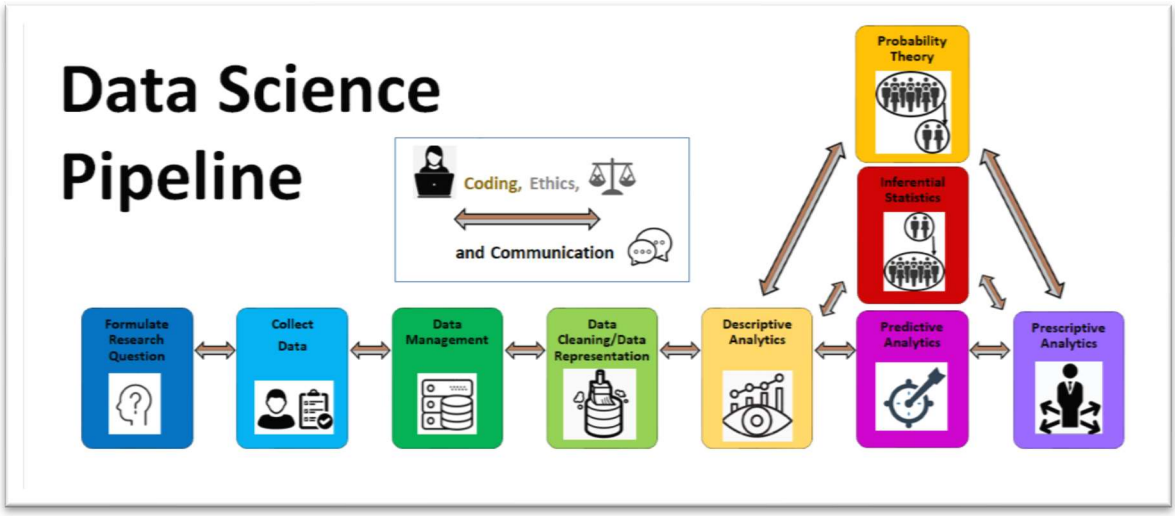
Unit 12: Simple Linear Regression Modeling

Case Studies:

- To introduce the concept of **simple linear regression model** between **two numerical variables (where one is a response variable and one is an explanatory variable)** we will examine the relationship between mother and daughter heights.

Purpose of this Lectures



1. **Analyses for Associations**
2. **Association Analyses Summary:** Numerical Explanatory Variable-> Numerical Response Variable
3. **Basic Descriptive Analytics for the Sample Data**
   - 3.1. Visualizations for the relationship between two numerical variables
   - 3.2. Summary statistics for the relationship between two numerical variables
     - 3.2.1. Covariance
     - 3.2.2. Correlation Coefficient (R)
4. **Modeling the Sample Data: Ordinary Least Squares Regression – Simple Linear Regression**
   - 4.1. Finding a Best Fit Line
   - 4.2. Evaluating the Model Fit
5. **Conducting Inference for the Population Slope(s) and Population Intercept of a Simple Linear Regression Line for the Population Data**
   - 5.1. Properties of the Sampling Distribution of Sample Slopes
   - 5.2. Checking the Conditions for Population Slope(s)/Coefficient Inference
   - 5.3. Creating a Confidence Interval for a Population Slope
   - 5.4. Conducting a Hypothesis Test for a Population Slope, Testing the Claim H_A: $\beta_i \neq 0$ with a p-value
6. **Making a Prediction with a Simple Linear Regression**

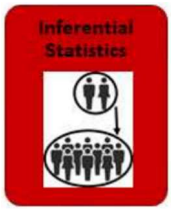
Additional Resources

Chapter 7 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: NUMERICAL

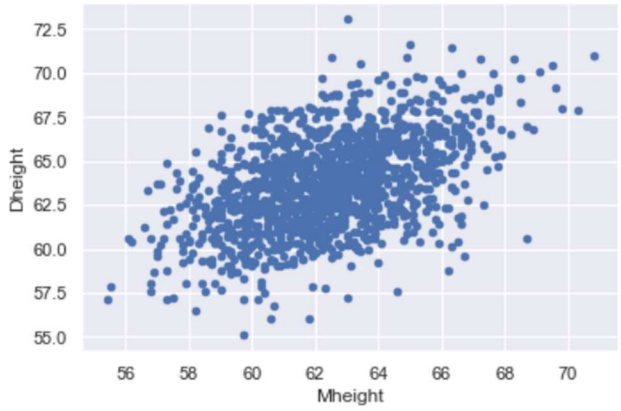
Research Questions about Associations	**Type of Variables Involved in the Association Test**	**Explanatory Variable:** Numerical Variable **Response Variable:** Numerical Variable
	Example	Is there an association between **mother height** and **daughter height**? 
	Type of Association *(Way to Quantify Association)*	**Simple Linear Regression Model** *(linear relationship between explanatory variable (x) and response variable (y))*
Descriptive Analytics	**How to Describe an Association in a Sample?**	1. Covariance 2. Correlation 3. Simple Linear Regression Model: • $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$ • R^2 of the model
	When is this analysis *(for the sample)* appropriate to use?	Linearity condition is met
Inferential Statistics	**How to Infer an Association for a Population?**	Conduct inference on the population parameter β_1, where $\hat{y} = \beta_0 + \beta_1 x$ is the simple linear regression for the population.
	When is this analysis *(for the population)* 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.
Predictive Analytics	**Making Predictions**	Use your simple linear regression line to make predictions $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$
	How to quantify the performance of your prediction(s)?	• Individual Data Point: residual • All Data: root mean square error (RMSE)

3. Basic Descriptive Analytics for Sample Data – two numerical variables

3.1. Visualizations

We can use a **scatterplot** to visualize the association between two numerical variables.

- The explanatory variable usually goes on the x-axis.
- The response variable usually goes on the y-axis.



There are four things we should always be prepared to **describe** about the relationship between two numerical variables in a dataset.

1. The **direction** of the relationship.

2. The **shape** of the relationship.

3. The **strength** of the relationship.

4. Are there any **outliers** in the data.

Ex: Describe the relationship between the mother heights and daughter heights using the scatterplot above.

3.2. SUMMARY STATISTICS

Ex: What is the **strength** of the linear relationship of mother and daughter heights by looking at the scatterplot above?

a. No association.
b. Weak association.
c. Moderate association.
d. Moderately strong association
e. Strong association.

Q: How can we quantify the association of a linear relationship such that this qualitative assessment is less contested?

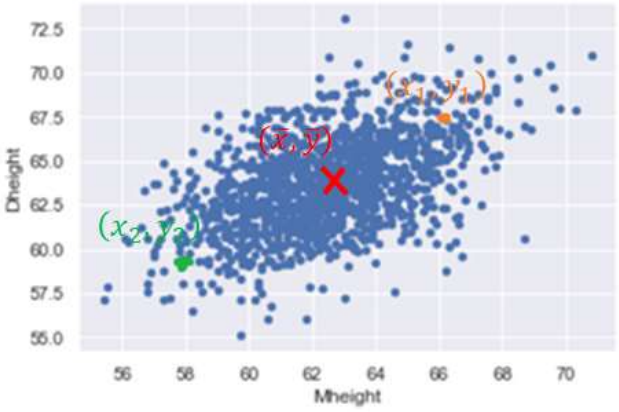
A: Use the **covariance** or the **correlation (R)** of the linear relationship.

Covariance between two numerical variables (in a sample)

If we have a set of numerical variables values $(x_1, x_2, \ldots, x_n)$ and another set of numerical variable values $(y_1, y_2, \ldots, y_n)$ The **sample covariance** between them is defined as:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Ex: Will $(x_1 - \bar{x})(y_1 - \bar{y})$ and $(x_2 - \bar{x})(y_2 - \bar{y})$ be positive or negative (use the image shown below)? Will we expect the covariance to be positive or negative?



Range:

How to interpret:

Ex: If we were to convert the heights to centimeters, would we expect the covariance to increase, decrease, or stay the same?

Would we expect the relationship to get stronger, get weaker, or stay the same?

Because want a standardized way to measure the strength of a linear relationship between two numerical variables, we introduce the **correlation coefficient (R).**

Covariance between two numerical variables (in a sample)

If s_x is the standard deviation of the numerical variable values $(x_1, x_2, \ldots, x_n)$ and s_y is the standard deviation of the numerical variable values $(y_1, y_2, \ldots, y_n)$, then we can define the **correlation** between these two numerical variables as:

$$R = \frac{s_{xy}}{s_x s_y}$$

Range:

How to interpret:

When to use:

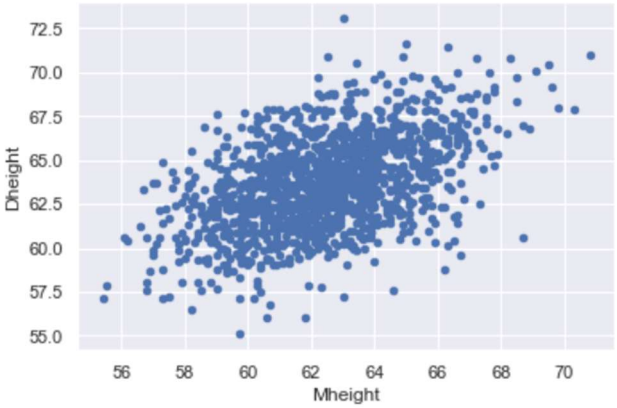
Ex: Use Python to calculate the covariance and correlation coefficient of the relationship between the mother and daughter heights. Are we allowed to use this correlation coefficient to quantify the strength of direction of the relationship between the daughter and mother heights?

4. Modeling the Sample Data: Ordinary Least Squares Regression – Simple Linear Regression –> Just one Slope

4.1. Finding a Best Fit Line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Q: What is one way to create a "best fit line" for this sample data?



Fitted Values

```
62.259733
61.447113
62.747306
62.801480
63.397402
```

Data

	Mheight	Dheight
0	59.7	55.1
1	58.2	56.5
2	60.6	56.0
3	60.7	56.8
4	61.8	56.0

A: We can fit an **Ordinary Least Squares Regression** line to this data.

Goal for Finding the Ordinary Least Squares Regression Line: Find an intercept value $\hat{\beta}_0$ and a slope $\hat{\beta}_1$ for the equation value $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ that **minimizes**

$$\sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$
$$= (y_1 - (\hat{\beta}_0 + \hat{\beta}_1 x_1))^2 + (y_2 - (\hat{\beta}_0 + \hat{\beta}_1 x_2))^2 + \cdots + (y_n - (\hat{\beta}_0 + \hat{\beta}_1 x_n))^2$$

General Idea: How do we find these optimal values of intercept value $\hat{\beta}_0$ and a slope $\hat{\beta}_1$ that minimize this equation?

Just for Simple Linear Regression:

When we are dealing with a simple linear regression, the optimal values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize $\sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$ end up being the following:

- $\hat{\beta}_1 = R\frac{s_y}{s_x}$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\,\bar{x}$

Where:

- $\bar{y}$: mean of the response variable
- $\bar{x}$: mean of the explanatory variable
- s_y: standard deviation of the response variable
- s_x: standard deviation of the explanatory variable

Ex: Go to the notebook to first calculate the slope and intercept of the ordinary least squares best fit line "by hand". Then use this to formulate your line.

Definition: We call $(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ the **residual** of a given sample data point (x_i, y_i).

Definition: After finding the optimal values of our intercept value $\hat{\beta}_0$ and a slope $\hat{\beta}_1$, we can define the **residual sum of squares** (or the **sum squared error)** as

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 .$$

A simple linear regression line is an ordinary least squares best fit line that has _____ slope.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Ex: Use the Python code (given in the notebook) to formulate the best fit simple linear regression line for mother heights and daughter heights.

OLS Regression Results

Dep. Variable:	Dheight	**R-squared:**	0.241
Model:	OLS	**Adj. R-squared:**	0.240
Method:	Least Squares	**F-statistic:**	435.5
Date:	Tue, 13 Oct 2020	**Prob (F-statistic):**	3.22e-84
Time:	00:26:58	**Log-Likelihood:**	-3075.0
No. Observations:	1375	**AIC:**	6154.
Df Residuals:	1373	**BIC:**	6164.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	29.9174	1.622	18.439	0.000	26.735	33.100
Mheight	0.5417	0.026	20.868	0.000	0.491	0.593

Omnibus:	1.412	**Durbin-Watson:**	0.126
Prob(Omnibus):	0.494	**Jarque-Bera (JB):**	1.353
Skew:	0.002	**Prob(JB):**	0.508
Kurtosis:	3.154	**Cond. No.**	1.66e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.66e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Notes on Notation for your regression lines:

4.2. EVALUATING THE MODEL FIT

Definition: We can use $\boldsymbol{R^2}$ of a given linear regression model to quantify what **percent** of the variability of the response variable was explained by the model.

Equation: We can calculate $\boldsymbol{R^2} = \frac{\boldsymbol{SST-SSE}}{\boldsymbol{SST}} = \mathbf{1} - \frac{\boldsymbol{SSE}}{\boldsymbol{SST}}$, where

- $\boldsymbol{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 =$_____

- $\boldsymbol{SSE} = \sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 =$_____

- $\boldsymbol{SSR} = \boldsymbol{SST} - \boldsymbol{SSE} =$_____

Special Case ONLY for Simple Linear Regression:

If your model is a simple linear regression model (one slope and hence one explanatory variable), then the $\boldsymbol{R^2}$ equivalently quantifies how much of the variability of the response variable was explained by the model.

You can also calculate the correlation coefficient of the model as:

$$\boldsymbol{R^2} = (\boldsymbol{R})^2$$

THIS ONLY WORKS FOR SIMPLE LINEAR REGRESSION!

Ex: Find the R^2 in the model output for your simple linear regression (that models mother and daughter heights). Use the R for these two numerical variables to also calculate this R^2.

OLS Regression Results

Dep. Variable:	Dheight	**R-squared:**	0.241
Model:	OLS	**Adj. R-squared:**	0.240
Method:	Least Squares	**F-statistic:**	435.5
Date:	Tue, 13 Oct 2020	**Prob (F-statistic):**	3.22e-84
Time:	00:26:58	**Log-Likelihood:**	-3075.0
No. Observations:	1375	**AIC:**	6154.
Df Residuals:	1373	**BIC:**	6164.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	29.9174	1.622	18.439	0.000	26.735	33.100
Mheight	0.5417	0.026	20.868	0.000	0.491	0.593

Omnibus:	1.412	**Durbin-Watson:**	0.126
Prob(Omnibus):	0.494	**Jarque-Bera (JB):**	1.353
Skew:	0.002	**Prob(JB):**	0.508
Kurtosis:	3.154	**Cond. No.**	1.66e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.66e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

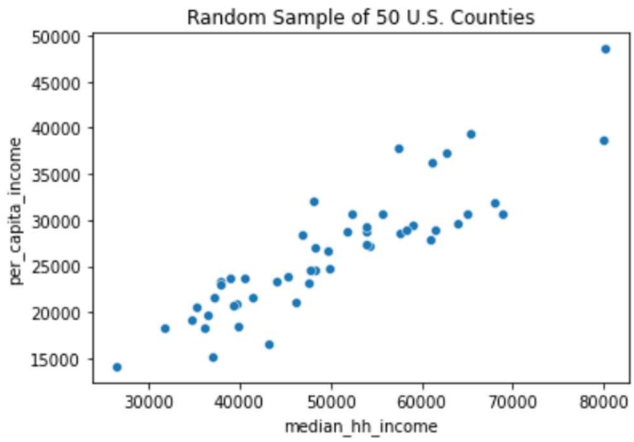
What percent of the total variability of daughter heights does this model explain?

What percent of total variability of daughter heights does this model NOT explain?

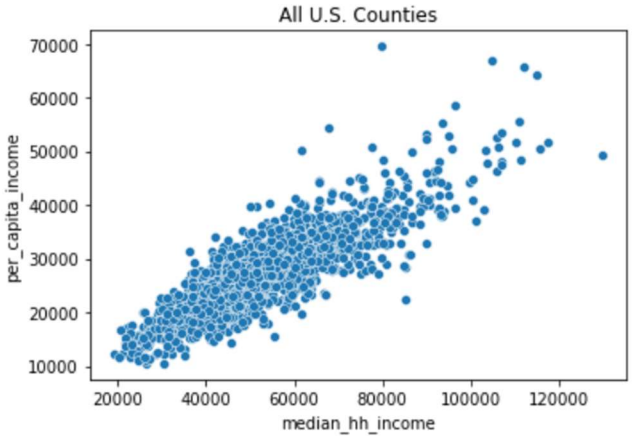
5. Conducting Inference for the Population Slope(s) and Population intercept of a Simple Linear Regression Line for the population data

We can fit a simple linear regression line for a sample of data or a population of data, however the notation is slightly different.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



$$\hat{y} = \beta_0 + \beta_1 x$$



Much like when conducting inference on other population parameters, we can use $\hat{\beta}_i$ as a point estimate (or a sample statistic) to conduct inference on the population parameter β_i.

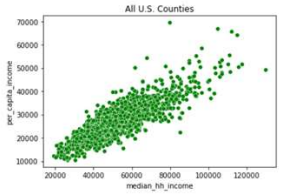
Similarly, we can do this by:

- Creating a **confidence interval** for β_i to attain a plausible range of values for β_i.

- Conducting a **hypothesis test** for β_i to test the claim (for instance) $\beta_i \neq 0$ where we can make a decision about this claim by using either a:
  - Confidence interval or
  - a p-value.

5.1. Properties of the Sampling Distribution of Sample Slopes

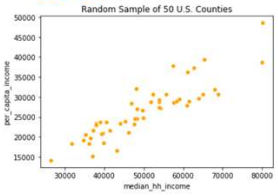
Population of Numerical



Best Fit Ordinary Least Squares Regression Line for Population:

$$\hat{y} = \beta_0 + \beta_1 x$$

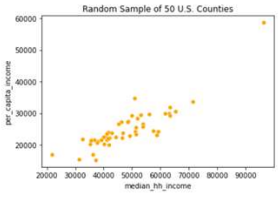
Sample of Numerical Data

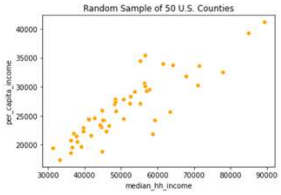


Best Fit Ordinary Least Squares Regression Line for Sample (of size **n**):

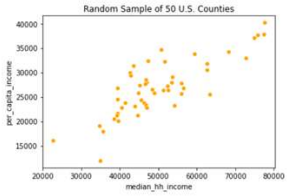
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Collect Many Random Samples (all of size same size n) drawn with replacement.

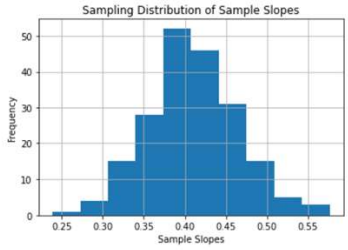




...



Sampling Distribution of Sample Slopes



Sample Slopes
0.40
0.40
...
0.31

Need to know the following to make an inference about *Unknown* Population Slope β_i:

Mean of **Sampling** Distribution $\approx E[\widehat{\beta_i}] =$ _____

Standard deviation of **Sampling** Distribution $\approx SD[\widehat{\beta_i}] =$ _____

Shape of **Sampling** Distribution is _____ when:

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

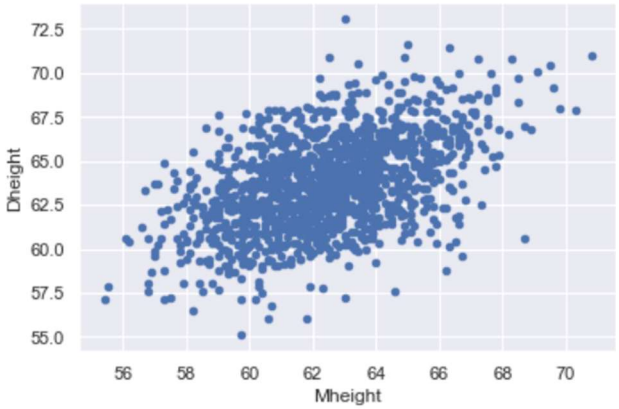
5.2. Checking the Conditions for Population Slope/Coefficient Inference

In order for our linear regression model to make accurate inferences (using the methods we will discuss in the next section), the following conditions must be satisfied.

1. Linearity Condition

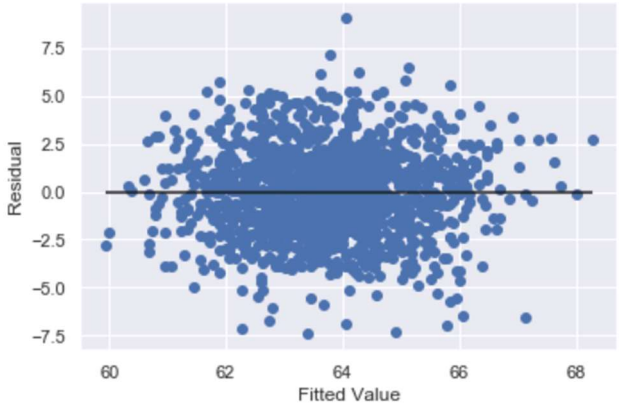
Is a linear regression model a good fit for our data? Or in other words is there a linear relationship between the explanatory variables and the response variable?

When we only have one explanatory variable, a linear relationship between the explanatory variable and the response variable is easy to see with a simple scatterplot of the two variables.



But what does it mean (and how can we detect) when multiple explanatory variables have a linear relationship with the response variable? **If there is a linear relationship, we would expect the residuals of all the sample data fitted to the model to be equally distributed above and below 0.** If not, it may suggest that a nonlinear regression model may be a better fit.

Rule of Thumb: If the "y-axis spread" of the points in this plot are roughly evenly distributed above and below the line as you move from left to right in this plot, then you can assume that this condition is met.



$$\hat{y} = 29.9174 + 0.5417(Mheight)$$

Residuals

```
-7.159733
-4.947113
-6.747306
-6.001480
-7.397402
```

Fitted Values

```
62.259733
61.447113
62.747306
62.801480
63.397402
```

Data

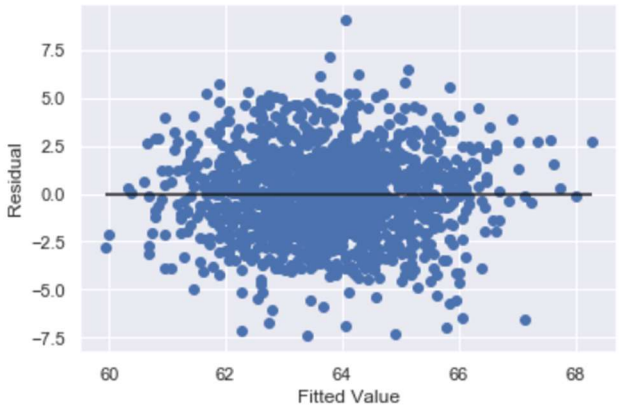
	Mheight	Dheight
0	59.7	55.1
1	58.2	56.5
2	60.6	56.0
3	60.7	56.8
4	61.8	56.0

Example of when this condition is not met.

2. Constant Variance of Residuals Condition

The next condition that must be satisfied is that the variance of the residuals must remain constant (for all fitted values. To check this condition, we can use the same plot as the one used for checking the linearity condition.

Rule of Thumb: If the "spread" (ie. y-axis) range of the points in this plot remain constant as you move from left to right in this plot, then you can assume that this condition is met.

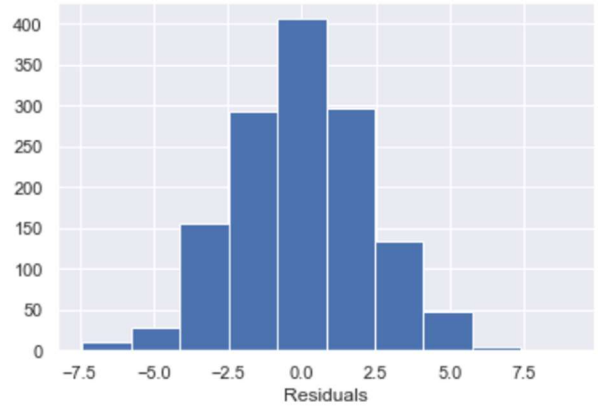


Example of when this condition is not met.

3. Residuals are Normal (with Mean of 0).

The next condition that must be satisfied is that the residuals must be normally distributed and must have a mean of zero.

To check this condition, we can look at a histogram of the residuals.



Example of when this condition is not met.

4. Residuals are independent.

There are many ways in which residuals may NOT be independent. You will discuss more of these ways (and other conditions to check) in later statistics classes. For now, we can say that the observations (at least) in the data must be independent in order for the residuals to be independent.

Thus the observations in the sample must be:

- randomly sampled and
- n<10% of the population size.

5.3. Creating a $(1-\alpha)\cdot 100\%$ Confidence interval for a population 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 β_i is calculated by**:

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

$$\widehat{\beta}_i \pm t^*_{\{n-p-1\}} SE_{\beta_i}$$

Notation:

Ex: Use the simple linear regression output table from Python to create a 95% confidence interval for the population slope.

Dep. Variable:	Dheight	**R-squared:**	0.241
Model:	OLS	**Adj. R-squared:**	0.240
Method:	Least Squares	**F-statistic:**	435.5
Date:	Wed, 24 Mar 2021	**Prob (F-statistic):**	3.22e-84
Time:	22:01:44	**Log-Likelihood:**	-3075.0
No. Observations:	1375	**AIC:**	6154.
Df Residuals:	1373	**BIC:**	6164.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	29.9174	1.622	18.439	0.000	26.735	33.100
Mheight	0.5417	0.026	20.868	0.000	0.491	0.593

Omnibus:	1.412	**Durbin-Watson:**	0.126
Prob(Omnibus):	0.494	**Jarque-Bera (JB):**	1.353
Skew:	0.002	**Prob(JB):**	0.508
Kurtosis:	3.154	**Cond. No.**	1.66e+03

5.4. Conducting a Hypothesis Test for a Population Slope, testing the claim $H_A: \beta_i \neq 0$ – with a p-value

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.

   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)**

   $$\widehat{\beta_i}$$

4. **Calculate the p-value**

$$\boldsymbol{p - value = 2P(T_{n-p-1} \geq |\frac{\widehat{\beta_i} - 0}{SE_{\widehat{\beta_i}}}|)}$$

5. **Make a Decision**

   a. If $\mathbf{p - value < \alpha}$, then we "reject the null hypothesis." And we say that "there IS sufficient evidence to suggest the alternative hypothesis."

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

Intuition behind p-value

$$p-value = P\left(\begin{array}{c} sample\ statistic\ that\ is\ at\ least \\ as\ suspicious\ (in\ favor\ of\ the\ alternative\ hypotheis)\ as\ the\ observed\ sample\ statistic \end{array} \middle|\ Null\ hypothesis\ is\ True\right)$$

$$= P\left(\begin{array}{c} \_\_\_\_\_\ that\ is\ at\ least \\ as\ suspicious\ (in\ favor\ of\ the\ alternative\ hypotheis)\ as\ \_\_\_\_\_ \end{array} \middle|\ Null\ hypothesis\ is\ True\right)$$

$$= P(\_\_\_\_\_ \geq \_\_\_\_\_\ OR\ \_\_\_\_\_ \leq \_\_\_\_\_)$$

$$= P(\_\_\_\_\_ \geq \_\_\_\_\_) + P(\_\_\_\_\_ \leq \_\_\_\_\_)$$

$$= \_\_\_\_\_ * P(\_\_\_\_\_ \geq \_\_\_\_\_)$$

Assuming that **Ho**: β_1 = _____

Ex: We would like to test whether there is a linear relationship in the population of ALL mother and daughter heights. **(See notebook for these questions and code).**

Use the simple linear regression output table to answer the following questions.

Dep. Variable:	Dheight	**R-squared:**	0.241
Model:	OLS	**Adj. R-squared:**	0.240
Method:	Least Squares	**F-statistic:**	435.5
Date:	Wed, 24 Mar 2021	**Prob (F-statistic):**	3.22e-84
Time:	22:01:44	**Log-Likelihood:**	-3075.0
No. Observations:	1375	**AIC:**	6154.
Df Residuals:	1373	**BIC:**	6164.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	29.9174	1.622	18.439	0.000	26.735	33.100
Mheight	0.5417	0.026	20.868	0.000	0.491	0.593

Omnibus:	1.412	**Durbin-Watson:**	0.126
Prob(Omnibus):	0.494	**Jarque-Bera (JB):**	1.353
Skew:	0.002	**Prob(JB):**	0.508
Kurtosis:	3.154	**Cond. No.**	1.66e+03

1. Set up the hypotheses for this test.

2. Make sure the conditions for this test hold.

3. Use the table to find the p-value for this test.

4. Use this p-value to make a conclusion for these hypotheses using a significance level of

5. What is the test statistic for this test?

6. Calculate this test statistic by hand.

7. Calculate the p-value by hand (and using your t-distribution object in Python).

8. Use your 95% confidence interval from the previous section to make a conclusion about your hypotheses.

6. MAKING A PREDICTION WITH A SIMPLE LINEAR REGRESSION

Ex: Use your simple linear regression equation to predict the height of a daughter whose mother is 66". **Go to the notebook for how to answer these questions by hand and by code.**

Ex: If this mother's daughter was *actually* 68", calculate the residual of this observation.