter $\beta_j$ is *greater than or equal* to some value $a$. To conduct a left-tailed T-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_j \geq a$ versus $H_A : \beta_j < a$.
2. Select a significance level $\alpha$.
3. Construct the T-statistic $T = \dfrac{\widehat{\beta}_j - a}{\text{se}\left[\widehat{\beta}_j\right]}$.
4. Find the t-critical value $t_{\alpha,n-k}$ that corresponds to $n-k$ degrees of freedom at the chosen significance level $\alpha$.
5. Reject the null hypothesis if and only if $T < t_{\alpha,n-k}$.

# Hypothesis Testing Terminology

**Definition 5: Hypothesis Testing Terminology**

- Confidence Level: How sure we are of something.

# Hypothesis Testing Terminology

**Definition 5: Hypothesis Testing Terminology**

- Confidence Level: How sure we are of something.
- Significance Level $\alpha$: $1 -$ Confidence Level.

# Hypothesis Testing Terminology

**Definition 5: Hypothesis Testing Terminology**

- Confidence Level: How sure we are of something.
- Significance Level $\alpha$: $1 -$ Confidence Level.
- Null Hypothesis $H_0$: What we assume to be true.

# Hypothesis Testing Terminology

**Definition 5: Hypothesis Testing Terminology**

- Confidence Level: How sure we are of something.
- Significance Level $\alpha$: $1 -$ Confidence Level.
- Null Hypothesis $H_0$: What we assume to be true.
- Alternative Hypothesis $H_A$: What we need evidence of.

# Hypothesis Testing Terminology

**Definition 5: Hypothesis Testing Terminology**

- Confidence Level: How sure we are of something.
- Significance Level $\alpha$: $1-$ Confidence Level.
- Null Hypothesis $H_0$: What we assume to be true.
- Alternative Hypothesis $H_A$: What we need evidence of.
- Type 1 Error (False Positive): Rejecting a null hypothesis that is true. $\alpha$ corresponds to the probability we are willing to accept this error.

# Hypothesis Testing Terminology

**Definition 5: Hypothesis Testing Terminology**

- Confidence Level: How sure we are of something.
- Significance Level $\alpha$: $1 -$ Confidence Level.
- Null Hypothesis $H_0$: What we assume to be true.
- Alternative Hypothesis $H_A$: What we need evidence of.
- Type 1 Error (False Positive): Rejecting a null hypothesis that is true. $\alpha$ corresponds to the probability we are willing to accept this error.
- Type 2 Error (False Negative): Failing to reject a null hypothesis that is false. Corresponds to the power of a test (probability of correctly rejecting a null hypothesis).

# t-Critical Value

## Definition 6: t-Critical Value

The t-critical value $t_{1-\frac{\alpha}{2},n-k}$ for a two-tailed test is the value such that the cumulative distribution function, $F_T$, of a t-distributed random variable, $T_{n-k}$, with $n-k$ degrees of freedom equals $1-\frac{\alpha}{2}$:

$$\begin{aligned} F_{T_{n-k}}\left(t_{1-\frac{\alpha}{2},n-k}\right) &= \mathbb{P}\left(T_{n-k} \leq t_{1-\frac{\alpha}{2},n-k}\right) \\ &= 1-\frac{\alpha}{2}. \end{aligned}$$

# t-Critical Value

**Definition 6: t-Critical Value**

The t-critical value $t_{1-\frac{\alpha}{2},n-k}$ for a two-tailed test is the value such that the cumulative distribution function, $F_T$, of a t-distributed random variable, $T_{n-k}$, with $n-k$ degrees of freedom equals $1-\frac{\alpha}{2}$:

$$F_{T_{n-k}}\left(t_{1-\frac{\alpha}{2},n-k}\right) = \mathbb{P}\left(T_{n-k} \leq t_{1-\frac{\alpha}{2},n-k}\right)$$
$$= 1 - \frac{\alpha}{2}.$$

- An analogous process applies for right-tailed and left-tailed tests:
  - For a right-tailed test, the t-critical value $t_{1-\alpha,n-k}$ satisfies $F_{T_{n-k}}\left(t_{1-\alpha,n-k}\right) = 1-\alpha$.
  - For a left-tailed test, the t-critical value $t_{\alpha,n-k}$ satisfies $F_{T_{n-k}}\left(t_{\alpha,n-k}\right) = \alpha$.
  - We use the inverse CDF of the T-distribution to get $t_{1-\frac{\alpha}{2},n-k}$.

# p-value

**Definition 7: p-Value**

The p-value $p$ corresponding to a test statistic is the smallest level of significance at which the null hypothesis $H_0$ can be rejected.

# p-value for T-Test

**Example 1: p-value for Two-Tailed T-Test**

The p-value $p$ corresponding to a two-tailed test T-test is given by

$$\begin{aligned} p &= 2 * \mathbb{P}(T_{n-k} > |T|) \\ &= 2 * (1 - F_{T_{n-k}}(|T|)). \end{aligned}$$

where $T_{n-k}$ is a $t$-distributed random variable with degrees of freedom $n-k$.

- We multiply by two because we are doing a two-tailed test and need the area of both the upper and lower tails.
  - The area in these lower and upper tails is identical because the t-distribution is symmetric around zero.

# Economic Versus Statistical Significance

**Definition 8: Economic Significance**

We say a parameter estimate $\widehat{\beta}_j$ is economically significant if it is sufficiently large.

# Economic Versus Statistical Significance

**Question 3**

Suppose we estimate the model

$$\ln(income_i) = \beta_0 + \beta_1 education_i + \beta_2 experience_i + u_i$$

where $income$ is a person's yearly income, $education$ is measured in years, and $experience$ is measured in years.

Suppose $\widehat{\beta}_1 = 0.01$ with a p-value of $0.002$ and $\widehat{\beta}_2 = 0.3$ with a p-value of $0.045$. Are these estimates statistically significant at the $\alpha = 0.05$ level. What about economically significant?

# Economic Versus Statistical Significance

**Answer to Question 3**

1. $\widehat{\beta}_1$ is statistically significant at the $\alpha = 0.05$ level because $0.002 < 0.05$. However, it is not economically significant because our model predicts that the effect of an additional year of education only amounts to a person's yearly income increasing by $1\%$.
2. $\widehat{\beta}_2$ is statistically significant at the $\alpha = 0.05$ level because $0.045 < 0.05$. It is economically significant because our model predicts an additional year of education amounts to a person's yearly income increasing by $30\%$ which is quite a big amount.

# Confidence Intervals

**Theorem 4: Confidence Interval Estimator**

The $100(1-\alpha)\%$ confidence interval for a parameter $\beta_j$ is given by

$$\left[\widehat{\beta}_j - t_{1-\frac{\alpha}{2},n-k} \cdot \text{se}\left(\widehat{\beta}_j\right),\ \widehat{\beta}_j + t_{1-\frac{\alpha}{2},n-k} \cdot \text{se}\left(\widehat{\beta}_j\right)\right].$$

- Same as saying $\widehat{\beta}_j \pm t_{1-\frac{\alpha}{2},n-k} * \text{se}\left[\widehat{\beta}_j\right]$.
- Gives us a region of values we can say the true $\beta_j$ lies in with $100(1-\alpha)\%$ confidence.
- Confidence interval widens as $\alpha$ increases.

# Confidence Intervals

## Proof 2: Confidence Interval Estimator Part 1

Let $t_{1-\frac{\alpha}{2},n-k}$ be the critical value for the t-distribution with $n-k$ degrees of freedom. Then, the probability that a t-distributed random variable $T_{n-k}$ with $n-k$ degrees of freedom is greater than $t_{1-\frac{\alpha}{2},n-k}$ is $1-F_{T_{n-k}}(t_{1-\frac{\alpha}{2},n-k}) = 1-\left(1-\frac{\alpha}{2}\right) = \frac{\alpha}{2}$. By the symmetry of the t-distribution, we have $F_{T_{n-k}}(-t_{1-\frac{\alpha}{2},n-k}) = F_{T_{n-k}}(t_{\frac{\alpha}{2},n-k}) = \frac{\alpha}{2}$. Consequently,

$$\begin{aligned}
&\mathbb{P}(-t_{1-\frac{\alpha}{2},n-k} \leq T_{n-k} \leq t_{1-\frac{\alpha}{2},n-k}) \\
&= F_{T_{n-k}}\left(t_{1-\frac{\alpha}{2},n-k}\right) - F_{T_{n-k}}\left(-t_{1-\frac{\alpha}{2},n-k}\right) \\
&= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} \\
&= 1 - \alpha.
\end{aligned}$$

# Confidence Intervals

## Proof 2: Confidence Interval Estimator Part 2

We know that $\frac{\widehat{\beta}_j - \beta_j}{\text{se}\left[\widehat{\beta}_j\right]}$ is a t-distributed random variable with $n-k$ degrees of freedom. So,

$$\mathbb{P}\left(-t_{1-\frac{\alpha}{2},n-k} \leq \frac{\widehat{\beta}_j - \beta_j}{\text{se}\left[\widehat{\beta}_j\right]} \leq t_{1-\frac{\alpha}{2},n-k}\right)$$
$$= \mathbb{P}\left(-\widehat{\beta}_j - t_{1-\frac{\alpha}{2},n-k} \cdot \text{se}\left[\widehat{\beta}_j\right] \leq -\beta_j \leq -\widehat{\beta}_j + t_{1-\frac{\alpha}{2},n-k} \cdot \text{se}\left[\widehat{\beta}_j\right]\right)$$
$$= \mathbb{P}\left(\widehat{\beta}_j - t_{1-\frac{\alpha}{2},n-k} \cdot \text{se}\left[\widehat{\beta}_j\right] \leq \beta_j \leq \widehat{\beta}_j + t_{1-\frac{\alpha}{2},n-k} \cdot \text{se}\left[\widehat{\beta}_j\right]\right)$$
$$= 1 - \alpha.$$

# Confidence Intervals

**Proof 2: Confidence Interval Estimator Part 3**

Thus, the probability that $\beta_j$ lies in the interval $\widehat{\beta}_j \pm t_{1-\frac{\alpha}{2},n-k} \cdot \text{se}\left[\widehat{\beta}_j\right]$ is $1-\alpha$ meaning we can say with $100(1-\alpha)\%$ confidence that $\beta_j$ is some number in this interval. $\square$
Hooray!

# Two-Sided Tests with Confidence Intervals

**Property 2: Two-Sided Tests with Confidence Intervals**

Suppose we wish to test $H_0 : \beta_j = a$ versus $H_A : \beta_j \neq a$ at the $\alpha$ significance level. If $a$ is not contained in the confidence interval

$$\left[\widehat{\beta}_j - t_{1-\frac{\alpha}{2},n-k} \cdot \text{se}\left(\widehat{\beta}_j\right),\ \widehat{\beta}_j + t_{1-\frac{\alpha}{2},n-k} \cdot \text{se}\left(\widehat{\beta}_j\right)\right].$$

then we reject the null hypothesis.

# F-Tests

## Question 4

Suppose we are interested in the model

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + u_i.$$

How do we test $H_0 : \beta_{k-q} = 0 = \ldots = \beta_k = 0$ versus $H_A : \beta_{k-q} \neq 0$ or $\ldots$ or $\beta_k \neq 0$?

## Answer to Question 4

Construct an F-Test! How do we do that?

# Unrestricted Model

**Definition 9: Unrestricted Model**

The unrestricted model is simply the full model:

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + u_i.$$

# Restricted Model

**Definition 9: Restricted Model**

The restricted model assumes the last $q$ parameters are zero so we get

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{k-q} x_{i,k-q} + u_i.$$

- The restricted model is "nested" within the unrestricted model because it has a subset of the covariates included in the unrestricted model.

# Restricted and Unrestricted Model

**Example 2**

Suppose we have the unrestricted model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$$

and we wish to test $H_0 : \beta_1 = \beta_2 = 0$. Then, the restricted model is

$$y_i = \beta_0 + \beta_3 x_{i3} + v_i.$$

# F-Statistic

**Theorem 5: F-Statistic**

The F-statistic

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k)} \sim F_{q,n-k}$$

follows an F-distribution with $q$ numerator degrees of freedom and $n-k$ denominator degrees of freedom.

- $SSR_r$ is the $SSR$ from the restricted model.
- $SSR_{ur}$ is the $SSR$ from the unrestricted model.
- $q$ is the number of restrictions.
- $n-k$ is the number of estimated parameters (intercept included).
- Proof follows from the fact the F-statistic can be written as the ratio of two $\chi^2$ random. variables which necessarily follows an F-Distribution.

# F-Test

## Definition 10: F-Test

To conduct an F-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_{k-q} = \ldots = \beta_k = 0$ versus $H_A : \beta_{k-q} \neq 0$ or $\ldots$ or $\beta_k \neq 0$.

# F-Test

## Definition 10: F-Test

To conduct an F-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_{k-q} = \ldots = \beta_k = 0$ versus $H_A : \beta_{k-q} \neq 0$ or $\ldots$ or $\beta_k \neq 0$.
2. Select a significance level $\alpha$.

# F-Test

## Definition 10: F-Test

To conduct an F-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_{k-q} = \ldots = \beta_k = 0$ versus $H_A : \beta_{k-q} \neq 0$ or $\ldots$ or $\beta_k \neq 0$.
2. Select a significance level $\alpha$.
3. Estimate the unrestricted model to obtain the $SSR_{ur}$.

# F-Test

## Definition 10: F-Test

To conduct an F-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_{k-q} = \ldots = \beta_k = 0$ versus $H_A : \beta_{k-q} \neq 0$ or $\ldots$ or $\beta_k \neq 0$.
2. Select a significance level $\alpha$.
3. Estimate the unrestricted model to obtain the $SSR_{ur}$.
4. Estimate the restricted model to obtain the $SSR_r$.

# F-Test

## Definition 10: F-Test

To conduct an F-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_{k-q} = \ldots = \beta_k = 0$ versus $H_A : \beta_{k-q} \neq 0$ or $\ldots$ or $\beta_k \neq 0$.
2. Select a significance level $\alpha$.
3. Estimate the unrestricted model to obtain the $SSR_{ur}$.
4. Estimate the restricted model to obtain the $SSR_r$.
5. Construct the F-statistic $F = \dfrac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k)}$.

# F-Test

## Definition 10: F-Test

To conduct an F-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_{k-q} = \ldots = \beta_k = 0$ versus $H_A : \beta_{k-q} \neq 0$ or $\ldots$ or $\beta_k \neq 0$.
2. Select a significance level $\alpha$.
3. Estimate the unrestricted model to obtain the $SSR_{ur}$.
4. Estimate the restricted model to obtain the $SSR_r$.
5. Construct the F-statistic $F = \dfrac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k)}$.
6. Find the F-critical value $f_{1-\alpha,q,n-k}$ that corresponds to $q$ numerator degrees of freedom and $n-k$ denominator degrees of freedom at the chosen significance level $\alpha$.

# F-Test

## Definition 10: F-Test

To conduct an F-Test, we:

1. Formulate a hypothesis test of $H_0 : \beta_{k-q} = \ldots = \beta_k = 0$ versus $H_A : \beta_{k-q} \neq 0$ or $\ldots$ or $\beta_k \neq 0$.
2. Select a significance level $\alpha$.
3. Estimate the unrestricted model to obtain the $SSR_{ur}$.
4. Estimate the restricted model to obtain the $SSR_r$.
5. Construct the F-statistic $F = \dfrac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k)}$.
6. Find the F-critical value $f_{1-\alpha,q,n-k}$ that corresponds to $q$ numerator degrees of freedom and $n-k$ denominator degrees of freedom at the chosen significance level $\alpha$.
7. Reject the null hypothesis if and only if $F > f_{1-\alpha,q,n-k}$.

# F-Test

**Definition 11: F-Test Terminology**

When we reject the null $H_0 : \beta_{k-q} = 0, \ \ldots, \beta_k = 0$ we say these parameters are jointly significant. When we fail to reject this null, we say these parameters are jointly insignificant.

- When we reject this null hypothesis, we still do not know which of these variables effects the outcome significantly.
  - Perhaps they all do, but maybe only one does. So we must carry out individual T-tests.

# F-Statistic for Intercept Only Model

**Definition 11: F-Statistic for Intercept Only Model**

The F-statistic for testing the null hypothesis $H_0 : \beta_1 = 0, \ldots, \beta_k = 0$ is given by

$$F = \frac{R^2/(k-1)}{\left(1-R^2\right)/(n-k)}.$$

- This is reported in R's summary() function.
- No need to estimate unrestricted and restricted models to carry out this test.

# f-Critical Value

**Definition 12: f-Critical Value**

The f-critical value $f_{1-\alpha,q,n-k}$ for an F-test is the value such that the cumulative distribution function, $F_{f_{q,n-k}}$, of a F-distributed random variable, $f_{q,n-k}$, with $q$ and $n-k$ degrees of freedom equals $1-\alpha$:

$$\begin{aligned} F_{f_{q,n-k}}\left(f_{1-\alpha,q,n-k}\right) &= \mathbb{P}\left(f_{q,n-k} \leq f_{1-\alpha,q,n-k}\right) \\ &= 1-\alpha. \end{aligned}$$

- We use the inverse CDF of the F-distribution to get $f_{1-\alpha,q,n-k}$.

# p-value for F-Test

**Example 3: p-value for F-Test**

The p-value $p$ corresponding to an F-test is given by

$$\begin{aligned} p &= \mathbb{P}(f_{q,n-k} > F) \\ &= 1 - F_{f_{q,n-k}}(F). \end{aligned}$$

where $f_{q,n-k}$ is a F-distributed random variable with degrees of freedom $q$ and $n-k$.

- No need to multiply by two because the F-distribution only has a right tail.

# Thank You!