ANOVA

CIVL 7012/8012

ANOVA

- ***ANOVA = Analysis of Variance***
- *A statistical method used to compare means among various datasets (2 or more samples)*
- *Can provide summary of any regression analysis in a table called* ***ANOVA Table***
- *Developed by statistician and evolutionary biologist* ***Ronald Fisher*** *in 1921*

ANOVA Table

- *Basic Information contains Estimates of Variance*
- *Estimates used to answer Inferential questions of regression analysis*
- *Analysis of Variance and regression analysis are closely related*
- *Usually employed in comparisons involving several population means*

Why the name, "ANOVA"

- *Why Not ANOME, where ME=Means*
- *Although means are compared, but Comparisons are made using estimates of variances*
- *The ANOVA test statistics used are actually ratios of estimates of variance*

ANOVA vs. REGRESSION

- *Independent Variables*
  - ***ANOVA:*** *must be treated as nominal*
  - ***REGRESSION:*** *can be of any mixture (nominal, ordinal, interval)*
- *ANOVA is a special case of regression analysis*
- *For multivariable analysis or regression, the technique is called Analysis of Covariance (ANACOVA)*

FACTORS AND LEVELS

- Assume a nominal (categorical) variable with *k* categories:
  - *Then number of dummy variables = k – 1*
- These (k – 1) variables collectively describe the *basic* nominal variable
- The basic nominal variable is called **FACTOR**
- The different categories of the FACTOR are referred to as its LEVELS

FIXED AND RANDOM FACTORS

- RANDOM FACTOR
  - Whose LEVELs may be regarded as a sample from some large population of levels
  - Example, Subjects, Litters, Observers, Days, Weeks
- FIXED FACTOR
  - Whose LEVELs are the only ones of interest
  - Example, Gender, Age, Marital Status, Education
- BOTH: locations, treatments, drugs, exposures

Types of ANOVA

- Several types depending on experimental designs and situations for which they have been developed
  - One way (one factor, fixed effects)
  - Two way (two factors, random effects)
  - Two way with repeated measures (two factors, random effects)
  - Fully nested (hierarchical factors)
  - Kruskal-Wallis (non-parametric one way)
  - Friedman (non-parametric two way)

THE PROBLEM (One Way ANOVA)

- To Determine whether the population means are all equal or not.
- Given *k* means (denoted as $\mu_1, \mu_{2,\ldots,}\ \mu_k$), the basic null hypothesis of interest is:
  - $H_0\colon \mu_1 = \mu_2 = \cdots = \mu_k$
- The Alternate hypothesis is given by:
  - $H_A\colon \text{"}The\ k\ population\ means\ are\ not\ all\ equal\text{"}$

Assumptions (One Way ANOVA)

- All populations involved follow normal distribution
- Variance of the dependent variable is the same in each population
- Random samples have been selected from each populations or groups
- Each experimental unit sampled has been recorded with a specified dependent variable value

ANOVA Table

Source	Degrees of freedom (*df*)	Sum of Squares (SS)	Mean Square (MS)	F-value/F
Between groups/ Treatment groups/Model	$k-1$	SSE	$MSE = \frac{SSE}{k-1}$	$\frac{MSE}{MSR}$
Within Groups/Error	$N-k$	SSR	$MSR = \frac{SSR}{N-k}$	
Total	$N-1$	SST		

$k = number\ of\ population\ means$

$N = Total\ number\ of\ observations$

$SSE = Sum\ of\ squares\ between\ groups\ (Explained)$

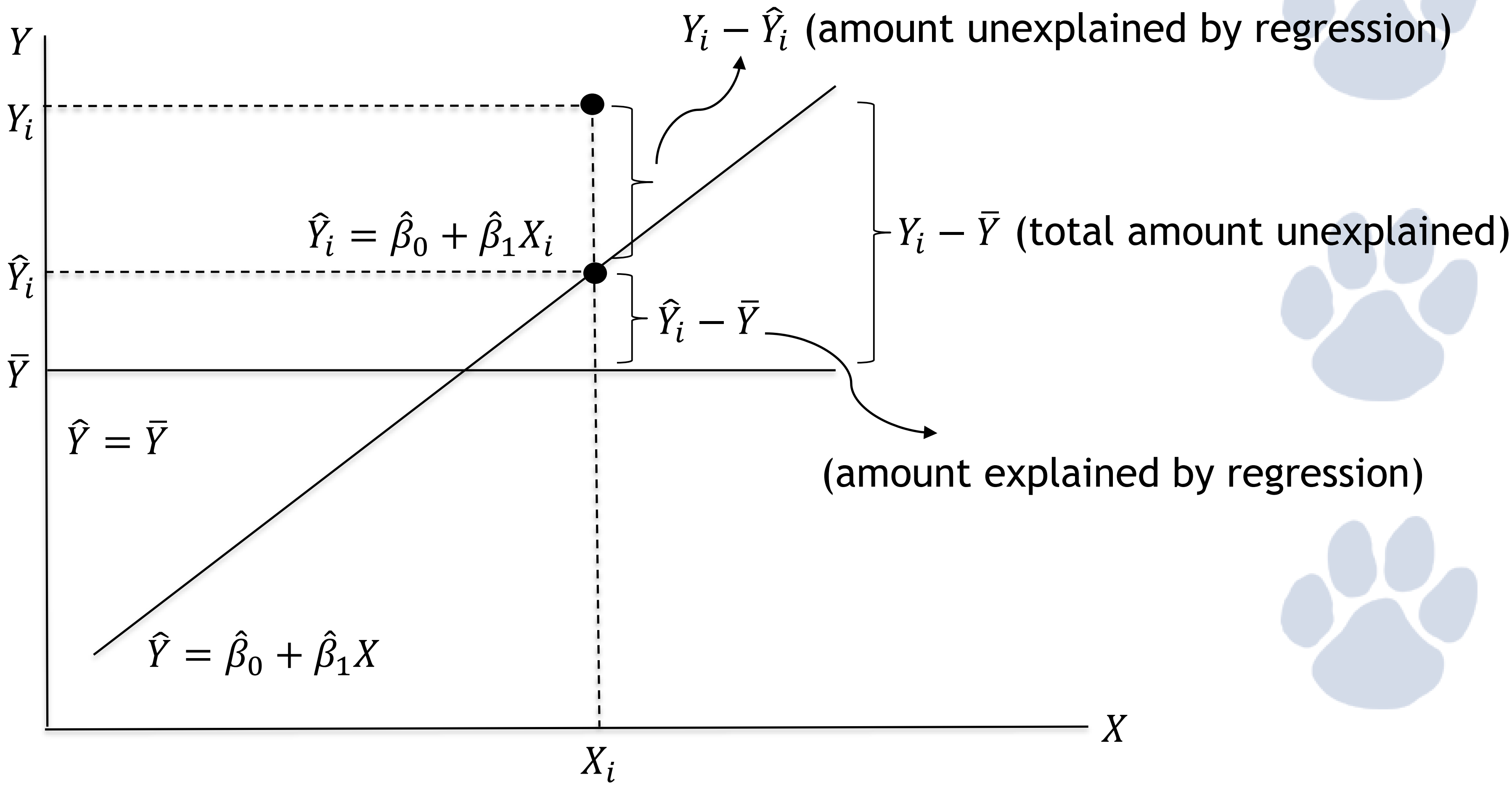
$SSR = Sum\ of\ squares\ within\ groups/Residual\ sum\ of\ squares/Error\ sum\ of\ Squares$

$SST = Total\ sum\ of\ squares$

$MST = Mean\ square\ Treatment/Mean\ Square\ between\ groups$

$MSE = Mean\ square\ Error$

Partition of Variance



Partition of Variance (Cont.)

Total unexplained variation	=	*Variation due to regression*	+	*Unexplained residual variation*
Variation in all observations	=	*Variation between each observation and its group mean*	+	*Variation between each group mean and the overall mean*

In other words,

$$SST = SSE + SSR$$

OR,

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{k} n_i(Y_i - \bar{Y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$

F Statistics

- For a one-way ANOVA, the test statistic is equal to the ratio of MST and MSE
- This ratio is known to follow an *F distribution*
- *The test statistics is calculated as,* $F = \frac{MSE}{MSR}$
- *If F (calculated) > F (Critical)*
  - *Reject Null hypothesis*
- *If F (calculated) ≤ F (Critical)*
  - *Fail to reject Null hypothesis*

F Distribution

- F distribution table is used to find the critical value
- Required:
  - Degrees of freedom of Numerator (MSE)
  - Degrees of freedom of Denominator (MSR)
  - Value of alpha (0.05, 0.1, …)
- See supplemental table on the website

EXAMPLE

- Suppose the National Transportation Safety Board (NTSB) wants to examine the safety of compact cars and full-size cars. It collects a sample of three for each of the treatments (cars types). Using the hypothetical data provided below, test whether the mean pressure applied to the driver's head during a crash test is equal for each types of car. Use α = 5%

Compact cars	Full size cars
643	484
655	456
702	402

EXAMPLE (Cont.)

- ***Step 1***

  State the null and alternate hypothesis

  - $H_0: \mu_1 = \mu_2$
  - $H_A: Atleast\ one\ mean\ pressure\ is\ not\ ststistically\ equal$
- ***Step 2***
  - Calculate the appropriate test statistic (Find sum of squares, mean squares) and critical value and then compare
- Example shown in Excel file (example_ANOVA.xlsx)

Example-1: Complete ANOVA Table

Source	SS	df	MS	F
Explained	18.9	3		
Error	72.0	16		
Total				

The Sum of Squares and Degrees of Freedom are given. Complete the table.

Example-1: Answer

Source	SS	df	MS	F
Explained	18.9	3	6.30	1.40
Error	72.0	16	4.50	
Total	90.9	19	4.78	

Example-2: Complete ANOVA Table

Source	SS	df	MS	F
Explained	106.6		21.32	2.60
Error		26		
Total				

Complete the table

Example-2: Solution

Source	SS	df	MS	F
Explained	106.6	5	21.32	2.60
Error	213.2	26	8.20	
Total	319.8	31	10.32	

Example-3

- N=20

Source	SS	df	MS	F
Explained	56.7			
Error		14	13.50	
Total				

Example-3: Solution

Source	SS	df	MS	F
Explained	56.7	5	11.34	0.84
Error	189.0	14	13.50	
Total	245.7	19	12.93	