



LECTURE 3

REGRESSION ANALYSIS

- MULTIPLE REGRESSION

AGENDA

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

FORMULATION OF MULTIPLE REGRESSION MODEL

- A multiple regression model is to relate **one dependent** variable with **two or more independent** variables in a **linear** function

Population Intercept

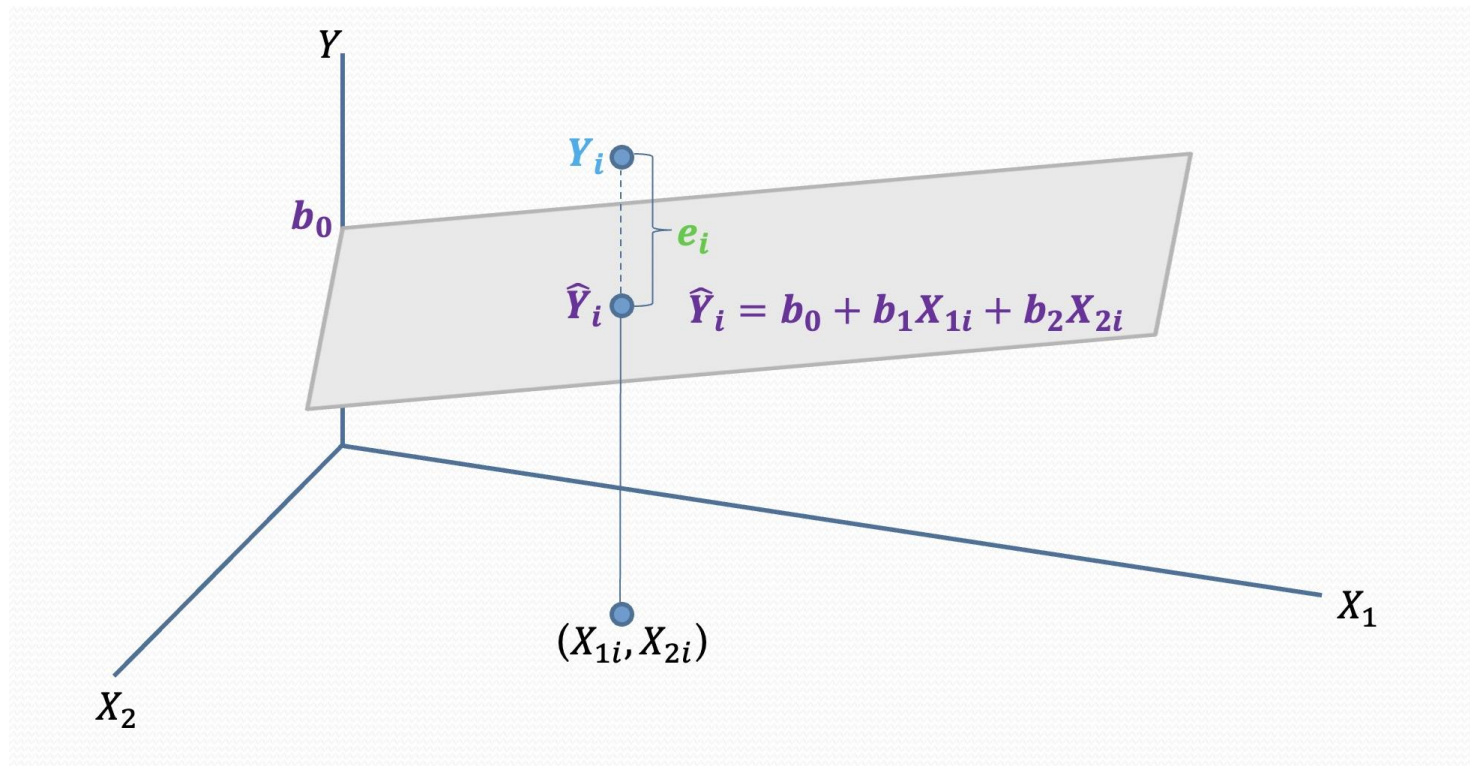
Population Slope Coefficients

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + \varepsilon_i$$

Dependent Variable Independent Variable Random Error

- K is the number of independent variables (e.g., $K = 1$ for simple linear regression)
- $\beta_0, \beta_1, \beta_2, \dots, \beta_K$ are the $K+1$ parameters in a multiple regression model with K independent variables
- $b_0, b_1, b_2, \dots, b_K$ are used to represent sample intercept and sample slope coefficients

FORMULATION OF MULTIPLE REGRESSION MODEL



FORMULATION OF MULTIPLE REGRESSION MODEL

- Coefficients in a multiple regression **net out the impact** of each independent variable in the regression equation
- The estimated slope coefficient, b_j , measures the change in the average value of Y as a result of a one-unit increase in X_j , **holding all other independent variables constant – “ceteris paribus effect”**

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + \mathbf{b_jX_j} + \cdots + b_KX_K.$$

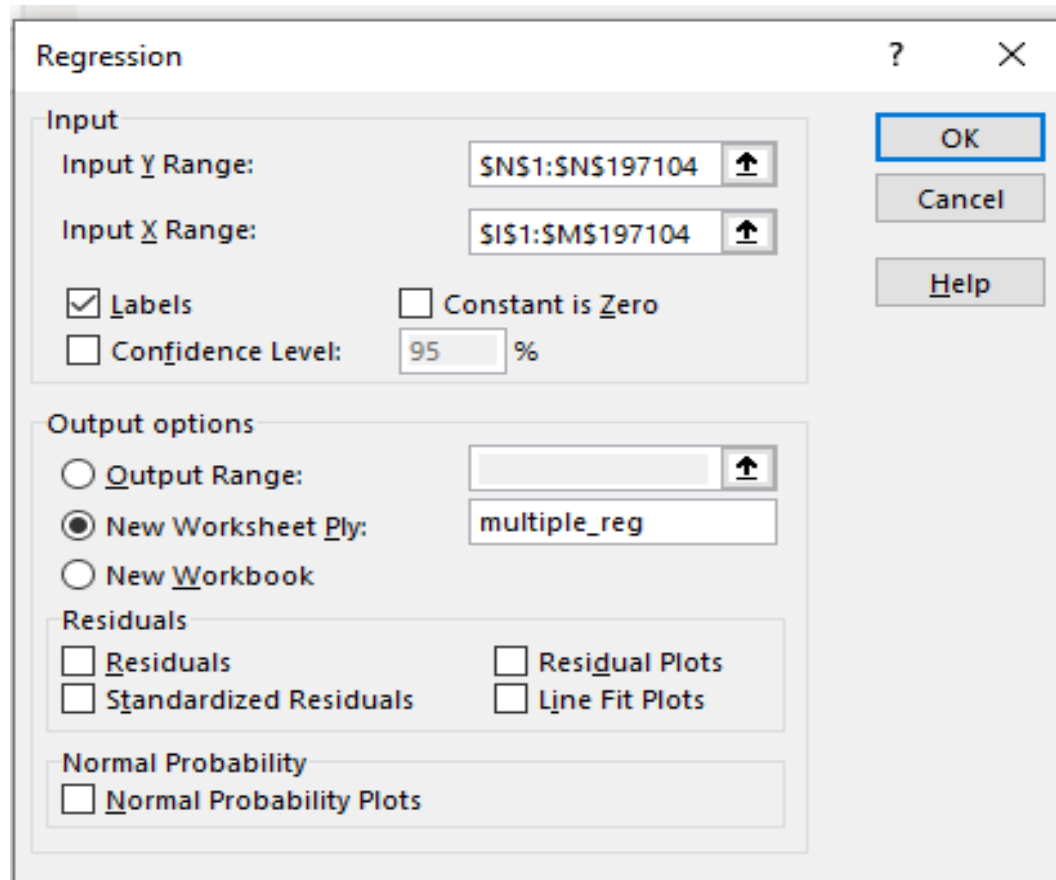
The diagram illustrates the concept of holding other variables constant. Four arrows originate from the text "remain constant" and point to the independent variables X_1 , X_2 , X_j , and X_K in the regression equation. This indicates that when measuring the effect of a change in X_j on \hat{Y} , the values of X_1 , X_2 , and X_K are held constant.

EXAMPLE

- Recall the example in the last topic, we wish to find possible factors affecting taxi tips in New York City (NYC). The relationship between the taxi fare and the size of the tip is estimated using a 2-variable regression model.
- Today we wish to include more factors that could possibly affect tips:
 - Area
 - number of riders
 - Holiday reasons
 -

MULTIPLE LINEAR REGRESSION

- Fill in the pop-up box:



The image shows the 'Regression' dialog box in Microsoft Excel. The dialog box is titled 'Regression' and has a question mark icon and a close button (X) in the top right corner. It is divided into several sections:

- Input:**
 - Input Y Range:** A text box containing '\$N\$1:\$N\$197104' and a selection icon (upward arrow).
 - Input X Range:** A text box containing '\$I\$1:\$M\$197104' and a selection icon (upward arrow).
 - ☒ **Labels**: A checkbox that is currently checked.
 - ☐ **Constant is Zero**: A checkbox that is currently unchecked.
 - ☐ **Confidence Level:** A checkbox that is currently unchecked, followed by a text box containing '95' and a '%' symbol.
- Output options:**
 - ☐ **Output Range:** A radio button that is currently unselected, followed by a text box and a selection icon (upward arrow).
 - ☒ **New Worksheet Ply:** A radio button that is currently selected, followed by a text box containing 'multiple_reg'.
 - ☐ **New Workbook**: A radio button that is currently unselected.
- Residuals:**
 - ☐ **Residuals**: A checkbox that is currently unchecked.
 - ☐ **Standardized Residuals**: A checkbox that is currently unchecked.
 - ☐ **Residual Plots**: A checkbox that is currently unchecked.
 - ☐ **Line Fit Plots**: A checkbox that is currently unchecked.
- Normal Probability:**
 - ☐ **Normal Probability Plots**: A checkbox that is currently unchecked.

On the right side of the dialog box, there are three buttons: 'OK' (highlighted with a blue border), 'Cancel', and 'Help'.

MULTIPLE LINEAR REGRESSION

■ Excel's Output:

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	<i>Regression Statistics</i>								
4	Multiple R	0.77943147							
5	R Square	0.60751341							
6	Adjusted R Squ	0.60750346							
7	Standard Error	1.52233845							
8	Observations	197103							
9									
10	ANOVA								
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
12	Regression	5	707022.924	141404.585	61015.6245	0			
13	Residual	197097	456775.124	2.31751434					
14	Total	197102	1163798.05						
15									
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
17	Intercept	1.26486478	0.0319418	39.5990405	0	1.20225961	1.32746995	1.20225961	1.32746995
18	Area	-0.92157055	0.02781653	-33.1303142	5.001E-240	-0.97609029	-0.8670508	-0.97609029	-0.8670508
19	# of riders	0.03822414	0.00552458	6.91892132	4.5648E-12	0.0273961	0.04905219	0.0273961	0.04905219
20	High tipper	17.2675481	0.10719446	161.086202	0	17.0574495	17.4776466	17.0574495	17.4776466
21	New year day	0.02886057	0.00874959	3.29850399	0.00097219	0.01171158	0.04600957	0.01171158	0.04600957
22	Pre-tip amount	0.14956192	0.00040026	373.660081	0	0.14877742	0.15034643	0.14877742	0.15034643

MULTIPLE LINEAR REGRESSION

- The estimated multiple regression equation:

$$\hat{Y} = 1.2649 - 0.9216X_1 + 0.0382X_2 + 17.2675X_3 + 0.0288X_4 + 0.1496X_5$$

where Y = Taxi tips in NYC in \$

X_1 = Area indicator (New York City = 1, John F Kennedy Airport = 0)

X_2 = Number of riders

X_3 = High tipper indicator (High=1, Normal = 0)

X_4 = New Year's Day indicator (Jan 1st = 1, Others = 0)

X_5 = Pre-tip amount in \$

INTERPRETATION OF ESTIMATES

- The estimated slope coefficient
 - $b_1 = -0.9216$ says that the estimated average tips decrease by \$0.9216 when the trip area switches from JFK to NYC, and given that other independent variables remain constant
 - $b_2 = 0.0382$ says that the estimated average tips increase by \$0.0382 for each additional rider, and given that other independent variables remain constant
 - $b_3 = 17.2675$ says that the estimated average tips increase by \$17.2675 if the rider is categorized as a high tipper, and given that other independent variables remain constant
 - $b_4 = 0.0288$ says that the estimated average tips increase by \$0.0288 if it is on New Year's Day, and given that other independent variables remain constant
 - $b_5 = 0.1496$ says that the estimated average tips increase by \$0.1496 for each \$1 increase in pre-tip taxi fare, and given that other independent variables remain constant

COMPARISON OF MODELS

- Suppose we run another linear regression model only used pre-tip taxi fare and # of riders as independent variables

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	<i>Regression Statistics</i>								
4	Multiple R	0.743946252							
5	R Square	0.553456026							
6	Adjusted R Square	0.553451494							
7	Standard Error	1.623781604							
8	Observations	197103							
9									
10	ANOVA								
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
12	Regression	2	644111.0423	322055.521	122144.95	0			
13	Residual	197100	519687.0058	2.6366667					
14	Total	197102	1163798.048						
15									
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
17	Intercept	0.270919756	0.009052735	29.9268419	2.434E-196	0.25317661	0.2886629	0.25317661	0.2886629
18	# of riders	0.046357224	0.005855885	7.91634823	2.4585E-15	0.03487983	0.05783462	0.03487983	0.05783462
19	Pre-tip amount	0.157555449	0.000321221	490.489434	0	0.15692586	0.15818503	0.15692586	0.15818503
20									

EVALUATE THE MODEL

- r^2 and adjusted r^2
- F-test for overall model significance
- t-test for a particular X -variable significance

MEASURES OF VARIATION -- r^2

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

$$SST = SSR + SSE$$

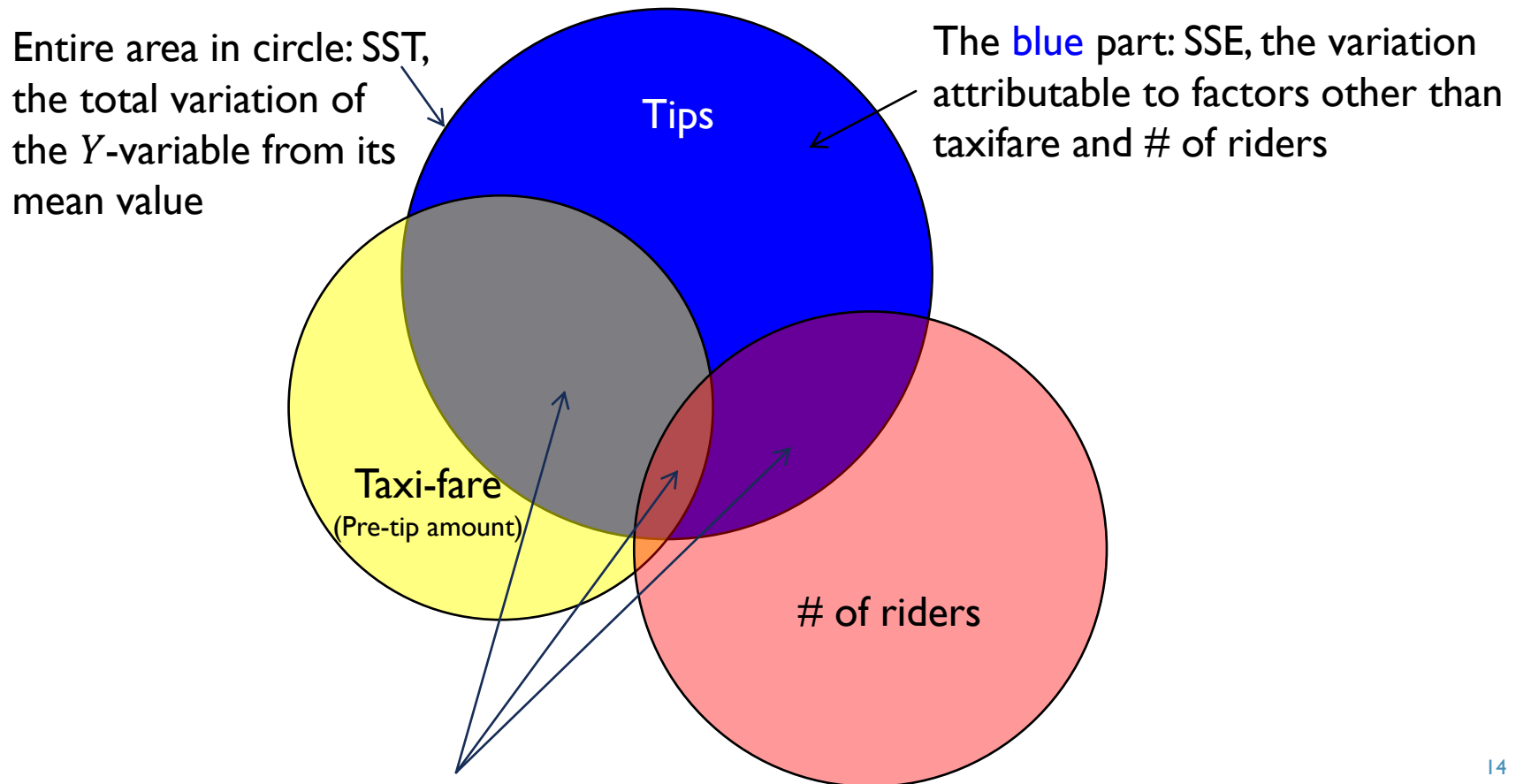
where

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

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

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

MEASURES OF VARIATION -- r^2



The grey, orange and purple parts: SSR, the total variation of Y -variable that being explained by the regression equation with independent variables

MEASURES OF VARIATION -- r^2

- What is the net effect of adding a new X -variable?
 - r^2 increases, even if the new X -variable is explaining an insignificant proportion of the variation of the Y -variable
 - Is it fair to use r^2 for comparing models with different number of X -variables?
 - A degree of freedom* will be lost, as a slope coefficient has to be estimated for that new X -variable
 - Did the new X -variable add enough explanatory power to offset the loss of one degree of freedom?

*Degrees of freedom: Number of independent pieces of information (data values) in the random sample.
If p parameters (intercept, slopes) must be estimated before the sum of squares errors, SSE, can be calculated from a sample of size n , the degrees of freedom are equal to $n - p$ ($= n - K - 1$ for multiple linear regression with $K+1$ coefficients of b_0, b_1, \dots, b_K).

MEASURES OF VARIATION – ADJUSTED r^2

(Recall: $r^2 = 1 - \frac{SSE}{SST}$)

- **Adjusted r^2** $= 1 - \frac{SSE/(n-K-1)}{SST/(n-1)} = 1 + \frac{(n-1)}{(n-K-1)} (r^2 - 1)$
- Measures the proportion of variation of the Y_i values that is explained by the regression equation with the independent variable X_1, X_2, \dots, X_K , **after the adjustment of** sample size (n) and the number of X -variables used (K)
- Smaller than or equal to r^2 , and can be negative
- **Penalize** the excessive use of X -variables
- Useful in **comparing among models** with different number of X -variables

EXAMPLE

- Compare the model that only used pre-tip amount (worksheet “Fare-Tip”) against the model using 5 independent variables (pp. 8), which one fits better?
 - Number of Observations: 197,103 vs 197,103
 - Degree of freedom (n-K-1): 197,101 vs 197,097
 - r^2 : 0.5533 vs 0.6075
 - Adjusted r^2 : 0.5533 vs 0.6075

INFERENCE: OVERALL MODEL SIGNIFICANCE

- **F-test** for the overall model significance

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

(none of the X -variables affects Y)

$$H_1: \text{At least one } \beta_i \neq 0$$

(at least one X -variable affects Y)

$$F = \frac{MSR}{MSE} = \frac{SSR(All)/K}{SSE(All)/(n-K-1)} \quad \text{with } K, (n - K - 1) \text{ degrees of freedom (d.f.)}$$

1. Rejection region approach

$$\text{Reject } H_0 \text{ if } F > \text{C. V.} = F_{\alpha, K, (n-K-1)}$$

or

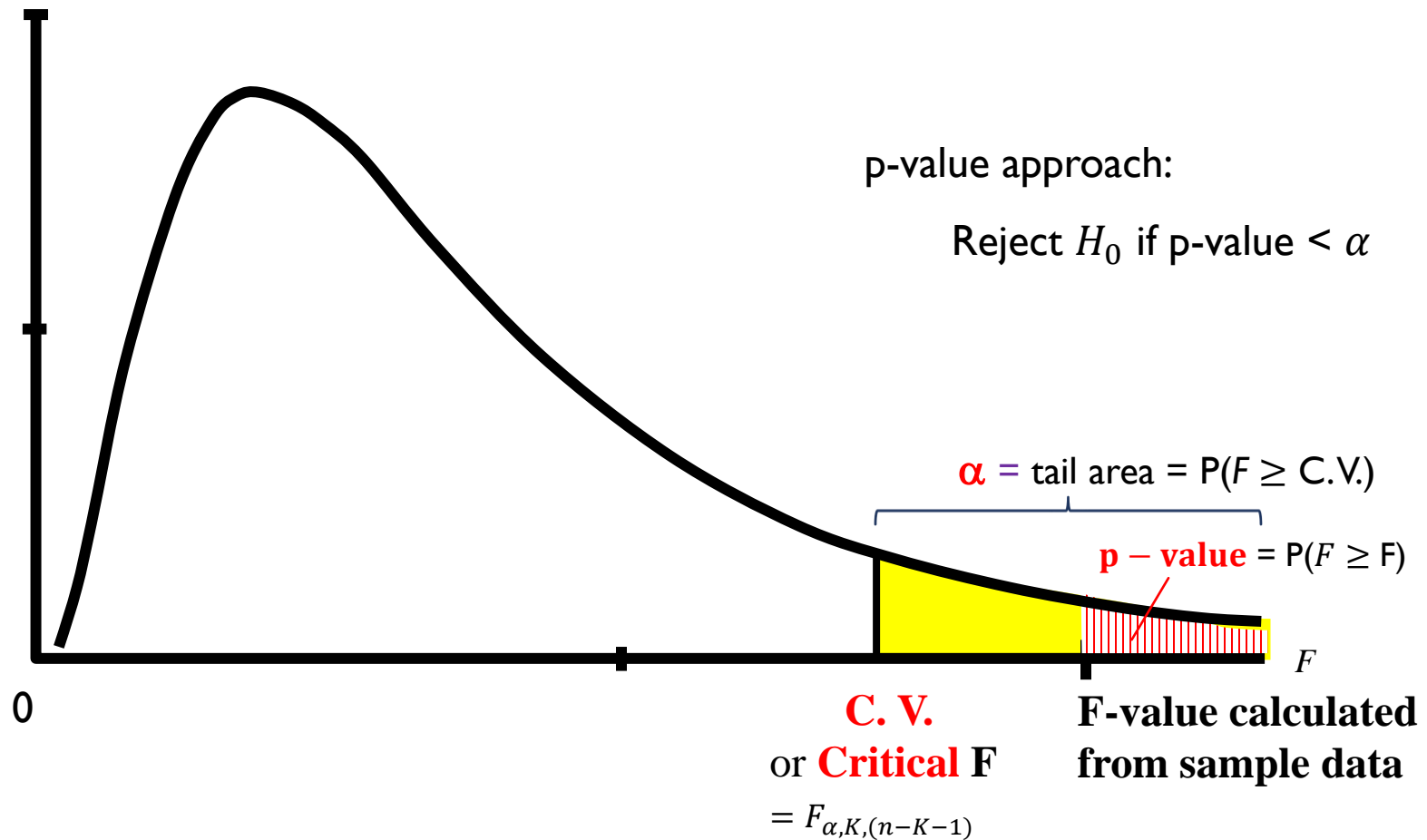
2. p-value approach

$$\text{p-value} = P(F \geq F)$$

$$\text{Reject } H_0 \text{ if p-value} < \alpha$$

INFERENCE: OVERALL MODEL SIGNIFICANCE

Probability distribution of F



INFERENCE: A PARTICULAR X-VARIABLE SIGNIFICANCE

- By rejecting the H_0 in F-test, we still cannot distinguish **which X-variable(s)** has the significant impacts on the Y-variable
- t-test** for a particular X-variable significance

$H_0: \beta_i = 0$ (X_i has no linear relationship with Y, given presence of other X-variable(s))

$H_1: \beta_i \neq 0$ (X_i is linearly related to Y, given presence of other X-variable(s))

$t = \frac{b_i - \beta_i}{s_{b_i}}$ with $(n - K - 1)$ degrees of freedom (d.f.)

1. Rejection region approach

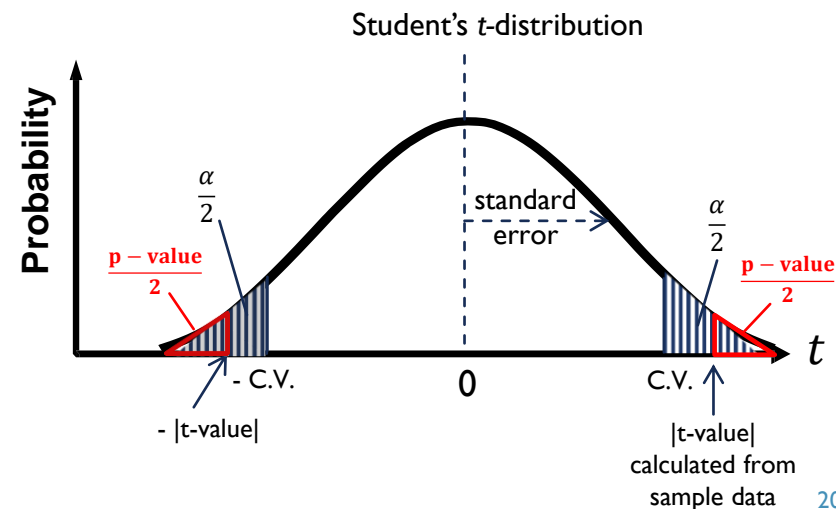
Reject H_0 if $|t| > \text{C.V.} = t_{\alpha/2, (n-2)}$

or

2. p-value approach

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

Reject H_0 if p-value $< \alpha$



EXAMPLE

- For the model using 5 independent variables, is the overall model significant?

10	ANOVA					p-value
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
12	Regression	5	707022.924	141404.585	61015.6245	0
13	Residual	197097	456775.124	2.31751434		
14	Total	197102	1163798.05			
15						

- $F = 61015.62$, p-value (*Significance F*) ≈ 0 ;
- At 5% significance level, p-value $\approx 0 < 5\%$. Therefore H_0 is rejected.

EXAMPLE

- At 5% level of significance, which X-variable(s), significantly affecting Y?

		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
16					
17	Intercept	1.26486478	0.0319418	39.5990405	0
18	Area	-0.92157055	0.02781653	-33.1303142	*5.001E-240
19	# of riders	0.03822414	0.00552458	6.91892132	*4.5648E-12
20	High tipper	17.2675481	0.10719446	161.086202	0
21	New year day	0.02886057	0.00874959	3.29850399	0.00097219
22	Pre-tip amoun	0.14956192	0.00040026	373.660081	0

- According to the t-test results, the p-value for each of the five independent variables is smaller than 5%, indicating each of them is significantly related to tips paid in NYC, given presence of other X-variable(s)).

*Scientific notation: $5.001\text{E-}240 = 5.001 \times 10^{-240} \approx 0$; $4.5648\text{E-}12 = 4.5648 \times 10^{-12} \approx 0$

VARIABLES SELECTION STRATEGIES

- In case some of the independent variables are insignificant based on t-test results, one may consider eliminating them using the following methods
 - All possible regressions
 - Backward elimination
 - Forward selection
 - Stepwise regression

ALL POSSIBLE REGRESSIONS

- To develop all the possible regression models between the dependent variable and all possible combinations of independent variables
- If there are K X -variables to consider using, there are $(2^K - 1)$ possible regression models to be developed
- The criteria for selecting the best model may include
 - Mean Sum of Squares Errors (MSE)
 - Adjusted r^2
- Disadvantages of all possible regressions
 - No unique conclusion, with different criteria, different conclusions will arise
 - Look at overall model performance, but not individual variable significance
 - When there is a large number of potential X -variables, computational time can be long

BACKWARD ELIMINATION

- Evaluate individual variable significance

Step 1: Build a model by using all potential X -variables

Step 2: Identify the least significant X -variable using t-test

Step 3: Remove this X -variable if its p-value is larger than the specified level of significance; otherwise terminate the procedure

Step 4: Develop a new regression model after removing this X -variable, repeat steps 2 and 3 until all remaining X -variables are significant

FORWARD SELECTION

- Evaluate individual variable significance

Step 1: Start with a model which only contains the intercept term

Step 2: Identify the most significant X -variable using t-test

Step 3: Add this X -variable if its p-value is smaller than the specified level of significance; otherwise terminate the procedure

Step 4: Develop a new regression model after including this X -variable, repeat steps 2 and 3 until all significant X -variables are entered

STEPWISE REGRESSION

- Evaluate individual variable significance
- An X -variable entering can later leave; an X -variable eliminated can later go back in

Step 1: Start with a model which only contains the intercept term

Step 2: Identify the most significant X -variable, add this X -variable if its p -value is smaller than the specified level of significance; otherwise terminate the procedure

Step 3: Identify the least significant X -variable from the model, remove this X -variable if its p -value is larger than the specified level of significance

Step 4: Repeat steps 2 and 3 until all significant X -variables are entered and none of them have to be removed

PRINCIPLE OF MODEL BUILDING

- A good model should
 - Have few independent variables
 - Have high predictive power
 - Have low correlation between independent variables
 - Be easy to interpret