

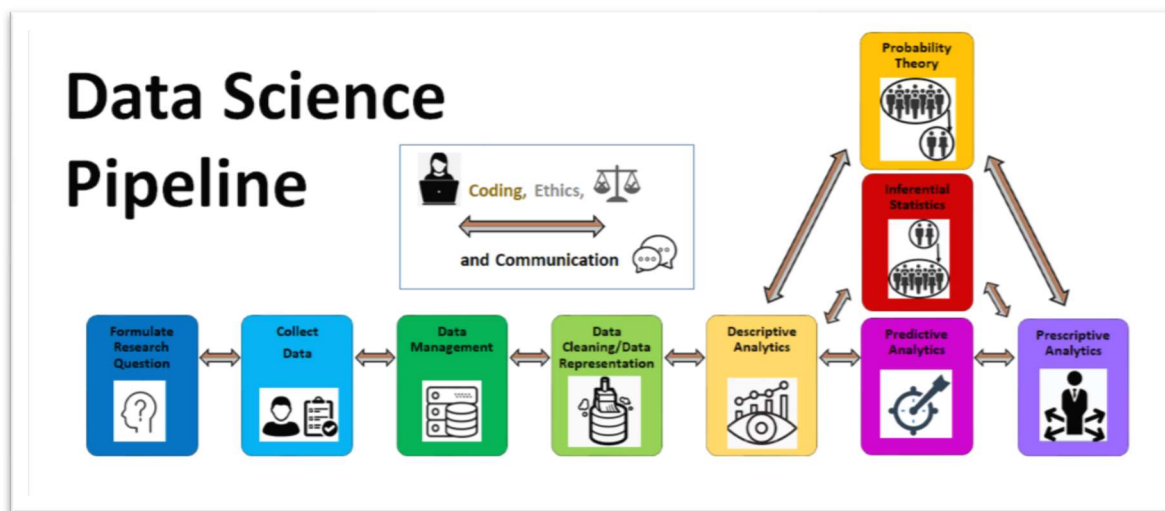


## Unit 12: Multiple Linear Regression Modeling

### Case Studies:

- Can we model the **weight** of healthy adult with a linear relationship of **height**, **sex**, and **age\_group**?
- Can we model the **mpg** of an old car using the **weight**?

### Purpose of this Lectures



1. Analyses for Associations
2. Association Analyses Summary: Numerical and/or Categorical Explanatory Variables-> Numerical Response Variable
3. Basic Descriptive Analytics for the Sample Data  
(Numerical Response Variable, Multiple Explanatory Variables – Numerical and Categorical)
  - 3.1. Visualizations
  - 3.2. Summary Statistics
    - 3.2.1. Python: .groupby() function
4. Multiple Linear Regression – Sample Data
5. Categorical Explanatory Variables
6. Interpreting Intercepts and Slopes of Regression Equations
7. Inference for Multiple Linear Regression Intercept and Slopes
  - 7.1. Conditions for Inference
  - 7.2. Inference of a Single Multiple Linear Regression Population Slope
    - 7.2.1. Confidence Intervals
    - 7.2.2. Hypothesis Test  $H_0: \beta_i = 0$
  - 7.3. Inference for ALL Multiple Linear Regression Population Slopes
    - 7.3.1. F-Distribution

7.3.2. Conducting the Test  $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$

7.3.3. Why would we want to conduct the test  $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$ .

7.4. Inference for a Subset of Multiple Linear Regression Population Slopes

7.4.1. Conducting the Test  $H_0: \beta_{p+1} = \beta_{p+2} = \dots = \beta_{p+q} = 0$ .

**8. Linear Regression Models with Interaction Variables**

**9. Making Predictions with Multiple Linear Regression Models**

**10. Linear Transformations: What to try when your multiple linear regression conditions aren't met.**

## Additional Resources

Chapter 8.1 and 8.3 in Diez, Barr, and Cetinkaya-Rundel, (2015), *OpenIntro*

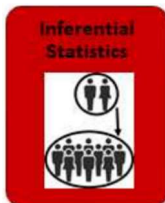
*Statistics* <https://www.openintro.org/download.php?file=os3&redirect=/stat/textbook/os3.php>

# 1. ANALYSES FOR ASSOCIATIONS

## Questions to consider, when selecting an analysis to test an association.



1. Which variable is the **response variable** in this association?
  - a. Is it a **categorical** or **numerical** variable?
  - b. If it's a categorical variable, **how many levels** does it have?
2. Which variable(s) is the **explanatory variable** in this association?
  - a. Is it a **categorical** or **numerical** variable?
  - b. If it's a categorical variable, **how many levels** does it have?
3. How would you **quantify this association**?
  - a. Difference between two summary statistics? What two summary statistics?
  - b. With a model? What kind of model?



4. Are you interested in an association in a **sample** or a **population**?
5. When is it **appropriate to use this test** for association?

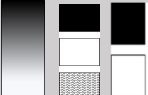




6. Can you use this model/test to **make predictions**?
  - a. How would you quantify the performance of your predictions?

## 2. ASSOCIATION ANALYSIS SUMMARY:

RESPONSE: NUMERICAL

EXPLANATORY(S): NUMERICAL AND/OR CATEGORICAL

Research Questions about Associations	Type of Variables Involved in the Association Test	 <b>Explanatory Variables:</b> Numerical and/or Categorical Variables   <b>Response Variable:</b> Numerical Variable
	Example	Is there an association between <b>weight</b> and the <b>height, age, and sex</b> of healthy adults? 
	Type of Association (Way to Quantify Association)	<b>Multiple Linear Regression Model</b> <i>(linear relationship between explanatory variable (x) and response variable (y))</i>
Descriptive Analytics	How to <u>Describe</u> an Association in a <u>Sample</u> ?	1. Multiple Linear Regression Model: <ul style="list-style-type: none"> <li><math>\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p</math></li> <li><math>R^2</math> of the model</li> </ul>
	When is this analysis ( <i>for the sample</i> ) appropriate to use?	Linearity condition is met
Inferential Statistics	How to <u>Infer</u> an Association for a <u>Population</u> ?	Conduct inference on: <ul style="list-style-type: none"> <li>A <u>single</u> population parameter <math>\beta_i</math></li> <li><u>All</u> population parameters <math>\beta_1, \dots, \beta_p</math></li> <li>A <u>subset</u> of population parameters <math>\beta_{p+1}, \dots, \beta_{p+q}</math></li> </ul>
	When is this analysis ( <i>for the population</i> ) appropriate to use?	1. Linearity condition is met 2. Constant variance of residuals condition is met. 3. Residuals are normal (and centered at 0). 4. Residuals are independent. 5. No-Multicollinearity condition is met.
Predictive Analytics	Making Predictions	Use your multiple linear regression line to make predictions $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$
	How to quantify the performance of your prediction(s)?	<ul style="list-style-type: none"> <li><u>Individual Data Point</u>: residual</li> <li><u>All Data</u>: root mean square error (RMSE)</li> </ul>

### 3. BASIC DESCRIPTIVE ANALYTICS FOR SAMPLE DATA – NUMERICAL RESPONSE VARIABLE – MULTIPLE EXPLANATORY VARIABLES (NUMERICAL AND CATEGORICAL)

See the Unit 13 notebook for examples of visualizations and summary statistics that involve a numerical response variable and two or more explanatory variables.

### 4. MULTIPLE LINEAR REGRESSION – SAMPLE DATA

A **multiple linear regression line** is generally a best fit line that has requires  $p > 1$  slopes.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

To find the optimal values of  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$  can similarly fit an **Ordinary Least Squares Regression** line to this data.

**Goal for Finding the Ordinary Least Squares Regression Line:** Find optimal values for  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$  for the equation value  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$  that **minimizes**

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{y})^2 &= \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_p x_{i,p}))^2 \\ &= (y_1 - (\hat{\beta}_0 + \hat{\beta}_1 x_{1,1} + \cdots + \hat{\beta}_p x_{1,p}))^2 + \cdots + (y_n - (\hat{\beta}_0 + \hat{\beta}_1 x_{n,1} + \cdots + \hat{\beta}_p x_{n,p}))^2 \end{aligned}$$

**Calculation of optimal values for  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ :**

- Based on the same idea for simple linear regression.
- Won't ask to calculate by hand in this class. The Python output tables will give you these optimal values of  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ .

In general, linear regression can incorporate \_\_\_\_\_ and \_\_\_\_\_ explanatory variables, but the response variable must be \_\_\_\_\_.

## 5. CATEGORICAL EXPLANATORY VARIABLES

How would we convert our sample dataset, with categorical explanatory variables into a linear regression equation?

Response Variable	Explanatory Variables		
weight	sex	age_group	height
73	Male	under_30	176.5
65.2	Female	under_30	168.5
84.5	Female	40 and above	162.6
58.4	Female	30-39	173.2
70.2	Male	under_30	175
...	...	...	...

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$$

$\hat{y}$	x1	x2	x3	x4	y
	1	0	1	176.5	73
	0	0	1	168.5	65.2
	0	1	0	162.6	84.5
	0	0	0	173.2	58.4
	1	0	1	175	70.2
	...	...	...	...	...
Predicted Weight	sex[T.Male]	age_Group[T.40 and above]	age_Group[T.under_30]	height	Observed Weight

### Definitions:

Categorical explanatory variables can be represented as 0/1 **indicator variables**.

The level of the categorical explanatory variable that is represented as a “1” in a given indicator variable is the **indicator of that variable**.

**Example:** Indicator variables can be translated as asking a (yes=1)/(no=0) question about the input observation.

- The variable ( $x_1$  also named **sex[T.male]**) can be translated as a yes/no question for an observation.
  - Is person male?
    - Yes = \_\_\_\_\_
    - No = \_\_\_\_\_
- The variable ( $x_2$  also named **age\_Group[T.40 and above]**) can be translated as a yes/no question for an observation.
  - Is person at least 40 years old?
    - Yes = \_\_\_\_\_
    - No = \_\_\_\_\_
- The variable ( $x_3$  also named **age\_Group[T.under\_30]**) can be translated as a yes/no question for an observation.
  - Is person under 30?
    - Yes = \_\_\_\_\_
    - No = \_\_\_\_\_

The level of an explanatory variable that has no corresponding indicator variable is called the **reference level**.

### Rules:

A categorical explanatory variable with  $w$  levels must be represented by exactly  $w-1$  levels. Otherwise when you try to calculate the optimal values of  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$  in your multiple linear regression equation, you will get “multiple solution errors.”

## 6. INTERPRETING INTERCEPTS AND SLOPES OF LINEAR REGRESSION EQUATIONS

Go to Unit 13 notebook to formulate the multiple linear regression line which predict weight, given height, sex, and age group.

### OLS Regression Results

Dep. Variable:	weight	R-squared:	0.594
Model:	OLS	Adj. R-squared:	0.590
Method:	Least Squares	F-statistic:	176.1
Date:	Sun, 28 Mar 2021	Prob (F-statistic):	7.71e-93
Time:	14:41:21	Log-Likelihood:	-1737.1
No. Observations:	487	AIC:	3484.
Df Residuals:	482	BIC:	3505.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	-59.0102	9.409	-6.272	0.000	-77.498	-40.523
sex[T.Male]	7.9128	1.086	7.286	0.000	5.779	10.047
age_group[T.40 and above]	3.4030	1.202	2.830	0.005	1.041	5.765
age_group[T.under_30]	-1.7432	0.949	-1.838	0.067	-3.607	0.121
height	0.7291	0.057	12.821	0.000	0.617	0.841

Omnibus:	91.450	Durbin-Watson:	1.991
Prob(Omnibus):	0.000	Jarque-Bera (JB):	182.551
Skew:	1.032	Prob(JB):	2.29e-40
Kurtosis:	5.176	Cond. No.	4.14e+03

### Notes:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 4.14e+03. This might indicate that there are strong multicollinearity or other numerical problems.



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

1.  $\hat{\beta}_0$ Intercepts: “If the explanatory variables were all 0, we would expect the response variable value, on average, to be  $\hat{\beta}_0$ .”
  - a. Note: Sometimes the interpretation of  $\hat{\beta}_0$  can be nonsensical to the application.
  
2.  $\hat{\beta}_i$ Slope of a Numerical Explanatory Variable: All else held equal, if we were to increase  $x_i$  by 1, then, on average, we would expect the response variable value to increase by  $\hat{\beta}_i$ .
  - a. Note: This language is important, as we do not want to phrase this interpretation to imply that  $x_i$  \_\_\_\_\_ a change in the response variable.
  
3.  $\hat{\beta}_i$ , Slope of a Categorical Indicator Variable: All else held equal, we would expect the difference in average response variable value for those observations in the indicator level of  $x_i$  and those observations in the reference level (for the corresponding explanatory variable) is  $\hat{\beta}_i$ .
  - a. Note: This language is important, as we do not want to phrase this interpretation to imply that  $x_i$  \_\_\_\_\_ a change in the response variable.

**Ex:** Interpret the intercept, height slope, and age\_group[t.under\_30] slope for the problem above.

## 7. INFERENCE FOR MULTIPLE LINEAR REGRESSION INTERCEPT AND SLOPES

Just like with simple linear regression, we can use:

- $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ , from the multiple linear regression equation  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1x_1 + \hat{\beta}_2x_2 + \dots + \hat{\beta}_px_p$  for the sample data, to **conduct inference on**
- $\beta_0, \beta_1, \dots, \beta_p$ , from the multiple linear regression equation  $\hat{y} = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p$  for the population data.

However, with multiple linear regression, we can conduct many types of hypothesis testing on the slopes.

### 1. Hypothesis Testing on One Population Slope

$$H_0: \beta_i = 0$$

$$H_A: \beta_i \neq 0$$

### 2. Hypothesis Testing on All Population Slopes

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_A: \text{at least one } \beta_i \neq 0 \text{ (for } i = 1, \dots, p)$$

### 3. Hypothesis Testing on a Subset of $q$ Population Slopes

$$H_0: \beta_{p+1} = \beta_{p+2} = \dots = \beta_{p+q} = 0$$

$$H_A: \text{at least one } \beta_i \neq 0 \text{ (for } i = p + 1, \dots, p + q)$$

## 7.1. CONDITIONS FOR INFERENCE ON MULTIPLE LINEAR REGRESSION

### INTERCEPT AND SLOPES

The conditions for multiple linear regression are the same as those for simple linear regression, except there is one more condition that must hold.

1. [SLR AND MLR]: Linearity condition
2. [SLR AND MLR]: Constant variability of residuals condition
3. [SLR AND MLR]: Normality of residuals (with mean of 0) condition
4. [SLR AND MLR]: Independence of residuals condition
5. **[JUST MLR]: No-Multicollinearity Condition**

In order to avoid the model producing \_\_\_\_\_ slope estimates, one should avoid having explanatory variables that are **collinear**. Two explanatory variables are collinear if they have a \_\_\_\_\_ between them.

**Ex:** Suppose now we also wanted to add 'elbow diameter' to our list of explanatory variables.

1. Fit a multiple linear regression model predicting weight with:
  - a. Height
  - b. Elbow diameter
  - c. Sex
  - d. Age group.
2. What is the  $R^2$  of this model?
3. Check the multiple linear regression conditions for inference for this model.
4. If we were to delete one of these numerical explanatory variables (because of the multicollinearity condition being violated), which one would you choose.
5. Fit a multiple linear regression model predicting weight with the following explanatory variables and check the conditions.
  - a. Height
  - ~~b. Elbow diameter~~
  - c. Sex
  - d. Age group.

**Go to the notebook (section 7.1) to answer these questions.**

## 7.2. INFERENCE FOR A SINGLE MULTIPLE LINEAR REGRESSION SLOPE

We can create confidence intervals and conduct hypothesis testing on population slopes in a multiple linear regression using the **same procedure** that we used for the one population slope in a **simple** linear regression.

### 7.2.1. CONFIDENCE INTERVAL FOR A SINGLE MULTIPLE LINEAR REGRESSION SLOPE

1. Check the conditions for conducting inference on a population slope/intercept.
  - a. The linearity condition holds.
  - b. The constant residuals condition holds.
  - c. The residuals are normal.
  - d. The residuals are independent.
  - e. The explanatory variables (if a multiple linear regression is used) are not collinear.

2. The confidence interval for  $\beta_i$  is calculated by:

*(point estimate)  $\pm$  (critical value)(standard error)*

$$\hat{\beta}_i \pm t_{\{n-p-1\}}^* SE_{\beta_i}$$

Notation:

Go to notebook for an example.

## 7.2.2. CONDUCTING A HYPOTHESIS TEST FOR A SINGLE POPULATION SLOPE, TESTING THE CLAIM $H_A: \beta_i \neq 0$

### 1. Set up the hypotheses

$$H_0: \beta_i = 0$$

$$H_A: \beta_i \neq 0$$

### 2. Check the conditions for conducting inference on a population slope/intercept.

- a. The linearity condition holds.
- b. The constant residuals condition holds.
- c. The residuals are normal (with mean 0).
- d. The residuals are independent.
- e. The explanatory variables (if a multiple linear regression is used) are not collinear.

### 3. Calculate the point estimate (observed sample statistic)

$$\hat{\beta}_i$$

### 4. Calculate the p-value (or calculate a confidence interval)

$$p - \text{value} = 2P(T_{n-p-1} \geq \left| \frac{\hat{\beta}_i - 0}{SE_{\hat{\beta}_i}} \right|)$$

## 6. Make a Decision

### With a p-value

- a. If **p – value** <  **$\alpha$** , then we “reject the null hypothesis.” And we say that “there IS sufficient evidence to suggest the alternative hypothesis.”
- b. If **p – value**  $\geq$   **$\alpha$** , then we “fail to reject the null hypothesis.” And we say that “there IS NOT sufficient evidence to suggest the alternative hypothesis.”

### With a confidence interval

- a. If the **null value (0)** is **not in the confidence interval**, then we “reject the null hypothesis.” And we say that “there IS sufficient evidence to suggest the alternative hypothesis.”
- b. If the **null value (0)** is **in the confidence interval**, then we “fail to reject the null hypothesis.” And we say that “there IS NOT sufficient evidence to suggest the alternative hypothesis.”

## Go to notebook for an example.

We are interested in testing the claim that the height slope in the multiple linear regression population model (that predicts the weight of all healthy adults with height, age group, and sex) is non-zero.

### OLS Regression Results

Dep. Variable:	weight	R-squared:	0.594
Model:	OLS	Adj. R-squared:	0.590
Method:	Least Squares	F-statistic:	176.1
Date:	Mon, 29 Mar 2021	Prob (F-statistic):	7.71e-93
Time:	17:33:17	Log-Likelihood:	-1737.1
No. Observations:	487	AIC:	3484.
Df Residuals:	482	BIC:	3505.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	-59.0102	9.409	-6.272	0.000	-77.498	-40.523
sex[T.Male]	7.9128	1.086	7.286	0.000	5.779	10.047
age_group[T.40 and above]	3.4030	1.202	2.830	0.005	1.041	5.765
age_group[T.under_30]	-1.7432	0.949	-1.838	0.067	-3.607	0.121
height	0.7291	0.057	12.821	0.000	0.617	0.841
Omnibus:	91.450	Durbin-Watson:	1.991			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	182.551			
Skew:	1.032	Prob(JB):	2.29e-40			
Kurtosis:	5.176	Cond. No.	4.14e+03			

### Notes:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.  
 [2] The condition number is large, 4.14e+03. This might indicate that there are strong multicollinearity or other numerical problems.

$$H_0: \beta_4 = 0$$

$$H_A: \beta_4 \neq 0$$

$$\text{test stat} = \frac{\hat{\beta}_4 - 0}{SE_{\hat{\beta}_4}}$$

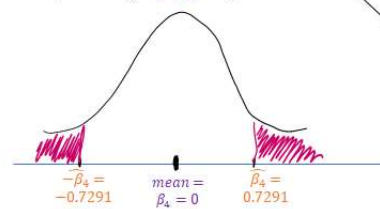
$$p\text{-value} = 2P(T_{n-p-1} \geq |\text{test stat}|)$$

Because the 5 multiple linear regression inference conditions hold.

### Sampling Distribution of

Sample Slopes  $\hat{\beta}_4$   
 (assuming  $H_0: \beta_4 = 0$ )

$$\hat{\beta}_4 \sim N(\text{mean} = \beta_4, \text{std} \approx SE_{\hat{\beta}_4})$$

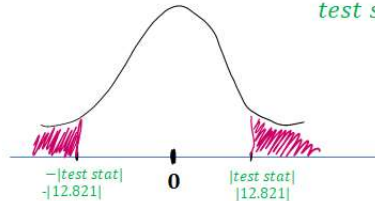


Because the sampling distribution is normal

### Sampling Distribution of Sample

Slope t-scores  $\text{test stat} = \frac{\hat{\beta}_4 - \beta_4}{SE_{\hat{\beta}_4}}$   
 (assuming  $H_0: \beta_4 = 0$ )

$$\text{test stat} = \frac{\hat{\beta}_4 - \beta_4}{SE_{\hat{\beta}_4}} \sim t_{n-p-1}$$



### 7.3. INFERENCE FOR ALL MULTIPLE LINEAR REGRESSION SLOPES

The hypothesis testing procedure for testing the following hypotheses uses a slightly different structure than what we've used up until now.

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_A: \text{at least one } \beta_i \neq 0 \text{ (for } i = 1, \dots, p)$$

#### What's different about this test?

1. First, this test does not have a confidence interval that corresponds to it. So we **must use a p-value**.
2. Next, when calculating our p-value in the past our **test statistic** represented either:
  - a. The z-score of the sample statistic (point estimate) OR
  - b. The t-score of the sample statistic (point estimate).

Now, we will use a test statistic to calculate the p-value, but it will not be a z-score or t-score.

3. Finally, when **calculating our p-value** in the past, our test statistic was considered to be either:
- a. an observation from the Z-distribution (ie. standard normal distribution) and our p-value was two tailed:  
ie.  $p - \text{value} = 2P(Z \geq |\text{test stat}|)$
  
  - b. an observation from the t-distribution and our p-value was two tailed:  
ie.  $p - \text{value} = 2P(T \geq |\text{test stat}|)$

Now, our test statistic will be an observation from a new distribution called the **F-distribution** and our p-value is right tailed.

ie.  $p - \text{value} = P(F \geq \text{test stat})$



### 7.3.1. F DISTRIBUTION

First, let's discuss this new distribution and some of its properties.

#### Random Variable that Follows the F-Distribution:

**Definition:** A continuous random variable is said to follow the **F-distribution** with  $d_1$  and  $d_2$  **degrees of freedom** if it has the following probability density function (pdf).

**Short-Hand:** \_\_\_\_\_

#### Probability Density Function:

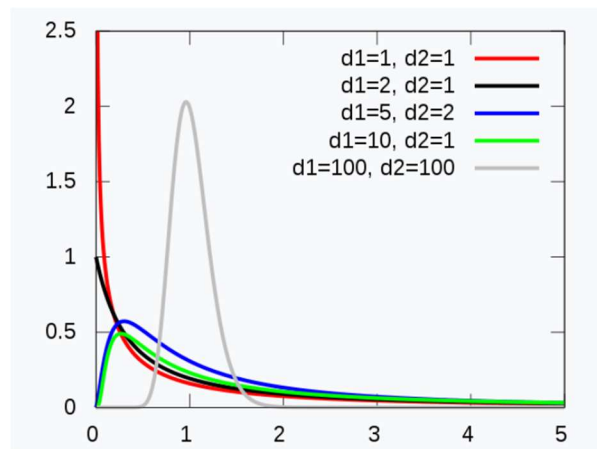
$$f(x) = \frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{xB(\frac{d_1}{2}, \frac{d_2}{2})}, \text{ for } x > 0$$

**Parameters that Dictates Shape:** \_\_\_\_\_

#### Properties:

- Always \_\_\_\_\_

**Shapes:** Can take on many different shapes, based on the parameter values.



**Ex:** Go to the Jupyter notebook to calculate the probability that an F-score is less than or equal to 4, (using df1=3 and df2=9).

$$P(F_{3,9} \leq 4) =$$

### 7.3.2. CONDUCTING THE TEST

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

1. Set up the hypotheses

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_A: \text{at least one } \beta_i \neq 0 \text{ (for } i = 1, \dots, p)$$

2. Check the conditions for conducting inference on a population slope/intercept.

- The linearity condition holds.
- The constant residuals condition holds.
- The residuals are normal (with mean 0).
- The residuals are independent.
- The explanatory variables (if a multiple linear regression is used) are not collinear.

3. Calculate the test statistic:

$$\text{test stat} = \frac{\frac{SSR}{p}}{\frac{SSE}{n - p - 1}}$$

4. Calculate the p-value

$$p - \text{value} = P(F_{p, n-p-1} \geq \text{test stat})$$

5. Make a Decision

- If **p – value** < **α**, then we “reject the null hypothesis.” And we say that “there IS sufficient evidence to suggest the alternative hypothesis.”
- If **p – value** ≥ **α**, then we “fail to reject the null hypothesis.” And we say that “there IS NOT sufficient evidence to suggest the alternative hypothesis.”

**Ex:** When using sex, height, and age\_group to predict weight in a linear regression equation, is there significant evidence to suggest that at least one of the slopes in the population linear regression model is non-zero? **Go to the Jupyter notebook.**

See notebook.

#### OLS Regression Results

Dep. Variable:	weight	R-squared:	0.594
Model:	OLS	Adj. R-squared:	0.590
Method:	Least Squares	F-statistic:	176.1
Date:	Sun, 28 Mar 2021	Prob (F-statistic):	7.71e-93
Time:	14:41:21	Log-Likelihood:	-1737.1
No. Observations:	487	AIC:	3484.
Df Residuals:	482	BIC:	3505.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	-59.0102	9.409	-6.272	0.000	-77.498	-40.523
sex[T.Male]	7.9128	1.088	7.288	0.000	5.779	10.047
age_group[T.40 and above]	3.4030	1.202	2.830	0.005	1.041	5.765
age_group[T.under_30]	-1.7432	0.949	-1.838	0.067	-3.607	0.121
height	0.7291	0.057	12.821	0.000	0.617	0.841

Omnibus:	91.450	Durbin-Watson:	1.991
Prob(Omnibus):	0.000	Jarque-Bera (JB):	182.551
Skew:	1.032	Prob(JB):	2.29e-40
Kurtosis:	5.176	Cond. No.	4.14e+03

#### Notes:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.  
 [2] The condition number is large, 4.14e+03. This might indicate that there are strong multicollinearity or other numerical problems.

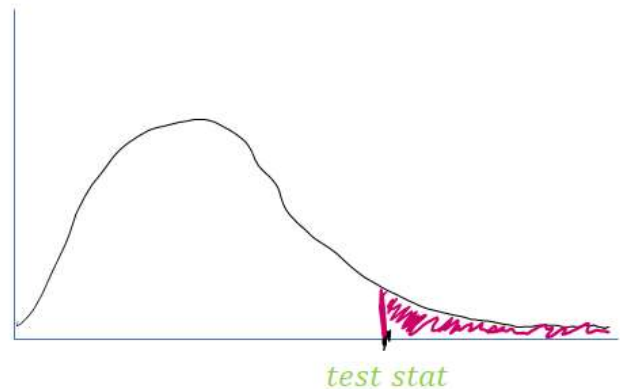
$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_A: \text{at least one } \beta_i \neq 0 \text{ (for } i = 1, \dots, p)$$

$$\text{test stat} = \frac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}}$$

$$p\text{-value} = P(F_{p, n-p-1} \geq \text{test stat})$$

**F distribution (with df1 = p, df2 = n-p-1)**



### 7.3.3. WHY WOULD WE WANT TO CONDUCT THIS TEST?

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

This test is vague!

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_A: \text{at least one } \beta_i \neq 0 \text{ (for } i = 1, \dots, p)$$

It does not tell us which slopes are non-zero.

**Q:** Why not conduct the following  $p$  tests instead to find out which slopes *specifically* are non-zero using a significance level of  $\alpha=0.05$ ?

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

$$H_0: \beta_2 = 0$$

$$H_A: \beta_2 \neq 0$$

...

$$H_0: \beta_p = 0$$

$$H_A: \beta_p \neq 0$$

**A:** The more tests we conduct, the more likely we are to have made a **Type 1 Error**.

To make a **Type 1 Error** in your hypothesis test decision is to have *incorrectly* rejected a null hypothesis that was *actually* true.

The probability of making a of Type 1 Error is highly related to the significance level. Actually they're the same!

$$P(\text{Type 1 Error}) = \alpha = \text{significance level}$$

**Ex:** If we were to use a significance level of  $\alpha = 0.05$  for each of the following individual population slope tests that corresponds to one of the slopes in the model below, what is the probability that at least one of these tests made a type 1 error?

#### OLS Regression Results

Dep. Variable:	weight	R-squared:	0.594
Model:	OLS	Adj. R-squared:	0.590
Method:	Least Squares	F-statistic:	176.1
Date:	Sun, 28 Mar 2021	Prob (F-statistic):	7.71e-93
Time:	14:41:21	Log-Likelihood:	-1737.1
No. Observations:	487	AIC:	3484.
Df Residuals:	482	BIC:	3505.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	-59.0102	9.409	-6.272	0.000	-77.498	-40.523
sex[T.Male]	7.9128	1.086	7.286	0.000	5.779	10.047
age_group[T.40 and above]	3.4030	1.202	2.830	0.005	1.041	5.765
age_group[T.under_30]	-1.7432	0.949	-1.838	0.067	-3.607	0.121
height	0.7291	0.057	12.821	0.000	0.617	0.841

Omnibus:	91.450	Durbin-Watson:	1.991
Prob(Omnibus):	0.000	Jarque-Bera (JB):	182.551
Skew:	1.032	Prob(JB):	2.29e-40
Kurtosis:	5.176	Cond. No.	4.14e+03

#### Notes:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 4.14e+03. This might indicate that there are strong multicollinearity or other numerical problems.

## 7.4. INFERENCE FOR A SUBSET OF MULTIPLE LINEAR REGRESSION SLOPES

The hypothesis testing structure for testing a subset of population slopes also requires a slightly different structure than before.

$$H_0: \beta_{p+1} = \beta_{p+2} = \cdots = \beta_{p+q} = 0$$

$$H_A: \text{at least one } \beta_i \neq 0 \text{ (for } i = p+1, \dots, p+q)$$

### What's different about this test?

1. First, this test also does not have a confidence interval that corresponds to it.  
So we **must use a p-value**.
2. Also, our test statistic will also be an observation the **F-distribution** and our p-value is right tailed.  
ie.  $p - \text{value} = P(F \geq \text{test stat})$

3. Finally, we need to define two “nested” sample models.

**Full Model** (contains all the slopes)

$$\widehat{y}_{full} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_p x_p + \widehat{\beta}_{p+1} x_{p+1} + \cdots + \widehat{\beta}_{p+q} x_{p+q}$$

We can calculate the sum squared error for this model:

$$SSE_{full} = \sum_{i=1}^n (y_i - \widehat{y}_{full})^2 = \sum_{i=1}^n (y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i,1} + \cdots + \widehat{\beta}_p x_p + \widehat{\beta}_{p+1} x_{p+1} + \cdots + \widehat{\beta}_{p+q} x_{p+q}))^2$$

**Reduced Model** (contains just the slopes that you aren't testing)

$$\widehat{y}_{red} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_p x_p$$

We can also calculate the sum squared error for this model:

$$SSE_{red} = \sum_{i=1}^n (y_i - \widehat{y}_{red})^2 = \sum_{i=1}^n (y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i,1} + \cdots + \widehat{\beta}_p x_p))^2$$

### 7.4.1. CONDUCTING THE TEST

$$H_0: \beta_{p+1} = \beta_{p+2} = \cdots = \beta_{p+q} = 0$$

1. Set up the hypotheses

$$H_0: \beta_{p+1} = \beta_{p+2} = \cdots = \beta_{p+q} = 0$$

$$H_A: \text{at least one } \beta_i \neq 0 \text{ (for } i = p + 1, \dots, p + q)$$

2. Check the conditions for conducting inference on a population slope/intercept.

- The linearity condition holds.
- The constant residuals condition holds.
- The residuals are normal (with mean 0).
- The residuals are independent.
- The explanatory variables (if a multiple linear regression is used) are not collinear.

3. Calculate the test statistic:

$$\text{test stat} = \frac{(SSE_{red} - SSE_{full})/q}{(SSE_{full})/(n - (p + q) - 1)}$$

4. Calculate the p-value

$$p - \text{value} = P(F_{q, n-(p+q)-1} \geq \text{test stat})$$

5. Make a Decision

- If  $p - \text{value} < \alpha$ , then we “reject the null hypothesis.” And we say that “there IS sufficient evidence to suggest the alternative hypothesis.”
- If  $p - \text{value} \geq \alpha$ , then we “fail to reject the null hypothesis.” And we say that “there IS NOT sufficient evidence to suggest the alternative hypothesis.”



**Ex:** Is there sufficient evidence to suggest that at least one of the slopes that correspond to the age\_group variable in the population model (predicting weight with height, age\_group, and sex) are non-zero? **Go to the Jupyter notebook to answer this.**

## 8. LINEAR REGRESSION MODELS WITH INTERACTION VARIABLES

By fitting a multiple linear regression model to predict weight with height and sex, we get

$$\widehat{weight} = -56.2457 + 8.7207sex[T.male] + 0.7087height.$$

However, given that  $sex[T.male]$  is an indicator variable, this multiple linear regression line can be equivalently broken down into two separate regression models.

1. One model just for females:  $ex[T.male] = 0$

$$\widehat{weight} = -56.2457 + 8.7207(0) + 0.7087height.$$

$$\widehat{weight} = -56.2457 + 0.7087height.$$

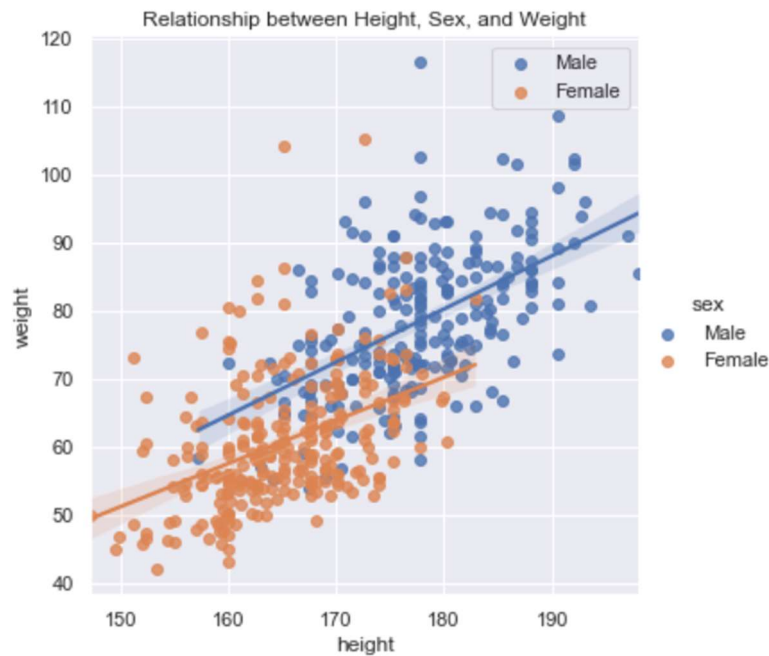
2. One model just for males:  $ex[T.male] = 1$

$$\widehat{weight} = -56.2457 + 8.7207(1) + 0.7087height.$$

$$\widehat{weight} = -47.525 + 0.7087height$$

These two models indicate the height and weight best fit lines have the same \_\_\_\_\_ for males and females, but a different \_\_\_\_\_.

However, the data below indicates that the slope of the height/weight best fit lines for males and females might be slightly different. **So how could we model these differing slopes for men and women?** If we suspect that the **interaction** of two explanatory variables  $x_i$  and  $x_j$  we can define an **interaction variable**  $x_i \cdot x_j$ , and add this to the linear regression model.



**Ex:** Go to the Jupyter notebook to set up a multiple linear regression model predicting weight with height, sex, and the interaction between height and sex.

#### OLS Regression Results

Dep. Variable:	weight	R-squared:	0.575
Model:	OLS	Adj. R-squared:	0.572
Method:	Least Squares	F-statistic:	217.5
Date:	Mon, 29 Mar 2021	Prob (F-statistic):	3.00e-89
Time:	19:52:45	Log-Likelihood:	-1748.3
No. Observations:	487	AIC:	3505.
Df Residuals:	483	BIC:	3521.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	-43.8193	13.791	-3.177	0.002	-70.917	-16.721
sex[T.Male]	-16.1357	19.885	-0.811	0.418	-55.208	22.936
height	0.6333	0.084	7.577	0.000	0.469	0.798
height:sex[T.Male]	0.1453	0.116	1.252	0.211	-0.083	0.373

Omnibus:	92.090	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	187.527
Skew:	1.030	Prob(JB):	1.90e-41
Kurtosis:	5.235	Cond. No.	1.11e+04

By modeling the interaction variable of height and sex, we get this model.

$$\widehat{weight} = -43.8193 - 16.1357sex[T.male] + 0.6333height + 0.1453(height \cdot sex[T.male])$$

Again, given that  $sex[T.male]$  is an indicator variable, this multiple linear regression line can be equivalently broken down into two separate regression models.

1. One model just for females:  $ex[T.male] = 0$

$$\widehat{weight} = -43.8193 - 16.1357(0) + 0.6333height + 0.1453(height \cdot 0)$$

$$\widehat{weight} = -43.8193 + 0.6333height.$$

2. One model just for males:  $ex[T.male] = 1$

$$\widehat{weight} = -43.8193 - 16.1357(1) + 0.6333height + 0.1453(height \cdot 1)$$

$$\widehat{weight} = -59.955 + 0.7786height$$

So now this new model indicates that the male and female relationships between height and weight have different intercepts AND slopes.

**Ex:** Is there sufficient evidence to suggest that the slope of the interaction variable (of height and sex) in the population model (that predicts weight with height, sex, and the interaction of height and sex) is non-zero?

## 9. MAKING A PREDICTION WITH A MULTIPLE LINEAR REGRESSION

Making predictions with multiple linear regression is the same procedure used by simple linear regression.

**Ex:** Set up the multiple linear regression model that predicts weight with height, sex, and age group. Then predict the weight of a 20 year old woman that is 170cm tall. **Go to Jupyter notebook.**

## 10. LINEAR TRANSFORMATIONS: WHAT TO TRY WHEN YOUR MULTIPLE LINEAR REGRESSION CONDITIONS AREN'T MET

See unit 13 notebook (section 10).