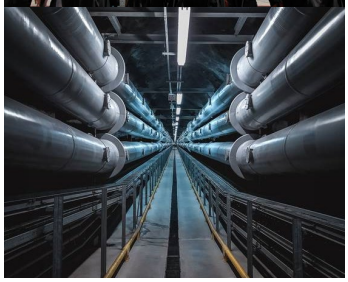

Unit 9: Linear Regression Modeling

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.

- To introduce the concept of **fitting linear regression using log-transformations** we will examine the relationship between brain and body weights of animal species.

- To introduce the concept of **multiple linear regression model of a numerical response variable (and many other explanatory variables)** we will examine the relationship between the "prestige score" of a career and the "type of career", "education level of the career" and the "income level of the career".

- To introduce the concept of **fitting a multiple linear regression model which has explanatory variables that are categorical** we will examine the relationship IQ score, age, and lead-level exposure.

Summary of Concepts:

- **Research questions we will be able to answer**
  - Is there a linear association between two numerical variables?
  - Is there a linear association between a numerical response variable and one or more explanatory variables?
- **Definitions**
  - Response vs. explanatory variables
  - Population vs. sample linear regression
- **Descriptive Analytics**
  - Visualizations for the relationship between two numerical variables
    - Scatterplot
    - What are four things we should describe about the relationship between two numerical variables?
  - Summary statistics for the relationship between two numerical variables
    - Covariance
    - Correlation (R)
    - R^2
- **Modeling**
  - Fit a simple linear regression model

        - Between two numerical variables
    - Fit a multiple linear regression model
        - Between a numerical response variable and one or more explanatory variables (can be numerical or categorical)
    - Check the model conditions for linear regression:
        - Linearity condition
        - Constant variance of residuals condition
        - Normal residuals centered at 0 condition
        - Non-collinear explanatory variable condition (just for multiple linear regression)
    - Strategies to try when the linear regression model conditions are not met:
        - Try transforming some of the variables.
- **Predictive Analytics**
    - Use a linear regression model to make a prediction about a response variable value given explanatory variable value(s).
- **Inference**
    - Is there evidence to suggest there is a linear relationship between an explanatory variable and the response variable *in the population?*
        - Create a confidence interval for a population slope β_i
        - Conduct a hypothesis test for a population slope β_i.

SUMMARY RESEARCH QUESTIONS AND ANALYSES - *WITH RESPECT TO THE TYPES OF VARIABLES INVOLVED*

Type of Variables Involved	Research Questions you Can Ask	Analysis to Do:	Assumptions for Analysis:
Single Numerical Variable	**About the Sample:** * What is the mean or median of the sample?	Calculate sample mean or sample median.	
	About an Unknown Population: * What is a range of plausible values for the mean of this population?	Create a confidence interval for μ.	* Random sample * n<10% of population * n>30 or population distribution is normal.
	About an Unknown Population: * Is there sufficient evidence to suggest that the mean of this population is not equal to (some number)?	Conduct a hypothesis test: H0: μ = (some number) HA: μ ≠ (some number)	* Random sample * n<10% of population * n>30 or population distribution is normal.
Single Categorical Variable *with 2 levels : "success" and "failure"*	**About the Sample:** * What is the proportion of "successes" in the sample?	Calculate sample proportion of successes.	
	About an Unknown Population: * What is a range of plausible values for the proportion of successes in this population?	Create a confidence interval for p.	* Random sample * n<10% of population * np>=10 and n(1-p)>=10
	About an Unknown Population: * Is there sufficient evidence to suggest that the proportion of "successes" in this population is not equal to (some number)?	Conduct a hypothesis test: H0: p = (some number) HA: p ≠ (some number)	* Random sample * n<10% of population * np>=10 and n(1-p)>=10
A Numerical Variable and a Categorical Variable *with 2 levels: "success" and "failure"*	**About the Sample:** * What is the mean of the two samples? How far apart are they?	Calculate the mean of the two samples.	
	About an Unknown Population: * What is a range of plausible values for the difference between the mean of population 1 vs. the mean of population 2? (equivalent) * Is there evidence to suggest an association between the two categorical variables in the population?	Create a confidence interval for μ1-μ2. Is 0 in the confidence interval? If not, there is evidence to suggest there is there is an association.	* Both samples are random *n1<10% of population 1 and n2<10% of population 2 * n1>30 or population distribution1 is normal * n2>30 or population distribution2 is normal. * Samples are independent.
	About an Unknown Population: * Is there sufficient evidence to suggest that there is a difference between the mean of population 1 vs. the mean of population 2? (equivalent) * Is there evidence to suggest an association between the two categorical variables in the population?	Conduct a hypothesis test: H0: μ1-μ2 = 0 HA:μ1-μ2 ≠ 0	* Both samples are random *n1<10% of population 1 and n2<10% of population 2 * n1>30 or population distribution1 is normal * n2>30 or population distribution2 is normal. * Samples are independent.
A Categorical Variable and *with 2 levels: "success1" and "failure1"* **a Categorical Variable *with 2 levels: "success2" and "failure2"***	**About the Sample:** * What is the proportion of successes in each of the two samples? How far apart are they?	Calculate the proportion of successes of sample 1 and tthe proportion of successes of sample 2.	
	About an Unknown Population: * What is a range of plausible values for the difference between the proportion of successes of population 1 vs. the proportion of successes of population 2? (equivalent) * Is there evidence to suggest an association between the two categorical variables in the population?	Create a confidence interval for p1-p2. Is 0 in the confidence interval? If not, there is evidence to suggest there is there is an association.	* Both samples are random *n1<10% of population 1 and n2<10% of population 2 * n1p1>10 and n1(1-p1)>=10 * n2p2>10 and n2(1-p2)>=10 * Samples are independent.
	About an Unknown Population: * Is there sufficient evidence to suggest that there is a difference between the proportion of successes in population 1 vs. the the proportion of successes in of population 2? (equivalent) * Is there evidence to suggest an association between the two categorical variables in the population?	Conduct a hypothesis test: H0: p1-p2 = 0 HA:p1-p2 ≠ 0	* Both samples are random *n1<10% of population 1 and n2<10% of population 2 * n1p1>10 and n1(1-p1)>=10 * n2p2>10 and n2(1-p2)>=10 * Samples are independent.

THIS UNIT'S RESEARCH QUESTIONS AND ANALYSES - *WITH RESPECT TO THE TYPES OF VARIABLES INVOLVED*

Two Numerical Variables	**About the Sample:** * What is the relationship between two numerical variables in the sample?	<u>You can calculate the following between the numerical variables in the sample to answer this question:</u> * the covariance * the correlation * R^2	
Two Numerical Variables *(One is the response variable and one is the explanatory variable).*	**About the Sample:** * What is the relationship between two numerical variables in the sample?	<u>You can calculate the following between the numerical variables in the sample to answer this question:</u> * the covariance * the correlation (R) * R^2 <u>You can also fit a simple linear regression model between the two numerical variables.</u>	* There is a linear relationship between these two numerical variables.
	About a new observation: * What is the predicted the response variable value for a given explanatory variable value?	Use the simple linear regression model you fitted to make this prediction.	* There is a linear relationship between these two numerical variables.
	About the Population: * What is a plausible range of values for the slope of the best fit line between these two variables in the population? (equivalent) * Is there evidence to suggest a linear association between the two numerical variables in the population?	* Create a confidence interval for βi. * Is 0 in the confidence interval? If not, there is sufficient evidence to suggest a linear relationship between the two numerical variables in the population.	* There is a linear relationship between these two numerical variables.
	About the Population: * Is there evidence to suggest that slope of the best fit line between these two variables non-zero? (equivalent) * Is there evidence to suggest a linear association between the two numerical variables in the population?	Conduct a hypothesis test: H0: βi = 0 HA: βi. ≠ 0	* There is a linear relationship between these two numerical variables. * The variance of the residuals are the same.
Numerical Response Variable and Many Explanatory Variables	**About the Sample:** * What is the relationship between these variables in the sample, with this particular response variable?	Fit a multiple linear regression model between these variables.	
	About a new observation: * What is the predicted the response variable value for a given explanatory variable value?	Use the multiple linear regression model you fitted to make this prediction.	
	About the Population: * What is a plausible range of values for the slope of one of the explanatory variable predictors of the best fit line in the population? (equivalent) * Is there a linear association between this explanatory variable predictor in the population and the response variable?	* Create a confidence interval for βi. * Is 0 in the confidence interval? If not, there is sufficient evidence to suggest a linear relationship between the two numerical variables in the population.	
	About the Population: * Is there evidence to suggest that the slope of one of the explanatory variable predictors of the best fit line in the population is non-zero? (equivalent) * Is there a linear association between this explanatory variable predictor in the population and the response variable?	Conduct a hypothesis test: H0: βi = 0 HA: βi. ≠ 0	