LECTURE 2
REGRESSION ANALYSIS
- SIMPLE LINEAR REGRESSION

AGENDA

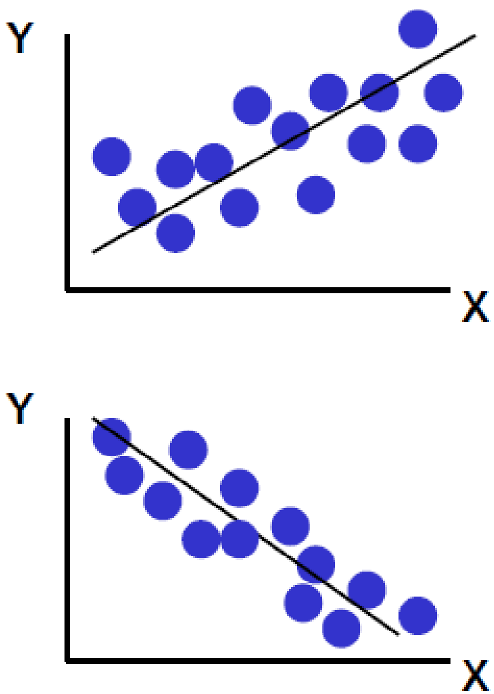
- Basic Concepts of Simple Linear Regression
- Data Analysis Using Simple Linear Regression Models
- Measures of Variation and Statistical Inference

ASSOCIATIONS BETWEEN TWO VARIABLES

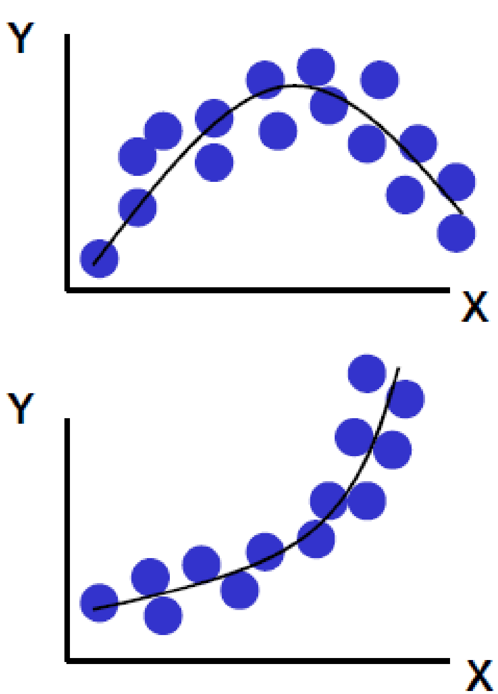
- To visualize the relationship between two numerical variables
  - Scatter plot (other name: X-Y plot)
- To measure the degree of linear association
  - Coefficient of Correlation (formal name: Pearson's correlation coefficient)
- To forecast one variable for given values of the other
  - Regression models
- Examples
  - Apartment price vs. Gross floor area
  - Weekly sales for chain stores vs. Number of customers

SCATTERPLOT

Linear relationships



Nonlinear relationships



COEFFICIENT OF CORRELATION

(Formal name: Pearson's correlation coefficient)

- (Sample) Linear correlation coefficient, r

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

  - Dimensionless
  - $-1 \leq r \leq +1$
  - "Sign" indicates the direction (positive / negative) of a linear relationship
  - "Magnitude" measures the strength of a linear relationship

LINEAR REGRESSION MODEL

- Input
  - Dependent / response variable, Y
    - The variable we wish to explain or predict
  - Independent / explanatory variable, X
    - The variable used to explain the dependent variable
- Output
  - A linear function that allows us to
    - Model causality*: Explain the variation of the dependent variable that is caused by the independent variable(s)
    - Provide prediction: Estimate the value of the dependent variable based on value(s) of the independent variable(s)

*Two other possibilities of causation even for a successful regression model:

1. Y is causing variation in X
2. There are other variables causing both Y and X to vary

FORMULATION OF SIMPLE LINEAR REGRESSION MODEL

- A simple linear regression model consists of two components
  - Regression line: A straight line that describes the dependence of the average value (conditional mean) of the Y-variable on one X-variable
  - Random error: The unexpected deviation of observed value from the expected value

Population Intercept*

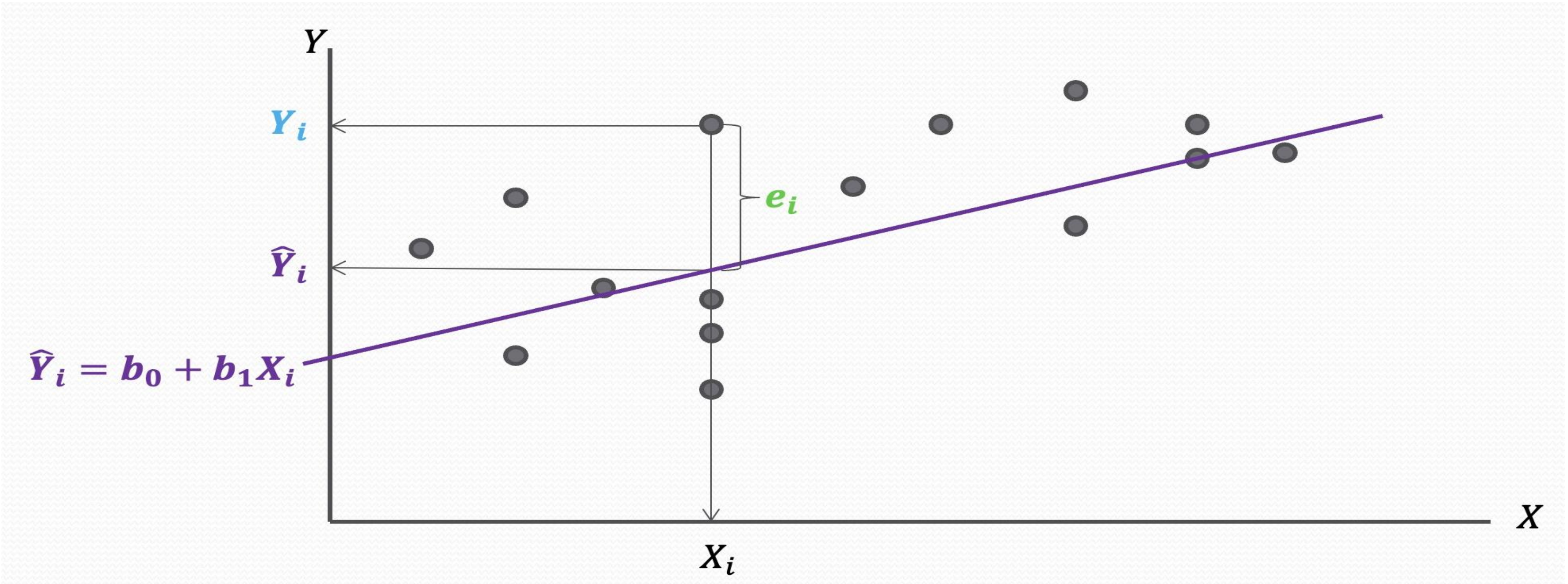
Population Slope Coefficient*

Independent Variable

Dependent Variable → $$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$ ← Random Error

*Population intercept (β_0) and population slope (β_1) are the 2 parameters in a simple linear regression model

FORMULATION OF LINEAR REGRESSION MODEL – CONT'D



- b_0 represents the sample intercept
- b_1 represents the sample slope coefficient
- e represents the random error

LEAST SQUARES METHOD

- b_0 and b_1 are estimated using the least squares method, which minimize the sum of squares errors (SSE)

$$SSE = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2$$

LEAST SQUARES METHOD

- The solution to b_0 and b_1 can be obtained by differentiating with respect to b_0 and b_1
- That is to solve for b_0 and b_1 in:

$$\frac{\partial \sum_{i=1}^{n} {e_i}^2}{\partial b_0} = -2\sum_{i=1}^{n}\left(Y_i - (b_0 + b_1 X_i)\right) = 0$$

and

$$\frac{\partial \sum_{i=1}^{n} {e_i}^2}{\partial b_1} = -2\sum_{i=1}^{n}X_i\left(Y_i - (b_0 + b_1 X_i)\right) = 0$$

simultaneously

LEAST SQUARES METHOD

- The solutions are

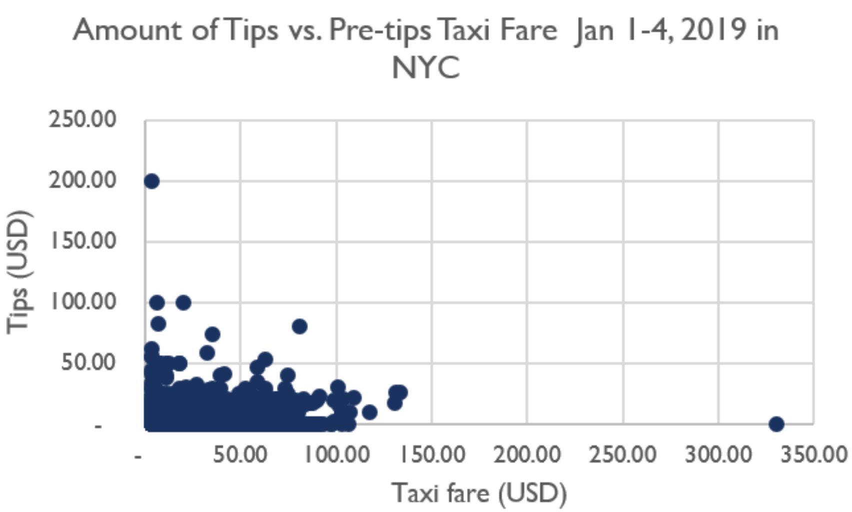
$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = r\frac{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = r(\frac{S_Y}{S_X})$$

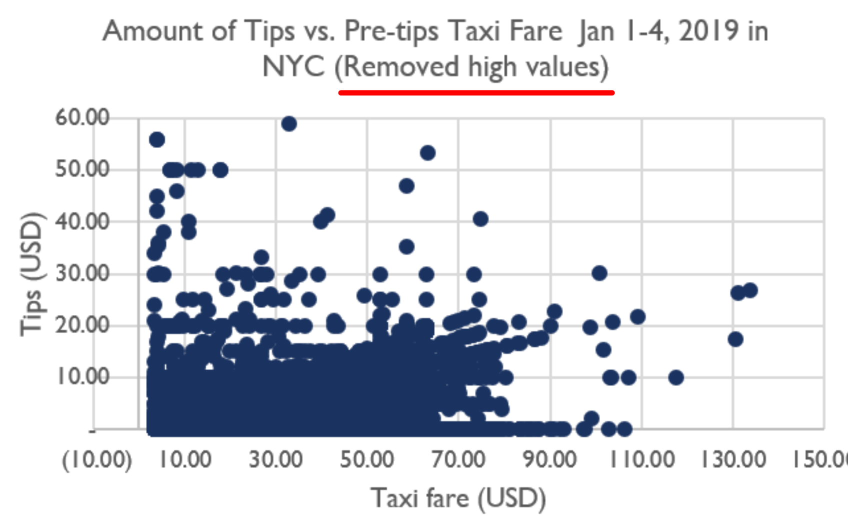
and

$$b_0 = \bar{Y} - b_1\bar{X}$$

EXAMPLE

- How much tips do riders pay their taxi driver in New York City (NYC)?
- Is there any relationship between the taxi fare and the size of the tip?

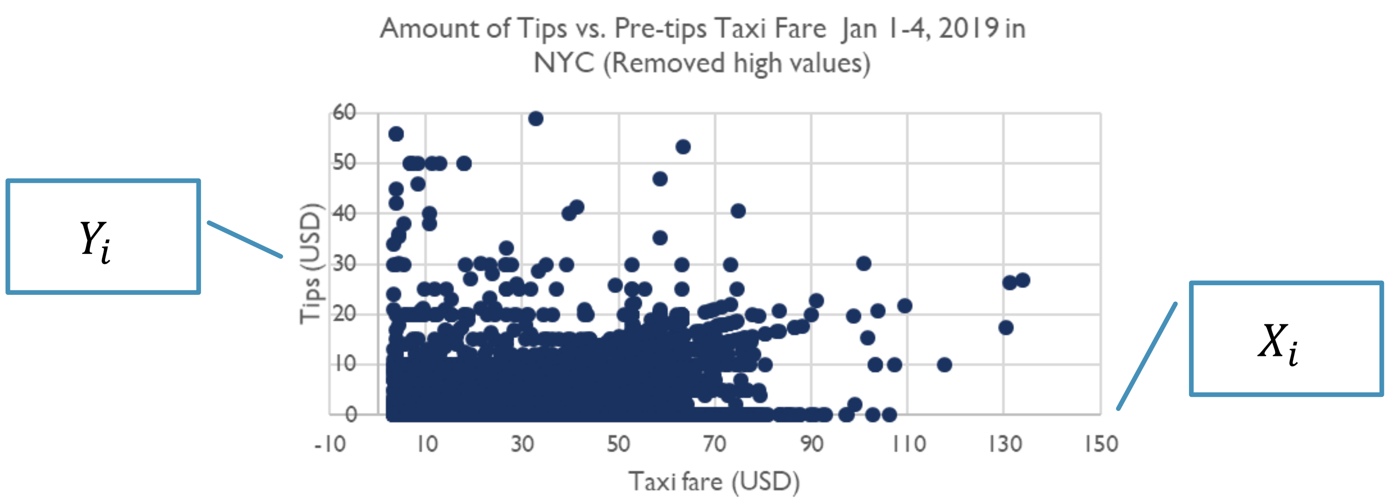




Source: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
TLC Trip Record Data: January 2019 Yellow Taxi Trip Records
Published by NYC Taxi & Limousine Commission
Accessed August 3, 2021

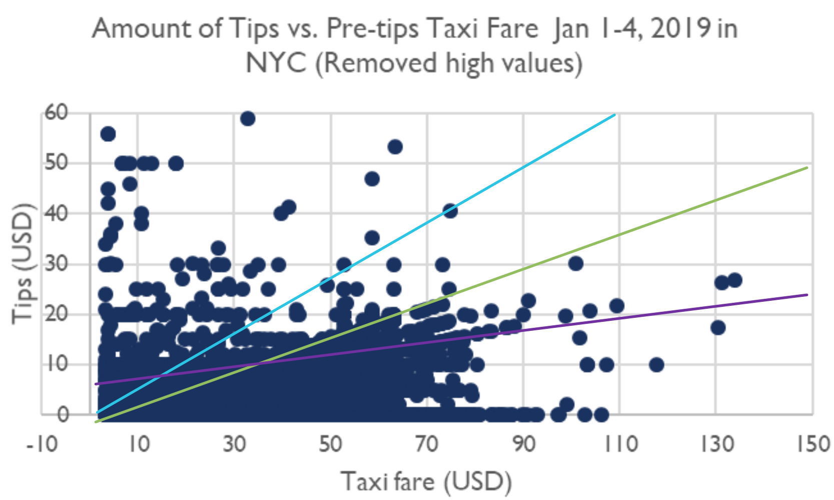
SIMPLE LINEAR REGRESSION

- Want to look at the relationship between two variables.
- Most common approach: consider a linear relationship between the two variables.
- Suppose our data is of the form (X_i, Y_i), where:
  - X_i is the pre-tip fare charged to the i-th customer,
  - Y_i is the tips paid by the i-th customer.



SIMPLE LINEAR REGRESSION

- Want to find values of b_0, b_1 such that $Y_i \approx b_0 + b_1 X_i$ for all customers.
- Implication: Tips (Y_i) increase by \$$b_1$ for each additional \$1 in taxi fare.
  - $b_1 < 0$ implies that tips decrease relative to the taxi fare.
- What are the right values of b_0, b_1 so that we can represent the data well?

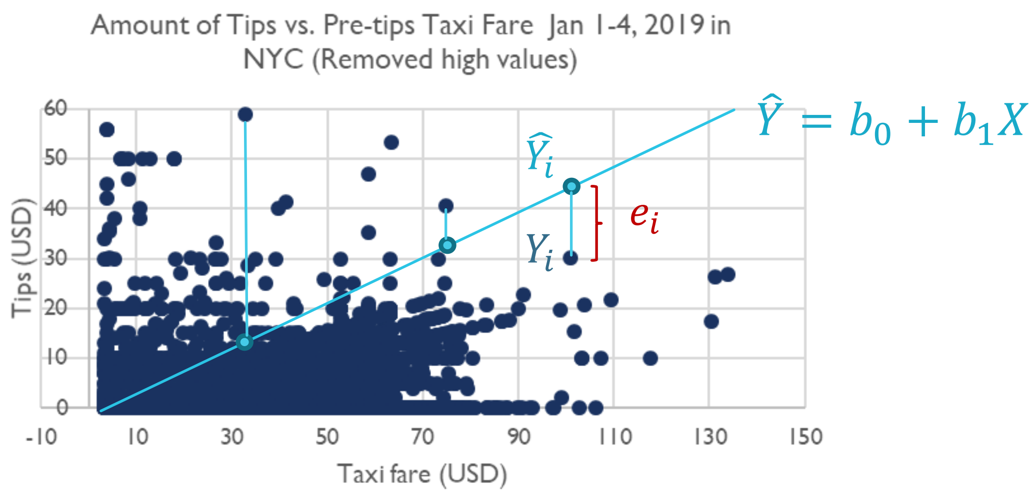


SIMPLE LINEAR REGRESSION

- Regardless of the values of b_0, b_1, there will be errors in our model because the data points don't lie on a straight line.
- Suppose we fix some values of b_0, b_1.
- Let $\widehat{Y_i}$ = the predicted value tips based on our model: $\widehat{Y_i} = \mathrm{b_0 + b_1 X_i}$.
- Then the error/residual for the i-th data point is $e_i = Y_i - \widehat{Y_i}$.
  - i.e. The true/observed value of the tips is $Y_i = b_0 + b_1 X_i + e_i$.

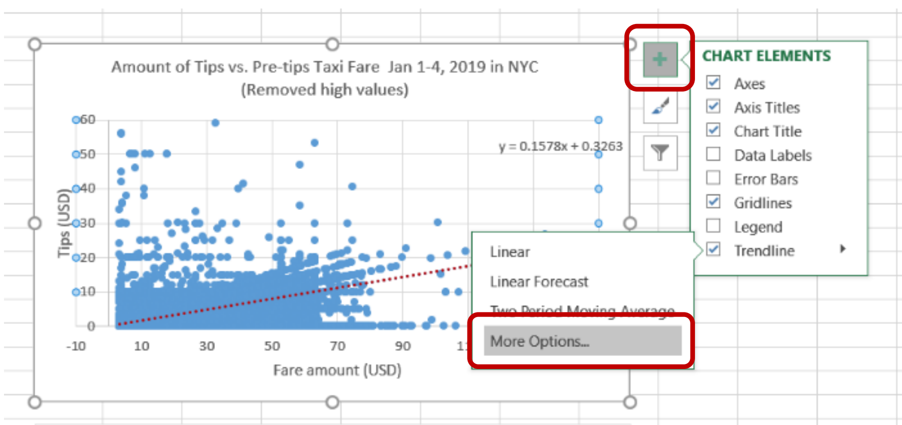
SIMPLE LINEAR REGRESSION

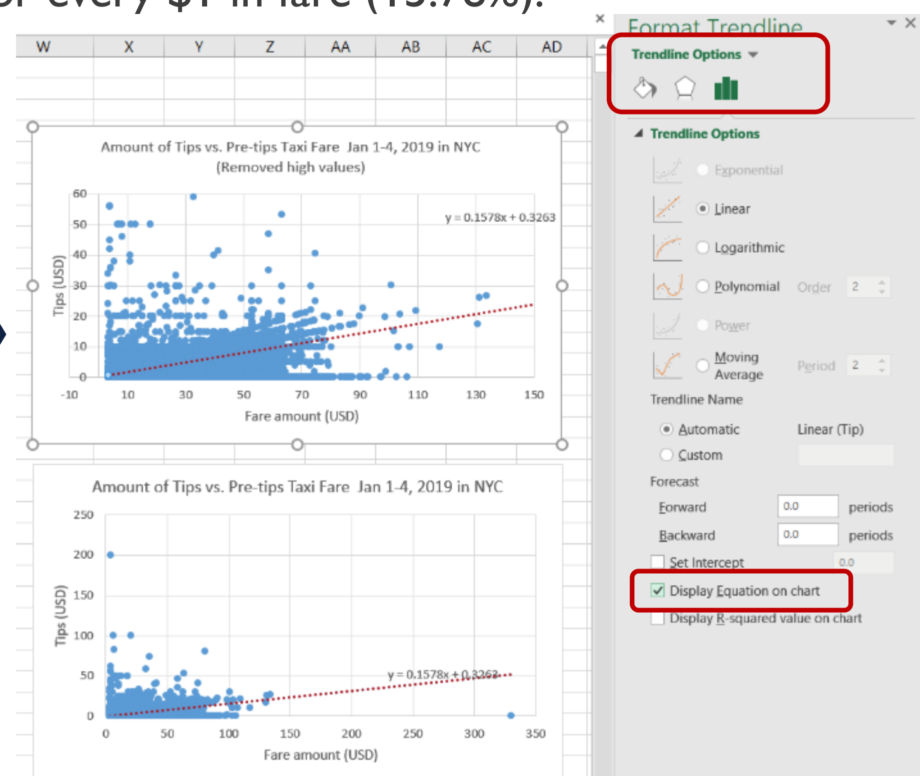
- Idea: We should minimize the amount of errors e_i when we choose b_0, b_1.
- We can't use the sum of e_i; the negative and positive errors could cancel out.
- Minimize the sum of square-errors: $\min \sum e_i^2 = \min \sum [Y_i - (b_0 + b_1 X_i)]^2$.
- Also known as least-squares regression model.



SIMPLE LINEAR REGRESSION

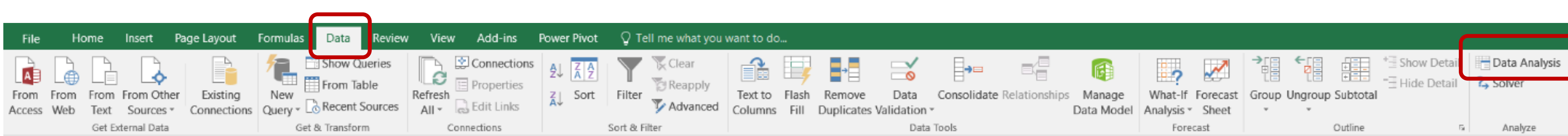
- How can we find b_0, b_1?
- Fast method: Use "trendline" function in Excel.
- $b_1 = 0.1578$; riders pay $0.16 in tips for every $1 in fare (15.78%).

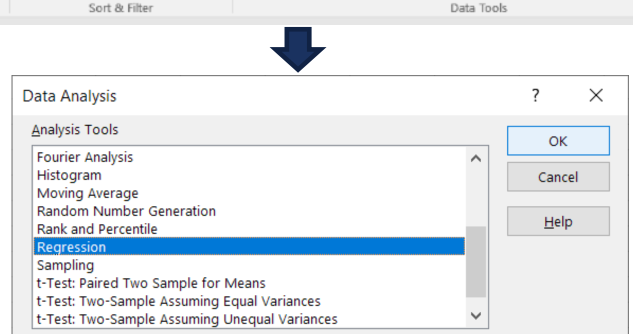




SIMPLE LINEAR REGRESSION

- Is the model "good"?
- Better/more informative method to find b_0, b_1: Use Regression tool in Excel.
- One-time step: File → Options → Add-ins → Analysis Toolpak (check and click OK).
- Subsequent access: Data → Analyze → Data Analysis → Regression.

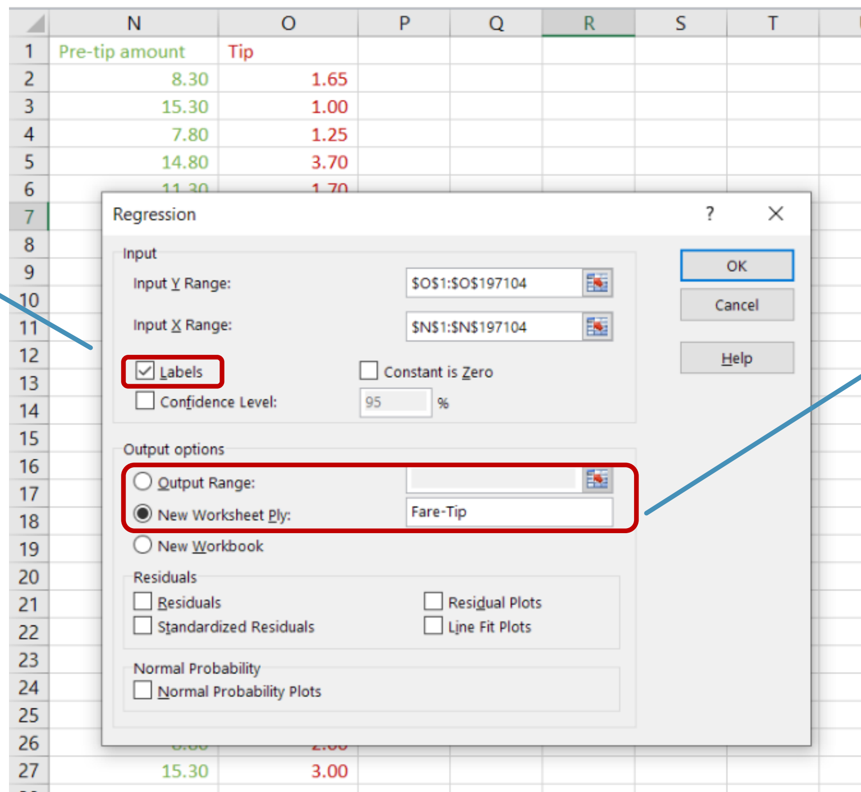




SIMPLE LINEAR REGRESSION

- Fill in the pop-up box:

Check this box if you have headers in your table.



You can choose to output on the same worksheet or on a new worksheet.

SIMPLE LINEAR REGRESSION

- Excel's Output:

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.743850822							
5	R Square	0.553314046							
6	Adjusted R Square	0.553311779							
7	Standard Error	1.624035606							
8	Observations	197103							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	643945.8062	643945.8062	244150.8416	0			
13	Residual	197101	519852.2419	2.637491651					
14	Total	197102	1163798.048						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	0.32631227	0.005744657	56.80273722	0	0.315052879	0.337571661	0.315052879	0.337571661
18	Pre-tip amount	0.157828366	0.000319415	494.1162227	0	0.157202319	0.158454412	0.157202319	0.158454412
19									

Sample intercept (b_0) and sample slope coefficient (b_1)

Sample estimates of the population intercept (β_0) and population slope (β_1)

SIMPLE LINEAR REGRESSION

SUMMARY OUTPUT				
Regression Statistics				
Multiple R	0.743850822			
R Square	0.553314046			
Adjusted R Square	0.553311779			
Standard Error	1.624035606			
Observations	197103			
ANOVA				
	df	*SS*	*MS*	*F*
Regression	1	643945.8062	643945.8062	24415(
Residual	197101	519852.2419	2.637491651	
Total	197102	1163798.048		
	Coefficients	*Standard Error*	*t Stat*	*P-valu*
Intercept	0.32631227	0.005744657	56.80273722	
Pre-tip amount	0.157828366	0.000319415	494.1162227	

- Multiple R*: Absolute value of Linear correlation coefficient "r".
- $-1 \leq r \leq 1$, no dimension or unit.
- $r > 0$: positive correlation (as X increases, then Y also increases).
- $r < 0$: negative correlation.
- Magnitude of r (without +/-) indicates the strength of the relationship.
  - $|r| \to 1$ means a stronger relationship.

*Correlation coefficient between observed Y and predicted $\hat{Y}$

MEASURES OF VARIATION

- Total variation of the Y-variable is made up of two parts

$$SST = SSR + SSE$$

where

Sum Squares Total, $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

- Variation of the Y_i values around their mean, $\bar{Y}$

Sum Squares Regression, $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$

- Variation of the Y_i values explained by the regression equation relating Y with X

Sum Squares Errors, $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

- Variation attributable to factors other than those considered in the regression equation

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.743850822
R Square	0.553314046
Adjusted R Square	0.553311779
Standard Error	1.624035606
Observations	197103

ANOVA

	df	*SS*
Regression	1	SSR 643945.8062
Residual	197101	SSE 519852.2419
Total	197102	SST 1163798.048

	Coefficients	*Standard Error*
Intercept	0.32631227	0.005744657
Pre-tip amount	0.157828366	0.000319415

SIMPLE LINEAR REGRESSION

SUMMARY OUTPUT				
Regression Statistics				
Multiple R	0.743850822			
R Square	0.553314046			
Adjusted R Square	0.553311779			
Standard Error	1.624035606			
Observations	197103			
ANOVA				
	df	*SS*	*MS*	*F*
Regression	1	SSR 643945.8062	643945.8062	24415(
Residual	197101	SSE 519852.2419	2.637491651	
Total	197102	SST 1163798.048		
	Coefficients	*Standard Error*	*t Stat*	*P-valu*
Intercept	0.32631227	0.005744657	56.80273722	
Pre-tip amount	0.157828366	0.000319415	494.1162227	

- R Square: Coefficient of determination $= r^2$.
- $0 \leq r^2 \leq 1$.
- $r^2 \to 1$ means a stronger relationship.
- In fact, let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ (average).
- $r^2 = \frac{\sum(\widehat{Y_i} - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

MEASURES OF VARIATION

- Coefficient of determination, r^2

$$r^2 = \frac{SSR}{SST}$$

- $0 \leq r^2 \leq 1$
- Measures the proportion of variation of the Y_i values that is explained by the regression equation with the independent variable X
- Measures the goodness of fit of the regression model

SIMPLE LINEAR REGRESSION

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.743850822
R Square	0.553314046
Adjusted R Square	0.553311779
Standard Error	1.624035606
Observations	197103

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	643945.8062	643945.8062	244150.8416	0
Residual	197101	519852.2419	2.637491651		
Total	197102	1163798.048			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.32631227	0.005744657	56.80273722	0	0.315052879	0.337571661	0.315052879	0.337571661
Pre-tip amount	0.157828366	0.000319415	494.1162227	0	0.157202319	0.158454412	0.157202319	0.158454412

INFERENCE ABOUT THE PARAMETERS

- **t-test** for **a slope** coefficient

  $H_0: \beta_1 = 0$ (no linear relationship)

  $H_1: \beta_1 \neq 0$ (linear relationship exists)

  $t = \frac{b_1 - \beta_1}{S_{b_1}}$ with $(n-2)$ degrees of freedom (d.f.)

  where S_{b_1} = standard error* of the slope

  1. Rejection region approach

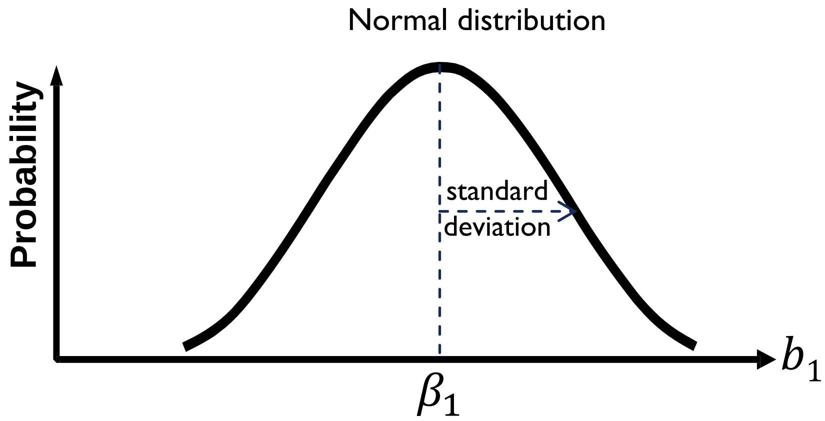
     Reject H_0 if |t| > $\text{C.V.} = t_{\alpha/2,(n-2)}$

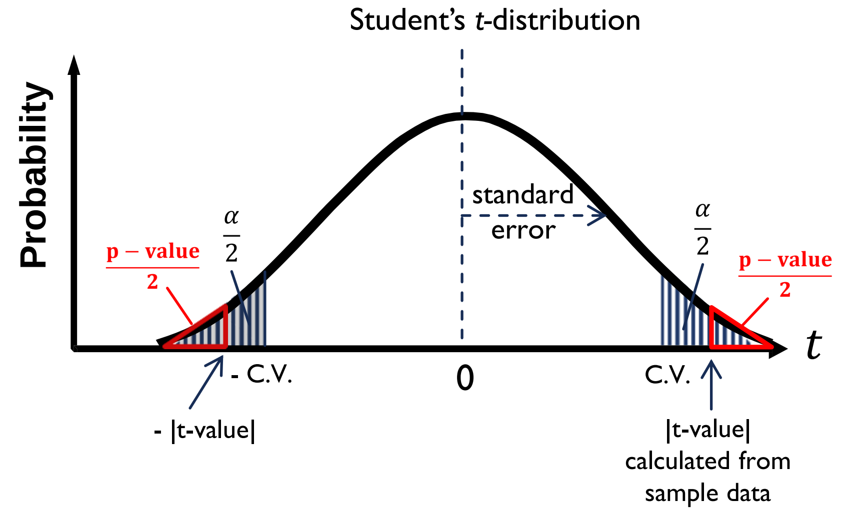
  or

  2. **p-value** approach

     p-value $= P(t \geq |t|)$

     Reject H_0 if **p-value** < α

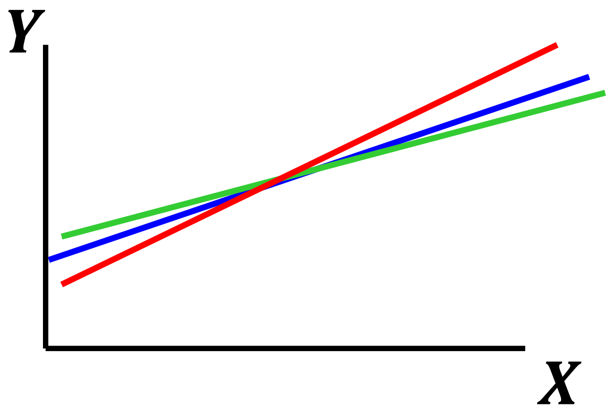


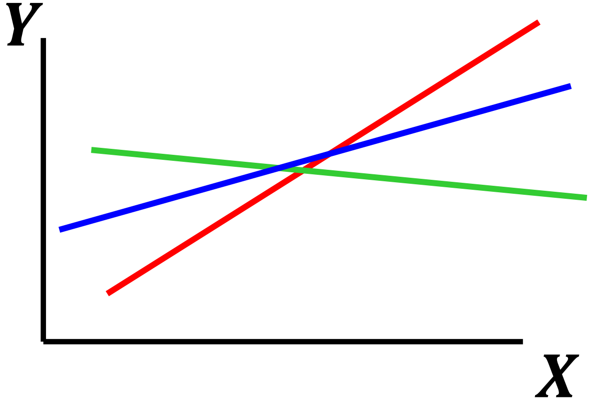


*standard error: standard deviation of the sampling distribution of a statistic (e.g., slope b_1, sample mean $\bar{X}$)

INFERENCE ABOUT THE PARAMETERS

- S_{b_1} measures the variation in the slope of regression lines from different possible samples





- $S_{b_1} = \sqrt{\dfrac{{S_e}^2}{\sum(X_i - \bar{X})^2}}$

  where S_e = variation of the errors around the regression line

SIMPLE LINEAR REGRESSION

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.743850822
R Square	0.553314046
Adjusted R Square	0.553311779
Standard Error	1.624035606
Observations	197103

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	643945.8062	643945.8062	244150.8416	0
Residual	197101	519852.2419	2.637491651		
Total	197102	1163798.048			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.32631227	0.005744657	56.80273722	0	0.315052879	0.337571661	0.315052879	0.337571661
Pre-tip amount	0.157828366	0.000319415	494.1162227	0	0.157202319	0.158454412	0.157202319	0.158454412

Confidence interval estimate

for slope coefficient

$$b_1 \pm t_{\alpha/2, n-2} S_{b_1}$$

= [0.1572, 0.1585]

CONFIDENCE INTERVAL

- Confidence interval estimate for slope coefficient

  $$b_1 \pm t_{\alpha/2, n-K-1} S_{b_1}$$

- Implication
  - The CI for slope coefficient does not include zero, indicating the independent variable significantly affects the dependent variable
  - Both boundaries of the CI are positive (negative), telling that the independent variable is very likely to be positively (negatively) related to the dependent variable

SUMMARY

- Scatter plot
- Coefficient of correlation
- Simple linear regression model
  - Model building
  - Model evaluation (coefficient of determination; t-test, confidence interval for slope coefficient)
- Next week: Multiple regression