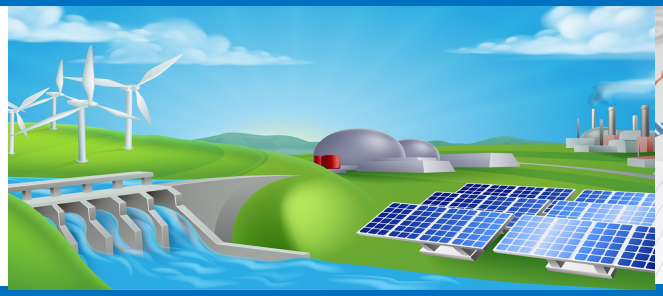



ENV872 - ENVIRONMENTAL DATA ANALYTICS

M6 – Generalized Linear Models (GLMs)

Luana Lima | John Fay

Master of Environmental Management Program
Nicholas School of the Environment - Duke University

Introduction

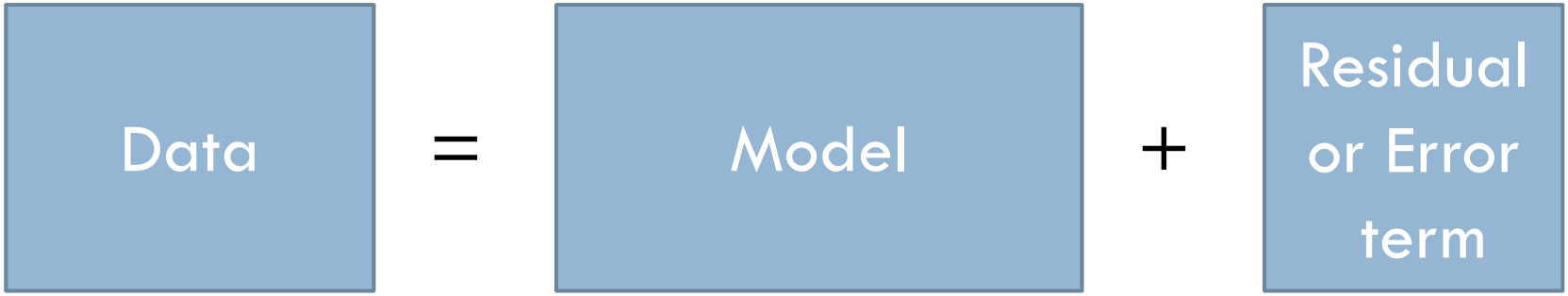
- In this module we will cover some introductory inferential statistics that uses sample data to answer research questions
  - estimating population parameters
  - estimate relationships
- We will use statistical model called Generalized Linear Models that include
  - T-test
  - Regression
  - ANOVA – Analysis of variance

Learning Goals

- Describe the components of the generalized linear model (GLM)
- Apply special cases of the GLM (linear regression) to real datasets
- Interpret and report the results of linear regressions
- Apply special cases of the GLM (ANOVA) to real datasets
- Interpret and report the results of ANOVA

Definition

The Generalized Linear Model (GLM) is a framework for comparing how variables affect different continuous variables.



Continuous variable Y

One or more variables $X_1, X_2, \ldots, X_n$

X could be continuous or categorical

GLM is based on the assumption that the data residuals approximate a normal distribution

Special cases of GLMs

- Analysis of variance
  - One-way ANOVA: Continuous response, **one categorical explanatory variable with more than two categories**
  - Two-way ANOVA: Continuous response, **two categorical explanatory variables**
- Linear Regression
  - Simple Linear Regression: Continuous response, **one continuous explanatory variable**
  - Multiple Linear Regression: Continuous response, **two or more continuous explanatory variables**

Special cases of GLMs – T-test

- One-sample t-test: continuous response, test the null hypothesis that the mean of the group is equal to a specific value

$$H0: \mu = \beta_0$$
$$H1: \mu \neq \beta_0$$

In R: **t.test**(EPAair$Ozone, mu = 50)

Under GLM: $Ozone = \beta_0 + \epsilon$

In R: **lm**(EPAair$Ozone ~ 1, EPAair)

Special cases of GLMs – T-test

- Two-sample t-test: continuous response, test the hypothesis that the mean of two samples is equivalent.
  - assumption that the variance of the two groups (response and explanatory) is equivalent

$$H0: \mu_{2018} = \mu_{2019}$$
$$H1: \mu_{2018} \neq \mu_{2019}$$

In R: **t.test**(EPAair$Ozone ~ EPAair$Ozone, var.equal = TRUE)

Under GLM: $Ozone = \beta_0 + \beta_1 * Year + \epsilon$

In R: lm(EPAair$Ozone ~ 1 + EPAair$Ozone)

Review: Hypothesis Testing

- Why do we use hypothesis testing?
  - To analyze evidence provided by data
  - To make decisions based on data
- What is a statistical hypothesis?
  - An assumption about a population parameter that may or may not be true
- In Hypothesis Testing we usually have

$$\begin{cases} H_0: & the\ null\ hypothesis \\ H_1: & the\ alternative\ hypothesis \end{cases}$$

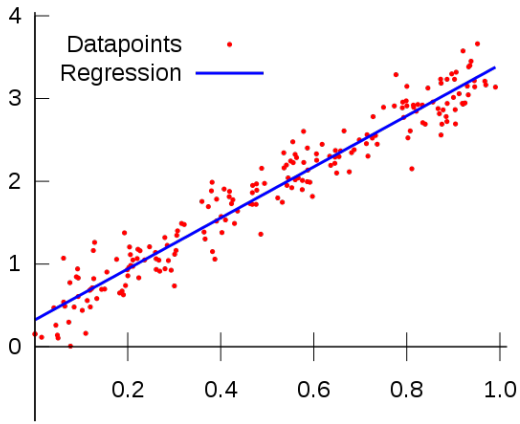
Review: Hypothesis Testing (cont'd)

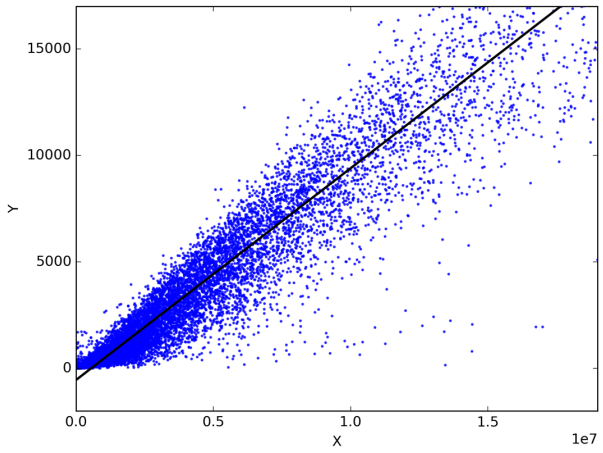
- Procedure
  1. State the hypotheses and identify the claim
  2. Find the critical value(s) from the appropriate table
  3. Compute the test value
  4. Make the decision to reject or not reject the null hypothesis

If P-value $\leq \alpha$, **reject** the null hypothesis.
If P-value $> \alpha$, **do not reject** the null hypothesis.

Simple Linear Regression

- **Regression** - a technique for fitting a line to a set of data points
  - **Simple linear regression** - the simplest form of regression that **involves a linear relationship between two variables**
  - The object of simple linear regression is to obtain an **equation of a straight line that minimizes the sum of squared vertical deviations from the line** (i.e., the *least squares criterion*)





Standard Error

- Standard error of estimate
  - **A measure of the scatter of points around a regression line**
  - **If the standard error is relatively small**, the **predictions using the linear equation will tend to be more accurate** than if the standard error is larger

$$S_e = \sqrt{\frac{\sum (y - y_c)^2}{n - 2}}$$

where

S_e = standard error of estimate

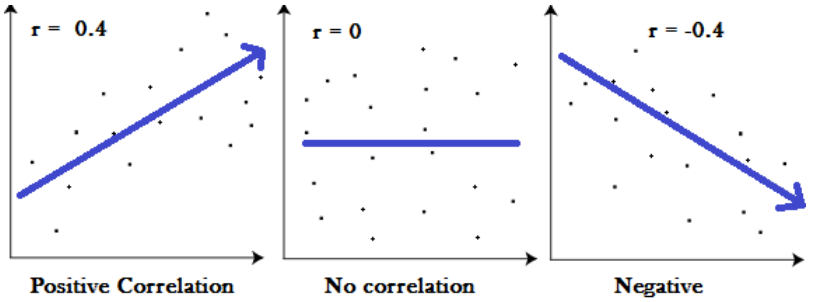
y = y value of each data point

n = number of data points

Correlation Coefficient

- **Correlation, *r***
  - **A measure of the strength and direction of relationship between two variables**
  - Ranges between -1.00 and +1.00

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2] * [n(\sum y^2) - (\sum y)^2]}}$$



- **r^2, square of the correlation coefficient**
  - **A measure of the percentage of variability in the values of *y* that is "explained" by the independent variable**
  - Ranges between 0 and 1.00

Residuals

- After fitting a regression model, check the residual plots first to be sure that you have unbiased estimates

P-values and coefficients in regression analysis

- The p-values for the coefficients indicate whether these relationships are statistically significant
  - i.e. determine whether the relationships that you observe in your sample also exist in the larger population

- The p-value for each independent variable tests the null hypothesis:
  - H_0: *independent variable has* ***no correlation*** *with the dependent variable*

- If the p-value for a variable is less than your significance level, your sample data provide enough evidence to **reject the null hypothesis for the entire population**
  - P-value < alpha, there is significant correlation with dependent variable
  - Addition to model is worthwhile

One-Way Analysis of Variance

It is a common situation when someone needs to determine whether three or more populations have equal means. To answer that question, we need to conduct a formal hypothesis test.

$$H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \ldots\ldots \mu_k$$

$$H_A : \text{ not all } \mu_j \text{ are equal}$$

The statistical method for conducting this test is called **Analysis of Variance - ANOVA**

One-Way Analysis of Variance - Layout

- Single Factor
- 4 Levels (populations)
- Equal Sample Size for each level – Balanced Design

		Factor 1		
Observation	Level 1	Level 2	Level 3	Level 4
1	x_{11}	x_{21}	x_{31}	x_{41}
2	x_{12}	x_{22}	x_{32}	x_{42}
3	.	.	.	.
4	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	x_{1n}	x_{2n}	x_{3n}	x_{4n}

One-Way Analysis of Variance - Example

A sandwich shop is interested in testing to see whether the mean value for customer orders is equal at the company's four store locations. Eight orders from each store are randomly selected

Single Factor - Customer Order Value

Four Levels - The Four Store Locations

Sample Size – 8 at each store

One-Way Analysis of Variance - Example

	Store Locations			
Customer	1	2	3	4
1	$4.10	$6.90	$4.60	$12.50
2	5.90	9.10	11.40	7.50
3	10.45	13.00	6.15	6.25
4	11.55	7.90	7.85	8.75
5	5.25	9.10	4.30	11.15
6	7.75	13.40	8.70	10.25
7	4.78	7.60	10.20	6.40
8	6.22	5.00	10.80	9.20

$$H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_A : \text{ not all } \mu_j \text{ are equal}$$

Introduction to One-Way ANOVA

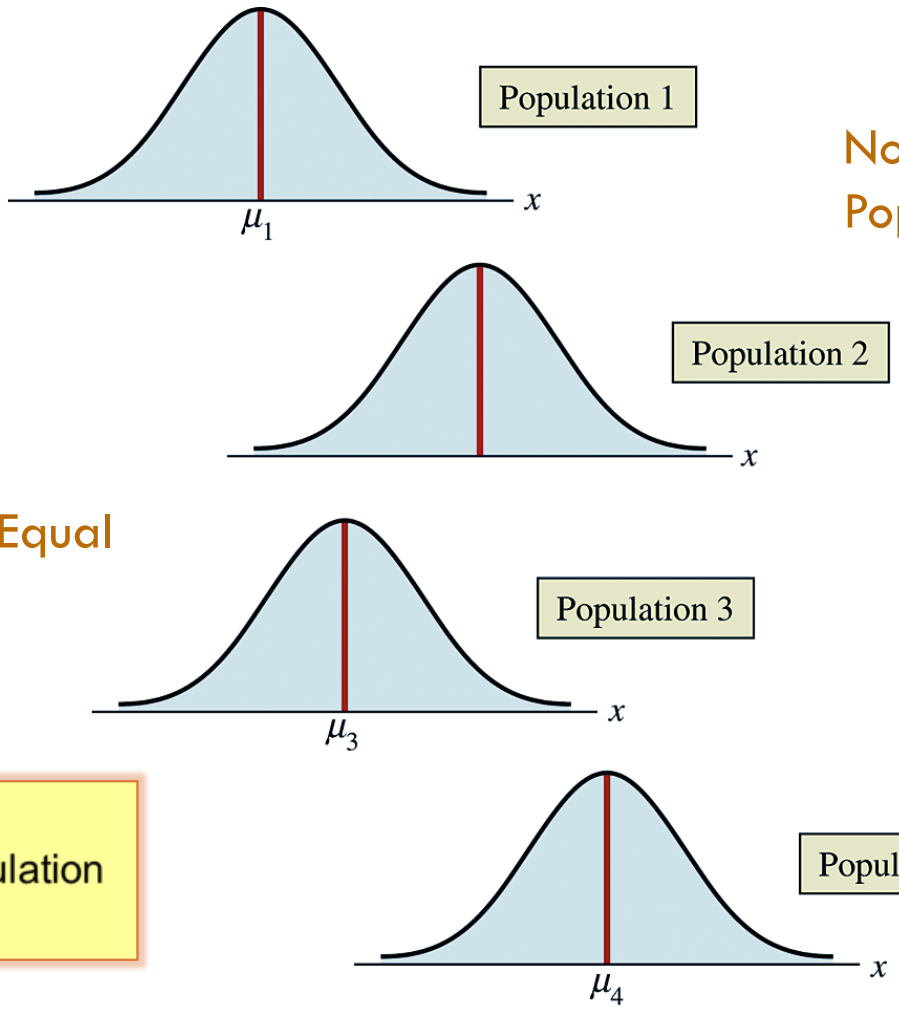
- Factor: A quantity under examination in an experiment as a possible cause of variation in the response variable
- Levels: The categories, measurements, or strata of a factor of interest in the current experiment – Also called **Populations**
- Balanced Design: An experiment has a balanced design if the factor levels have equal sample sizes

One-Way ANOVA Assumptions

- All **populations are normally distributed**
- The population **variances are equal**
- The **observations are independent** - that is, the occurrence of any one individual value does not affect the probability that any other observation will occur

One-Way ANOVA Assumptions

Null Hypothesis is False – **not all μ_i are equal**



Normally Distributed Populations

Populations have Equal Variances

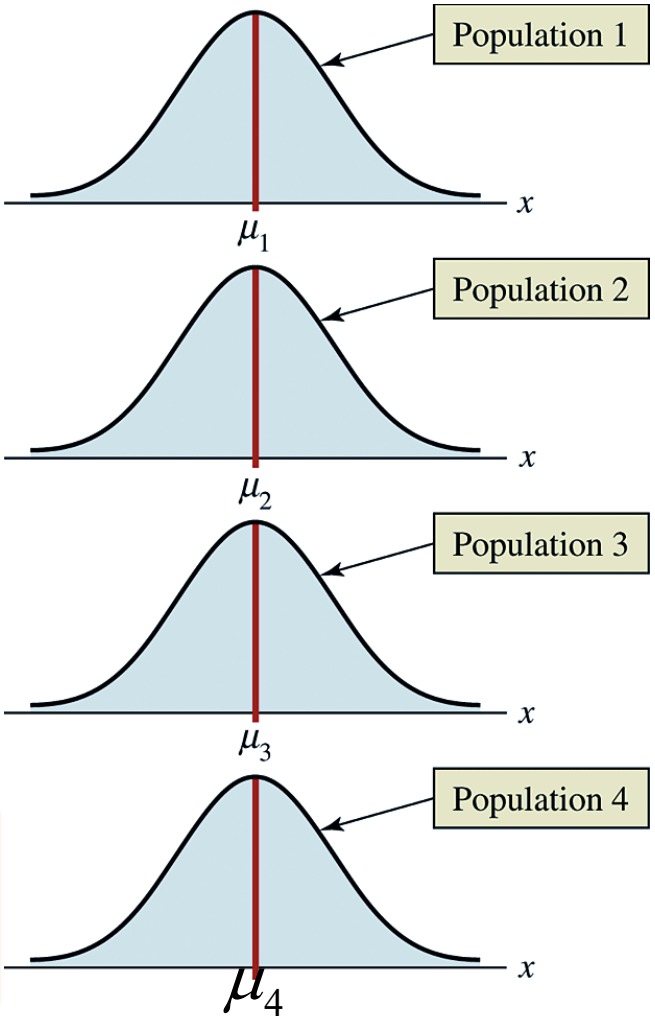
$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

H_A: At least two of the population means are different

One-Way ANOVA Assumptions

Null Hypothesis is True – **all μ_i are equal**

Populations have Equal Variances



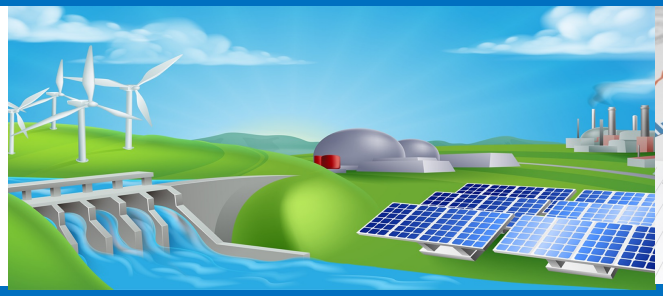
Normally Distributed Populations

$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

H_A: At least two of the population means are different

Post Hoc Test with ANOVA

- When you use ANOVA to test the equality of at least three group means, statistically significant results indicate that not all of the group means are equal
- However, ANOVA results do not identify which particular differences between pairs of means are significant.
- Use post hoc tests to explore differences between multiple group means
- Tukey HSD (Tukey Honest Significant Differences) method is the most common for comparing all possible group pairings.




THANK YOU !

luana.marangon.lima@duke.edu

Master of Environmental Management Program
Nicholas School of the Environment - Duke University