Regression Analysis

Multiple Regression

[Cross-Sectional Data]

Learning Objectives

- Explain the linear multiple regression model [for cross-sectional data]
- Interpret linear multiple regression computer output
- Explain multicollinearity
- Describe the types of multiple regression models

Regression Modeling Steps

- Define problem or question
- Specify model
- Collect data
- Do descriptive data analysis
- Estimate unknown parameters
- Evaluate model
- Use model for prediction

Simple vs. Multiple

- β represents the unit change in Y per unit change in X .
- Does not take into account any other variable besides single independent variable.

- β_i represents the unit change in Y per unit change in X_i.
- Takes into account the effect of other β_i s.
- "Net regression coefficient."

Assumptions

- **Linearity -** the Y variable is linearly related to the value of the X variable.
- **Independence of Error** - the error (residual) is independent for each value of X.
- **Homoscedasticity** - the variation around the line of regression be constant for all values of X.
- **Normality** - the values of Y be normally distributed at each value of X.

Goal

Develop a statistical model that can predict the values of a *dependent* (response) variable based upon the values of the *independent* (explanatory) variables.

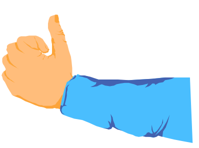
Simple Regression

A statistical model that utilizes <u>one</u> *quantitative* *independent* variable "X" to predict the *quantitative* *dependent* variable "Y."

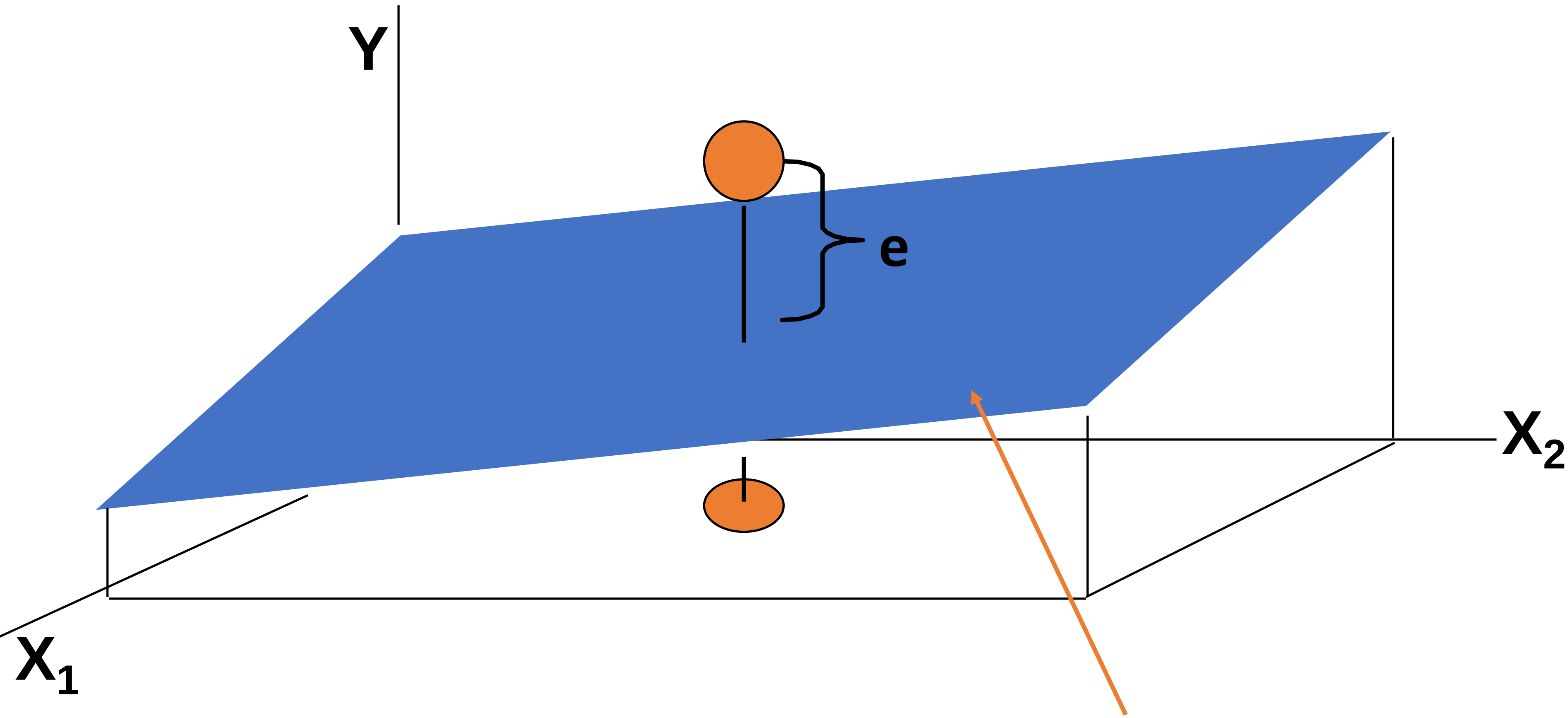
Multiple Regression

A statistical model that utilizes two or more *quantitative* and *qualitative* explanatory variables (x_1,..., x_p) to predict a *quantitative* dependent variable Y.

***Caution*: have at least two or more quantitative explanatory variables (rule of thumb)**



Multiple Regression Model



Hypotheses

- **H_0: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_P = 0$**
- **H_1: At least one regression coefficient is not equal to zero**

Hypotheses (alternate format)

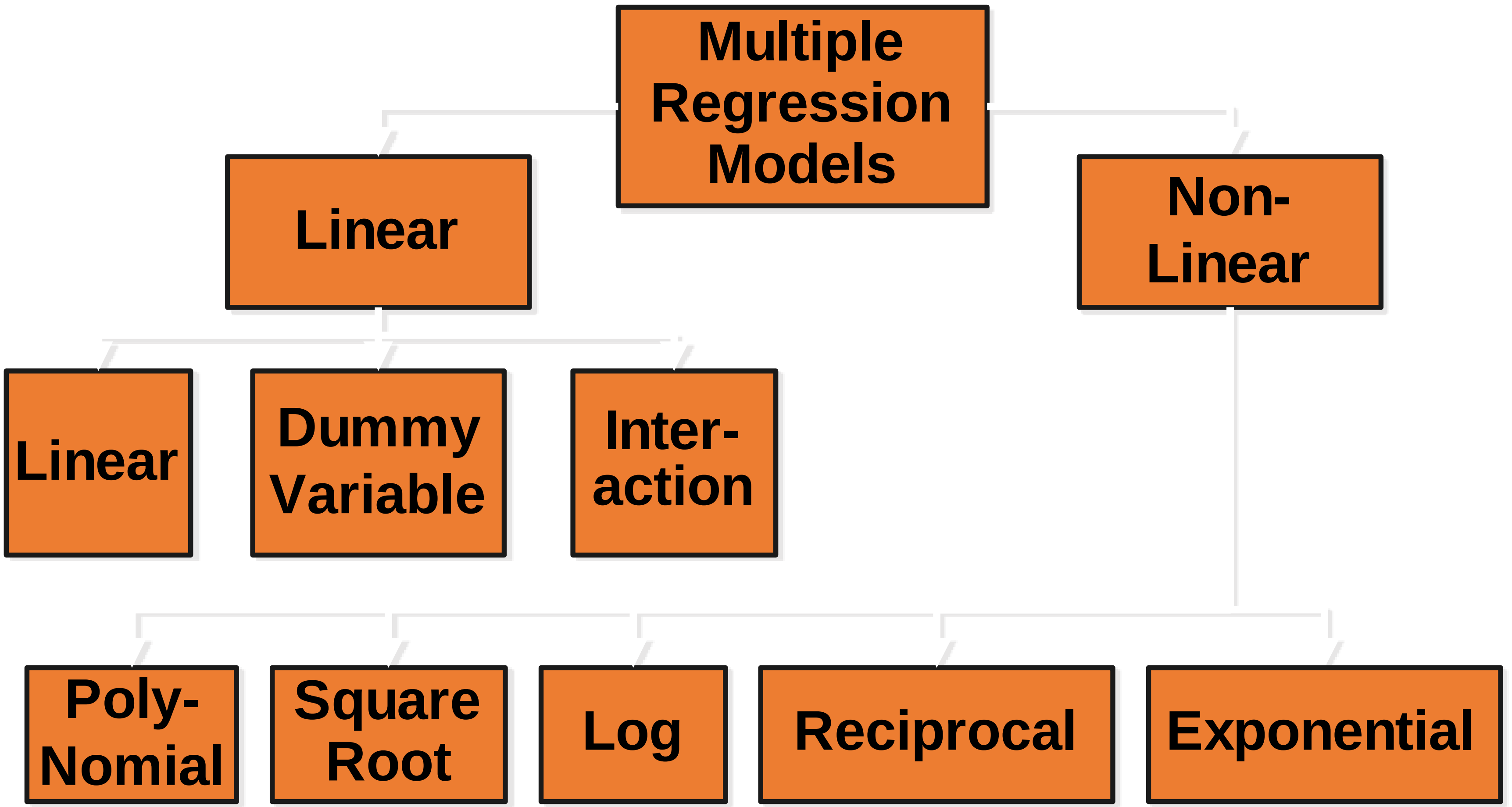
$$H_0: \ \beta_i = 0$$

$$H_1: \ \beta_i \neq 0$$

Types of Models

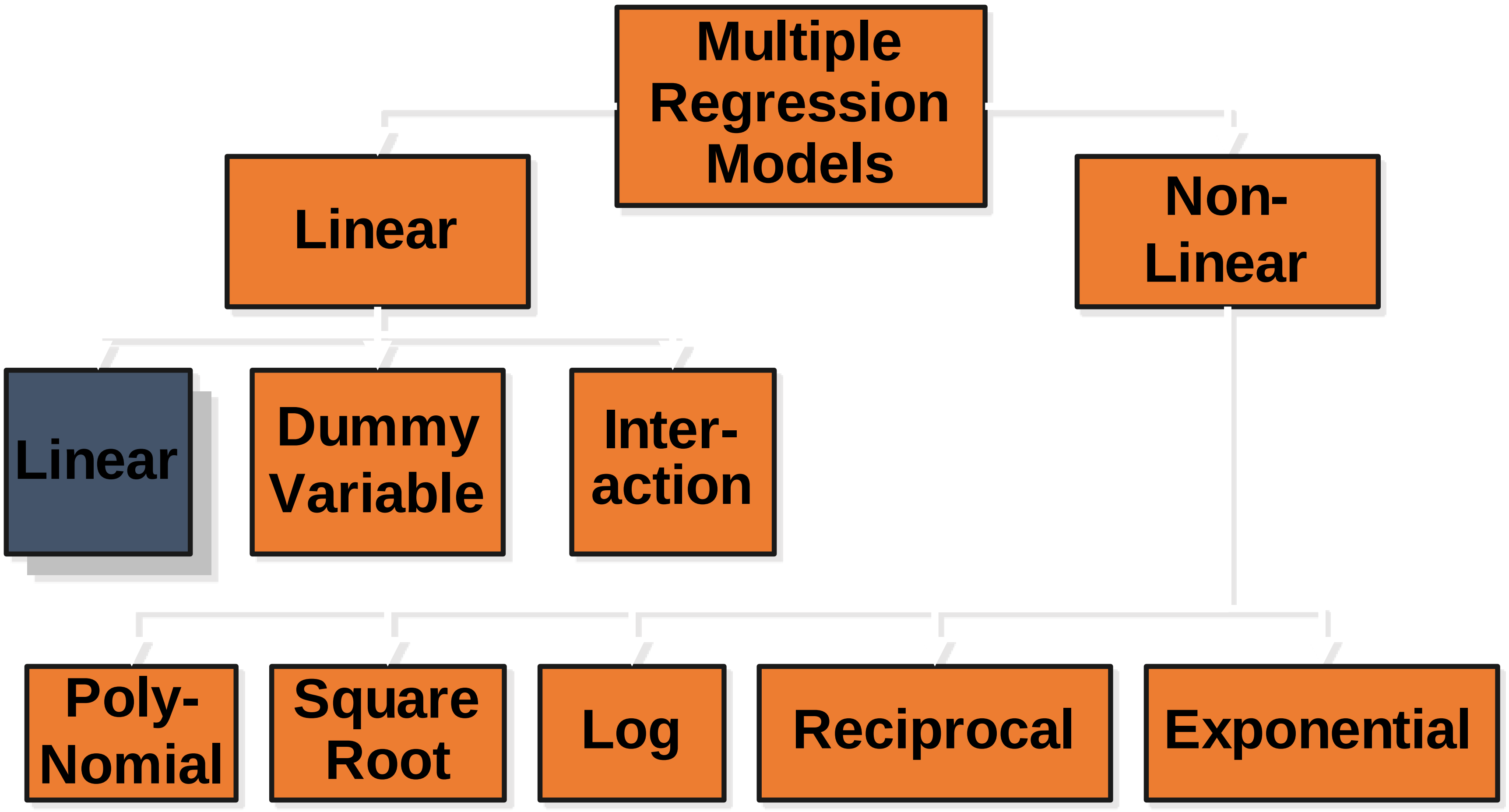
- **Positive linear relationship**
- **Negative linear relationship**
- **No relationship between X and Y**
- **Positive curvilinear relationship**
- **U-shaped curvilinear**
- **Negative curvilinear relationship**

Multiple Regression Models



Multiple Regression Equations

Multiple Regression Models



Linear Model

Relationship between one dependent & two or more independent variables is a linear function

Population Y-intercept

Population slopes

Random error

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P + \varepsilon$$

Dependent *(response)* variable

Independent *(explanatory)* variables

Method of Least Squares

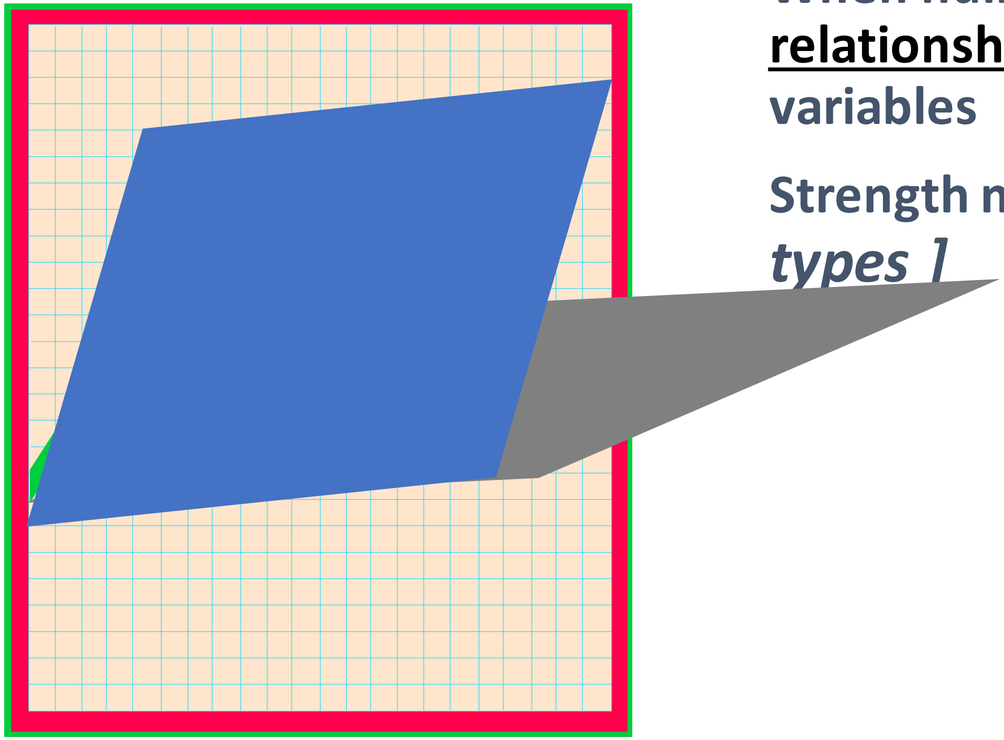
- **The straight line that *best fits* the data.**

- **Determine the straight line for which the differences between the actual values (Y) and the values that would be predicted from the fitted line of regression (Y-hat) are as small as possible.**

Measures of Variation

- **Explained variation (sum of squares due to regression)**
- **Unexplained variation (error sum of squares)**
- **Total sum of squares**

Coefficient of Multiple Determination



When null hypothesis is rejected, a <u>relationship</u> between Y and the X variables exists.

Strength measured by R^2 ***[several types]***

Coefficient of Multiple Determination

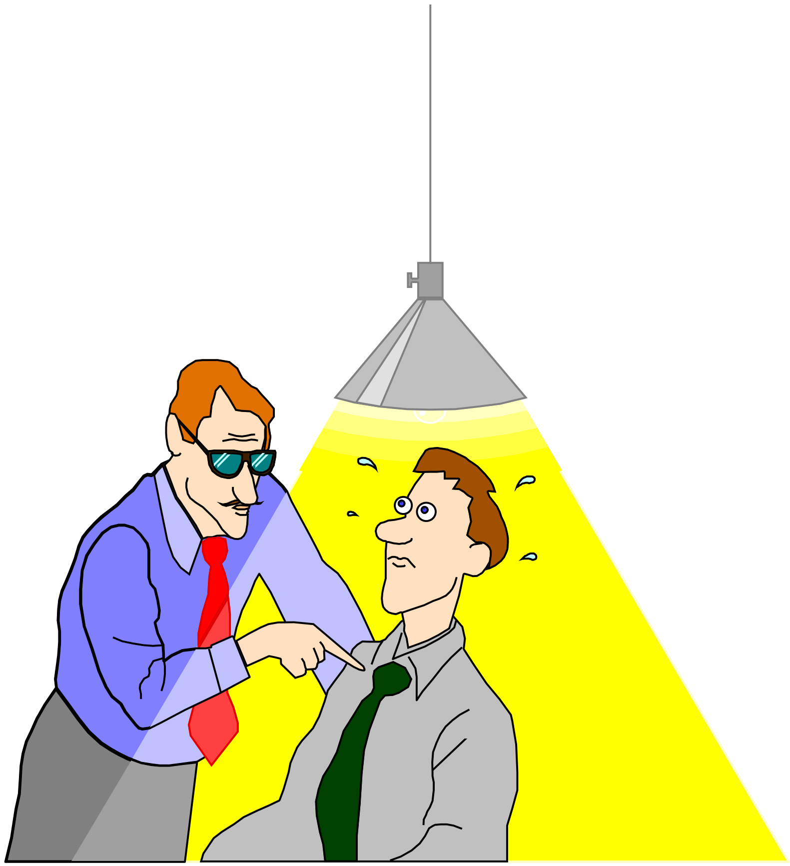
$$R^2_{y.123---P}$$

The proportion of Y that is explained by the set of explanatory variables selected

Standard Error of the Estimate

$S_{y.x}$

the measure of variability around the line of regression



Confidence interval estimates

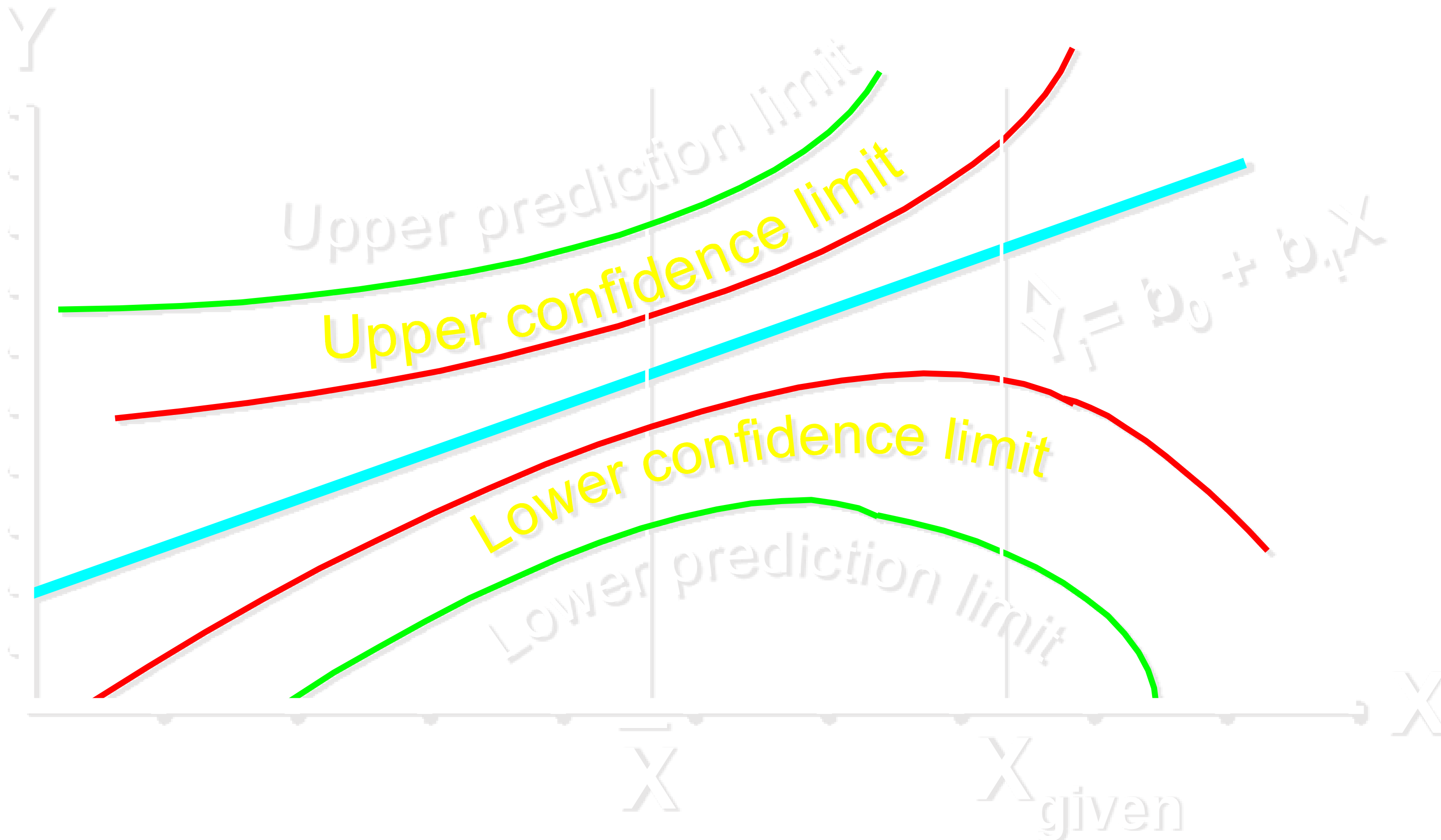
- **True mean**

$\mu_{Y.X}$

- **Individual**

Y-hat_i

Interval Bands [from simple regression]



Multiple Regression Equation

Y-hat = $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_P x_P + \varepsilon$

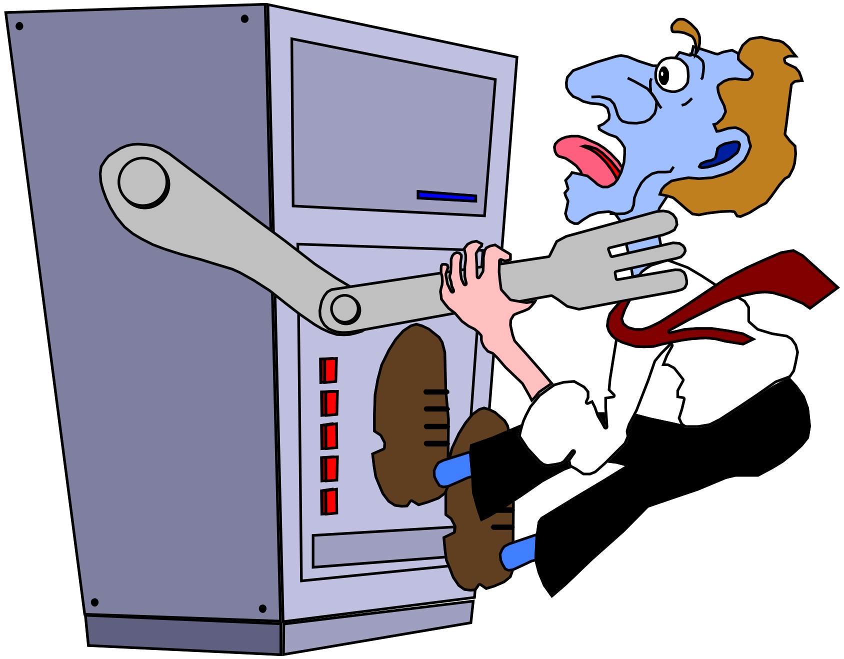
where:

β_0 = y-intercept {a constant value}

β_1 = slope of Y with variable x_1 holding the variables $x_2, x_3, \ldots, x_P$ effects constant

β_P = slope of Y with variable x_P holding all other variables' effects constant

Who is in Charge?

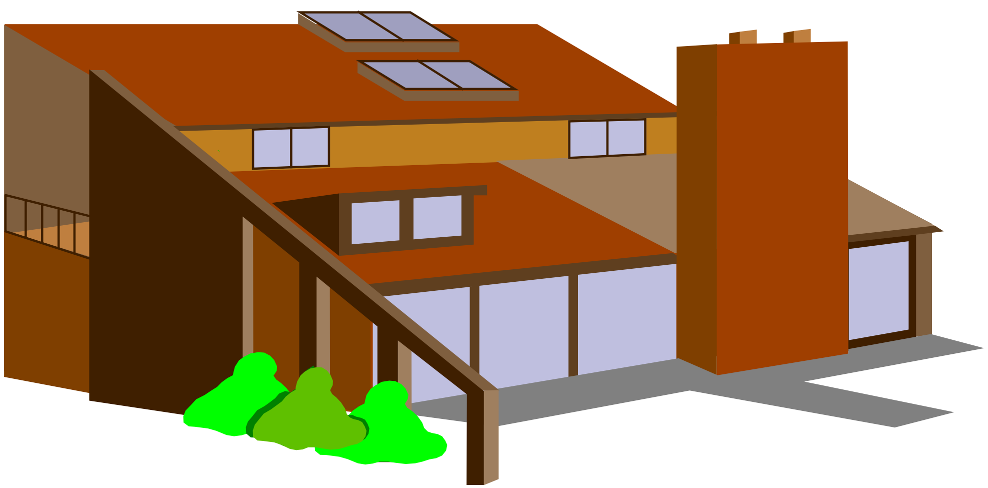


Mini-Case

Predict the consumption of home heating oil during January for homes located around Screne Lakes. Two explanatory variables are selected - - average daily atmospheric temperature (°F) and the amount of attic insulation (“).

Mini-Case

Develop a model for estimating heating oil used for a single family home in the month of January based on average temperature and amount of insulation in inches.



Oil (Gal)	Temp (^{0}F)	Insulation
275.30	40	3
363.80	27	3
164.30	40	10
40.80	73	6
94.30	64	6
230.90	34	6
366.70	9	6
300.60	8	10
237.80	23	10
121.40	63	3
31.40	65	10
203.50	41	6
441.10	21	3
323.00	38	3
52.50	58	10

Mini-Case

- **What preliminary conclusions can home owners draw from the data?**
- **What could a home owner expect heating oil consumption (in gallons) to be if the outside temperature is 15 ºF when the attic insulation is 10 inches thick?**

Multiple Regression Equation
[mini-case]

Dependent variable: Gallons Consumed

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	562.151	21.0931	26.6509	0.0000
Insulation	-20.0123	2.34251	-8.54313	0.0000
Temperature	-5.43658	0.336216	-16.1699	0.0000

R-squared = 96.561 percent

R-squared (adjusted for d.f.) = 95.9879 percent

Standard Error of Est. = 26.0138

+

Multiple Regression Equation
[mini-case]

$$\textbf{Y-hat} = 562.15 - 5.44x_1 - 20.01x_2$$

where: x_1 = temperature [degrees F]

x_2 = **attic** insulation [inches]

Multiple Regression Equation
[mini-case]

$$\text{Y-hat} = 562.15 - 5.44x_1 - 20.01x_2$$

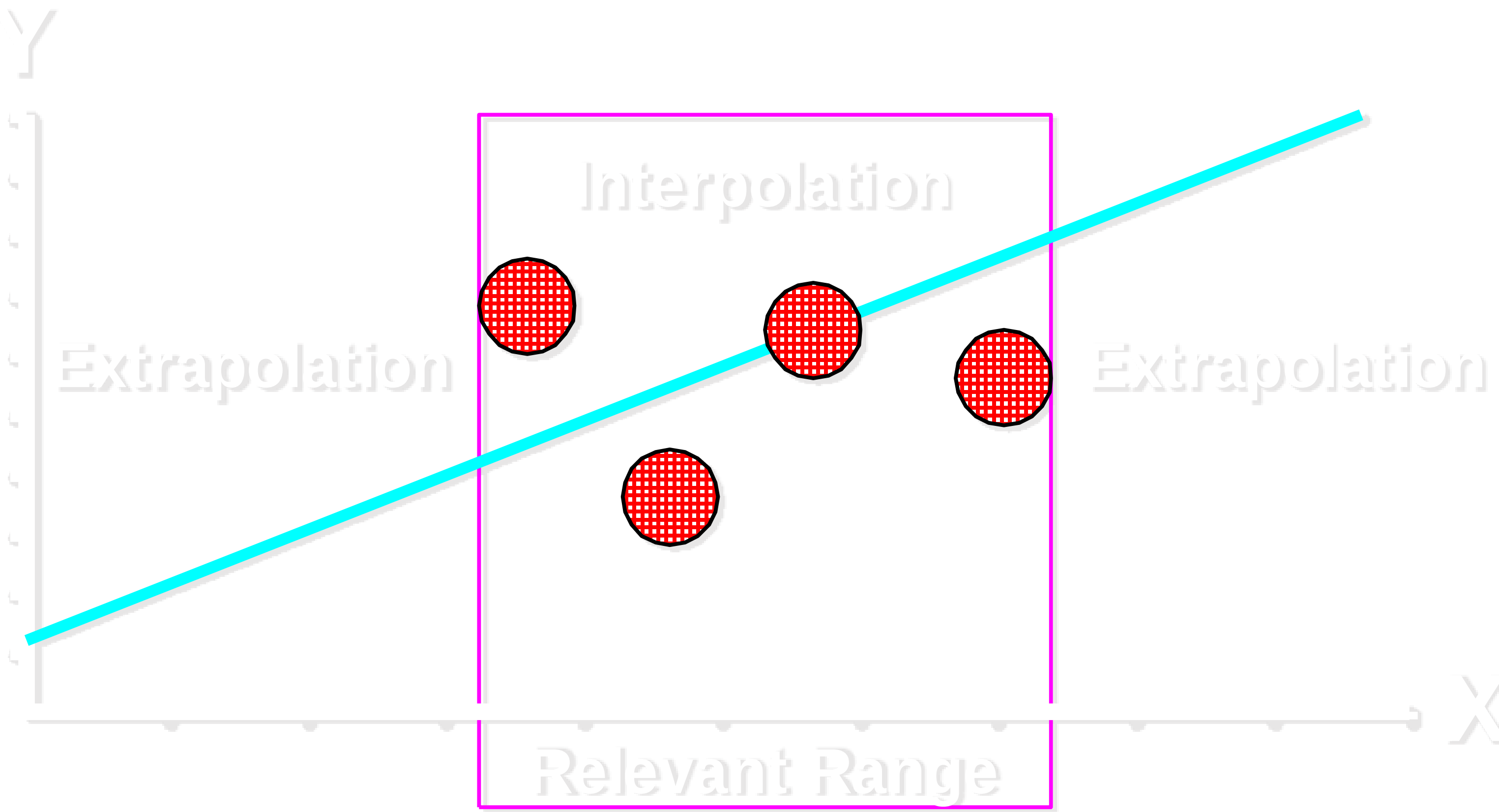
thus:

- For a home with zero inches of attic insulation and an outside temperature of 0 ^oF, 562.15 gallons of heating oil would be consumed.

[caution .. data boundaries .. extrapolation]

+

Extrapolation



Multiple Regression Equation
[mini-case]

$\hat{Y} = 562.15 - 5.44x_1 - 20.01x_2$

- For a home with zero attic insulation and an outside temperature of zero, 562.15 gallons of heating oil would be consumed. *[caution .. data boundaries .. extrapolation]*
- For each incremental increase in degree F of temperature, ***for a given amount of attic insulation,*** heating oil consumption drops 5.44 gallons.

+

Multiple Regression Equation
[mini-case]

$\hat{Y}$-hat = 562.15 - 5.44x_1 - 20.01x_2

- For a home with zero attic insulation and an outside temperature of zero, 562 gallons of heating oil would be consumed. ***[caution …]***
- For each incremental increase in degree F of temperature, for a given amount of attic insulation, heating oil consumption drops 5.44 gallons.

- **For each incremental increase in inches of attic insulation, *at a given temperature,* heating oil consumption drops 20.01 gallons.**

Multiple Regression Prediction
[mini-case]

Y-hat = 562.15 - 5.44x_1 - 20.01x_2

with x_1 = 15°F and x_2 = 10 inches

Y-hat = 562.15 - 5.44(15) - 20.01(10)

= 280.45 gallons consumed

Coefficient of Multiple Determination [mini-case]

$$R^2_{y.12} = .9656$$

96.56 percent of the variation in heating oil can be explained by the variation in temperature and insulation.

and

Coefficient of Multiple Determination

- Proportion of variation in *Y* 'explained' by all *X* variables taken together
- $R^2_{Y.12}$ = Explained variation = SSR ____ Total variation SST
- Never decreases when new *X* variable is added to model
  - Only *Y* values determine SST
  - Disadvantage when comparing models

Coefficient of Multiple Determination

Adjusted

- Proportion of variation in *Y* ‘explained’ by all *X* variables taken together
- Reflects
  - **Sample size**
  - **Number of independent variables**
- Smaller [more conservative] than $\boldsymbol{R^2_{Y.12}}$
- Used to compare models

Coefficient of Multiple Determination (adjusted)

$$R^2_{(adj)\ y.123---P}$$

The proportion of Y that is explained by the set of independent [explanatory] variables selected, adjusted for the number of independent variables and the sample size.

Coefficient of Multiple Determination (adjusted) [Mini-Case]

$$R^2_{adj} = 0.9599$$

95.99 percent of the variation in heating oil consumption can be explained by the model - adjusted for number of independent variables and the sample size

Coefficient of *Partial* Determination

- Proportion of variation in Y ‘explained’ by variable X_P holding all others constant
- Must estimate separate models
- Denoted $R^2{}_{Y1.2}$ in two X variables case
  - Coefficient of partial determination of X_1 with Y holding X_2 constant
- Useful in selecting X variables

Coefficient of Partial Determination
[p. 878]

$$R^2_{y1.234\,---\,P}$$

The coefficient of partial variation of variable Y with x_1 holding constant the effects of variables x_2, x_3, x_4, ... x_P.

Coefficient of Partial Determination [Mini-Case]

$$R^2_{y1.2} = 0.9561$$

For a fixed (constant) amount of insulation, 95.61 percent of the variation in heating oil can be explained by the variation in average atmospheric temperature. [p. 879]

Coefficient of Partial Determination [Mini-Case]

$$R^2_{y2.1} = 0.8588$$

For a fixed (constant) temperature, 85.88 percent of the variation in heating oil can be explained by the variation in amount of insulation.

Testing Overall Significance

- Shows if there is a linear relationship between all *X* variables together & *Y*
- Uses p-value
- Hypotheses
  - H_0: $\beta_1 = \beta_2 = \ldots = \beta_P = 0$
    - No linear relationship
  - H_1: At least one coefficient is not 0
    - At least one *X* variable affects *Y*

Testing Model Portions

- Examines the contribution of a *set* of *X* variables to the relationship with *Y*
- Null hypothesis:
  - Variables in set do not improve significantly the model when all other variables are included
- Must estimate separate models
- Used in selecting *X* variables

Diagnostic Checking

- H_0 retain or reject

  If **reject** - $\{p\text{-value} \leq 0.05\}$
- R^2_{adj}
- Correlation matrix
- Partial correlation matrix

Multicollinearity

- High correlation between *X* variables
- Coefficients measure combined effect
- Leads to unstable coefficients depending on *X* variables in model
- Always exists; matter of degree
- Example: Using both total number of rooms and number of bedrooms as explanatory variables in same model

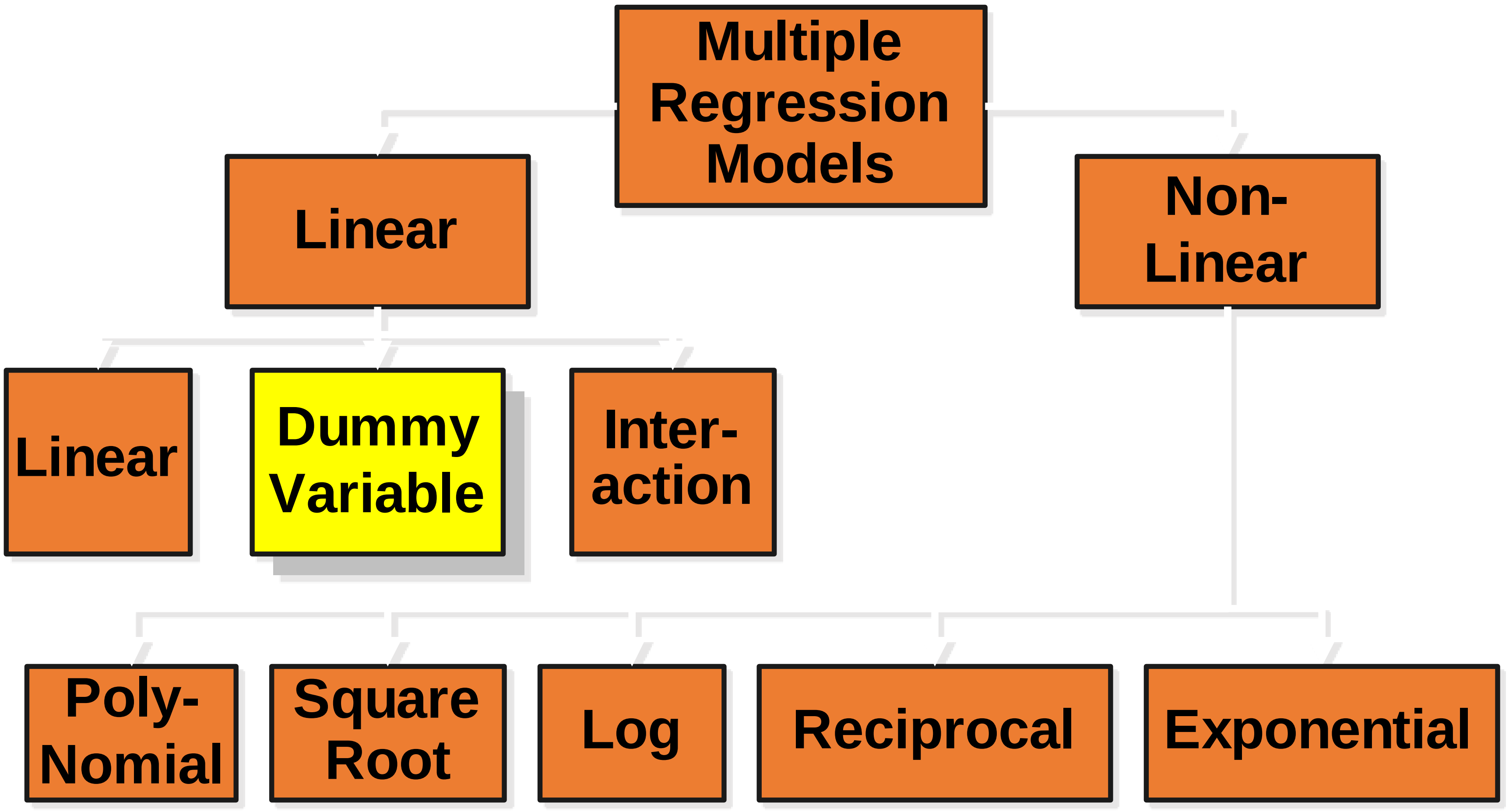
Detecting Multicollinearity

- Examine correlation matrix
  - Correlations between pairs of *X* variables are more than with *Y* variable
- Few remedies
  - Obtain new sample data
  - Eliminate one correlated *X* variable

Evaluating Multiple Regression Model Steps

- Examine variation measures
- Do residual analysis
- Test parameter significance
  - Overall model
  - Portions of model
  - Individual coefficients
- Test for multicollinearity

Multiple Regression Models



Dummy-Variable Regression Model

- Involves categorical *X* variable with two levels
  - e.g., female-male, employed-not employed, etc.

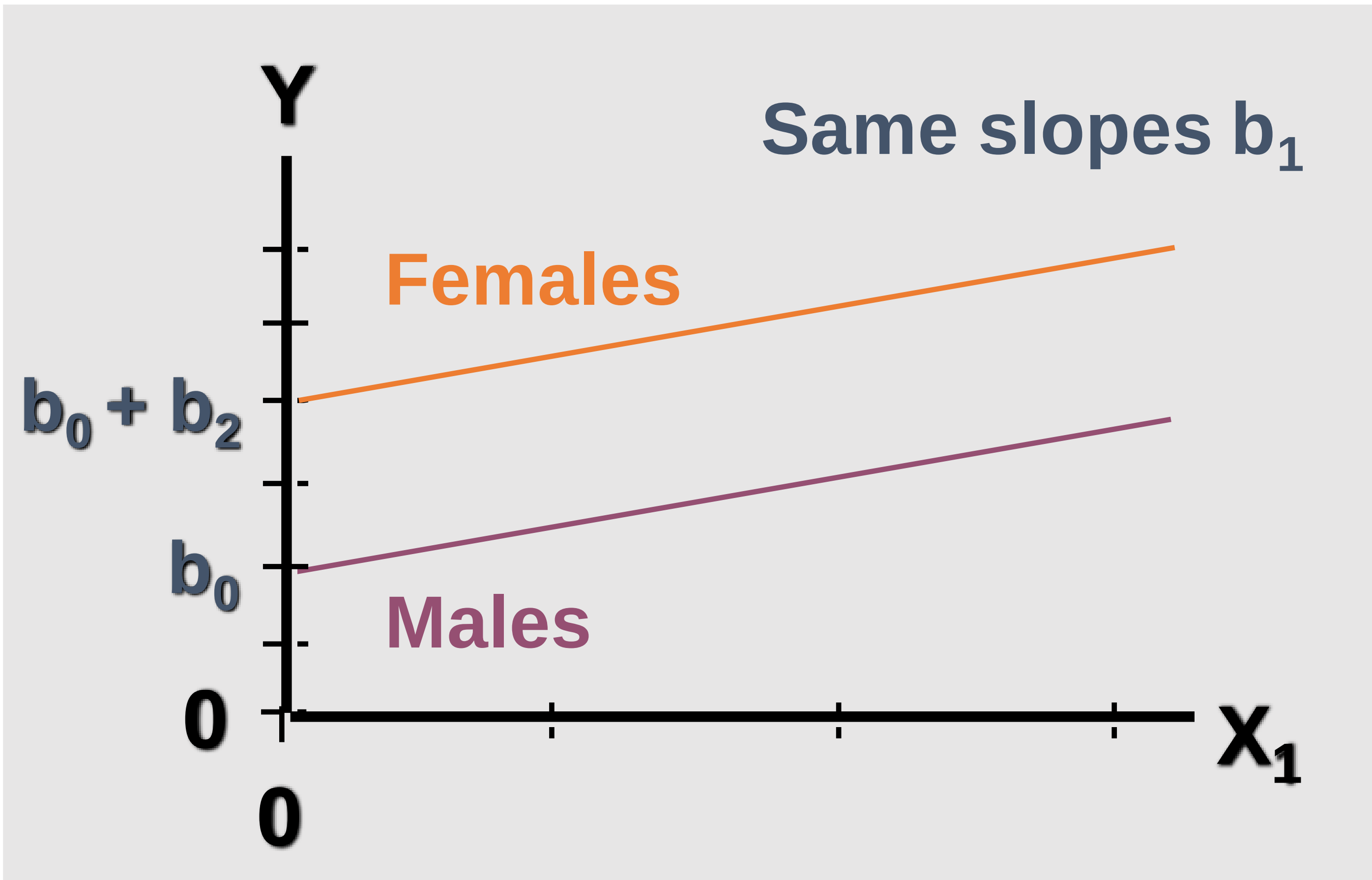
Dummy-Variable Regression Model

- Involves categorical *X* variable with two levels
  - e.g., female-male, employed-not employed, etc.
- Variable levels coded 0 & 1

Dummy-Variable Regression Model

- Involves categorical *X* variable with two levels
  - e.g., female-male, employed-not employed, etc.
- Variable levels coded 0 & 1
- Assumes only intercept is different
  - Slopes are constant across categories

Dummy-Variable Model Relationships



Dummy Variables

- **Permits use of qualitative data**

  (e.g.: seasonal, class standing, location, gender).

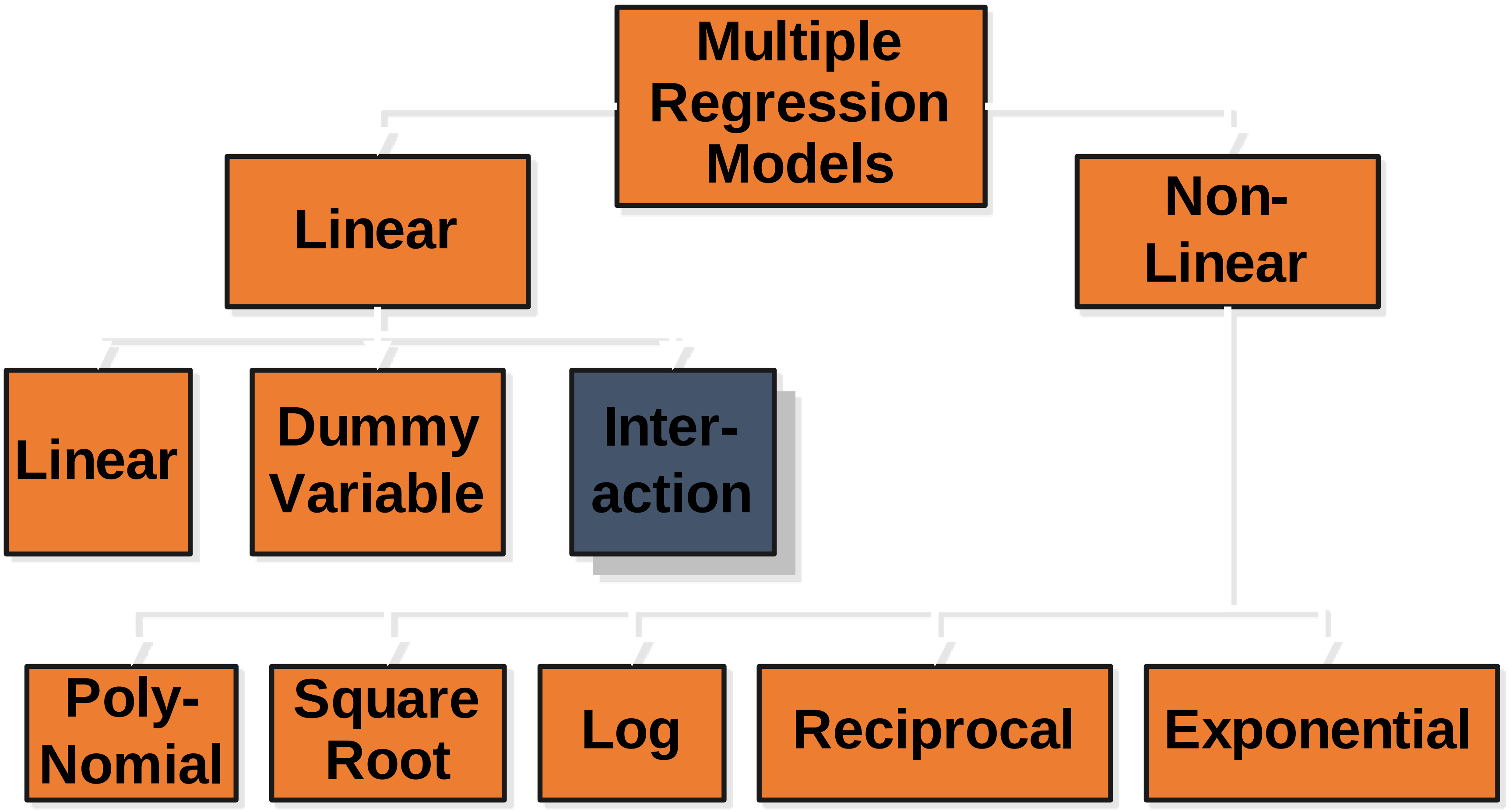
- **0, 1 coding (nominative data)**

- **As part of Diagnostic Checking;**

  incorporate outliers

  (i.e.: large residuals) and influence measures.

Multiple Regression Models



Interaction Regression Model

- Hypothesizes interaction between pairs of *X* variables
  - Response to one *X* variable varies at different levels of another *X* variable
- Contains two-way cross product terms

$$\mathbf{Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon}$$

- Can be combined with other models

  e.g. dummy variable models

Effect of Interaction

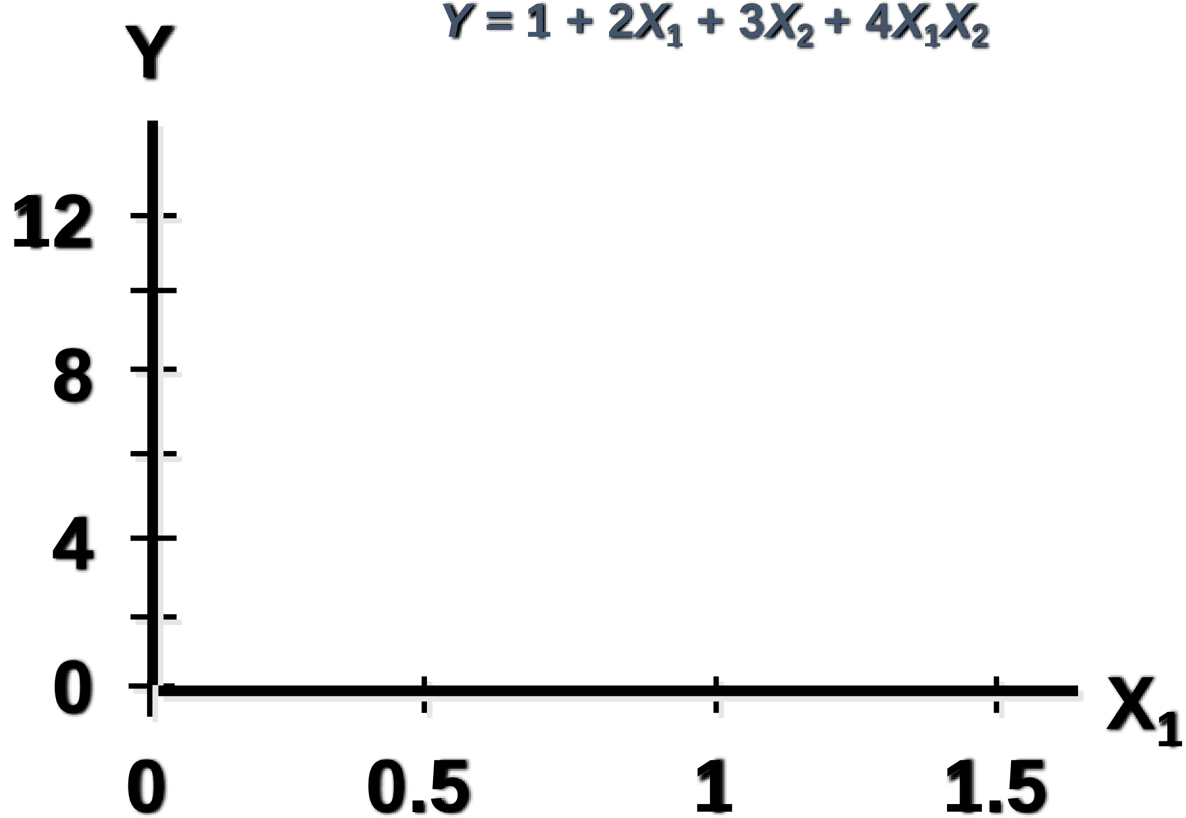
- Given:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$$

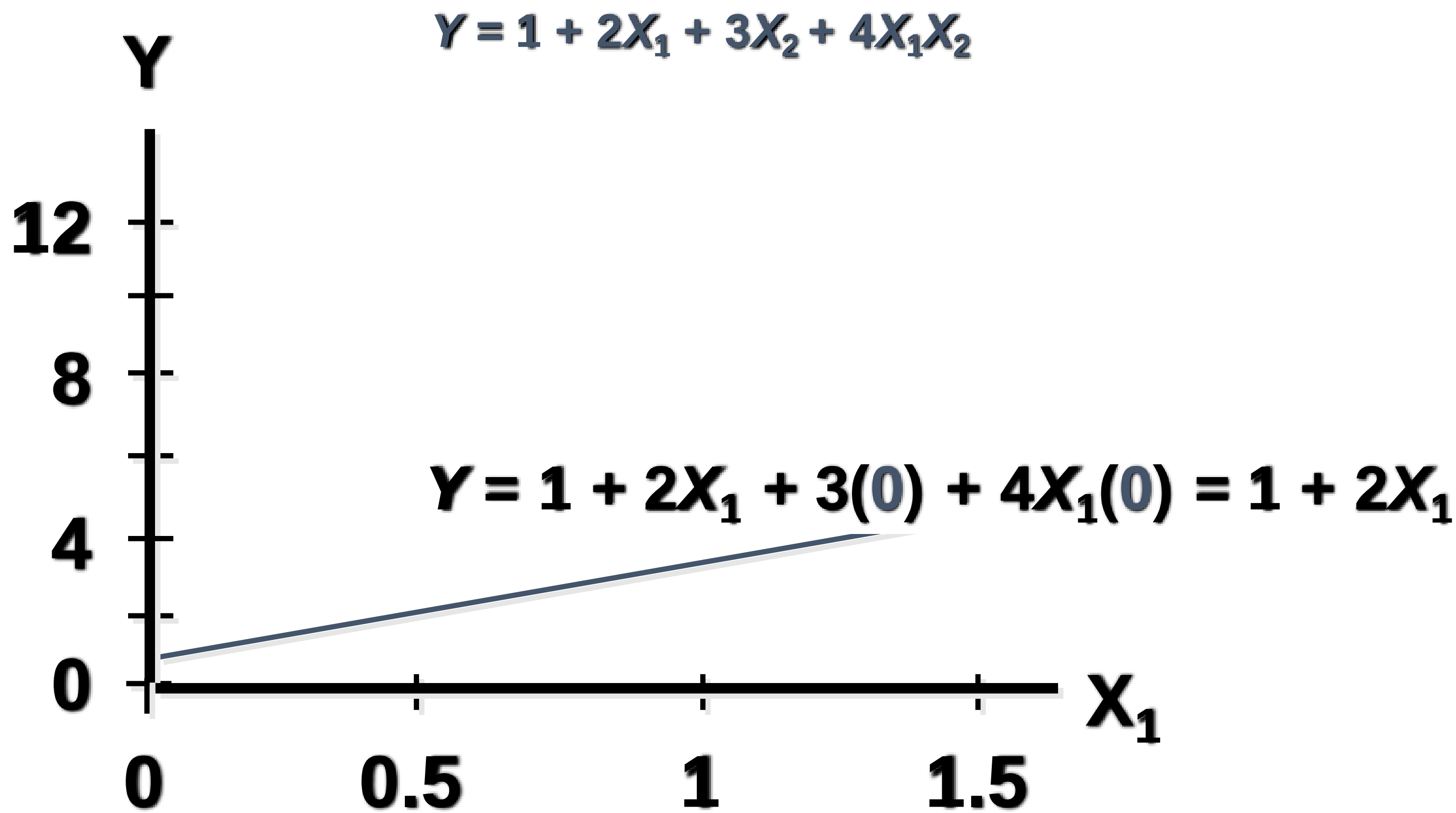
- Without interaction term, effect of X_1 on Y is measured by β_1
- With interaction term, effect of X_1 on Y is measured by $\beta_1 + \beta_3 X_2$
  - Effect increases as X_{2i} increases

Interaction Example

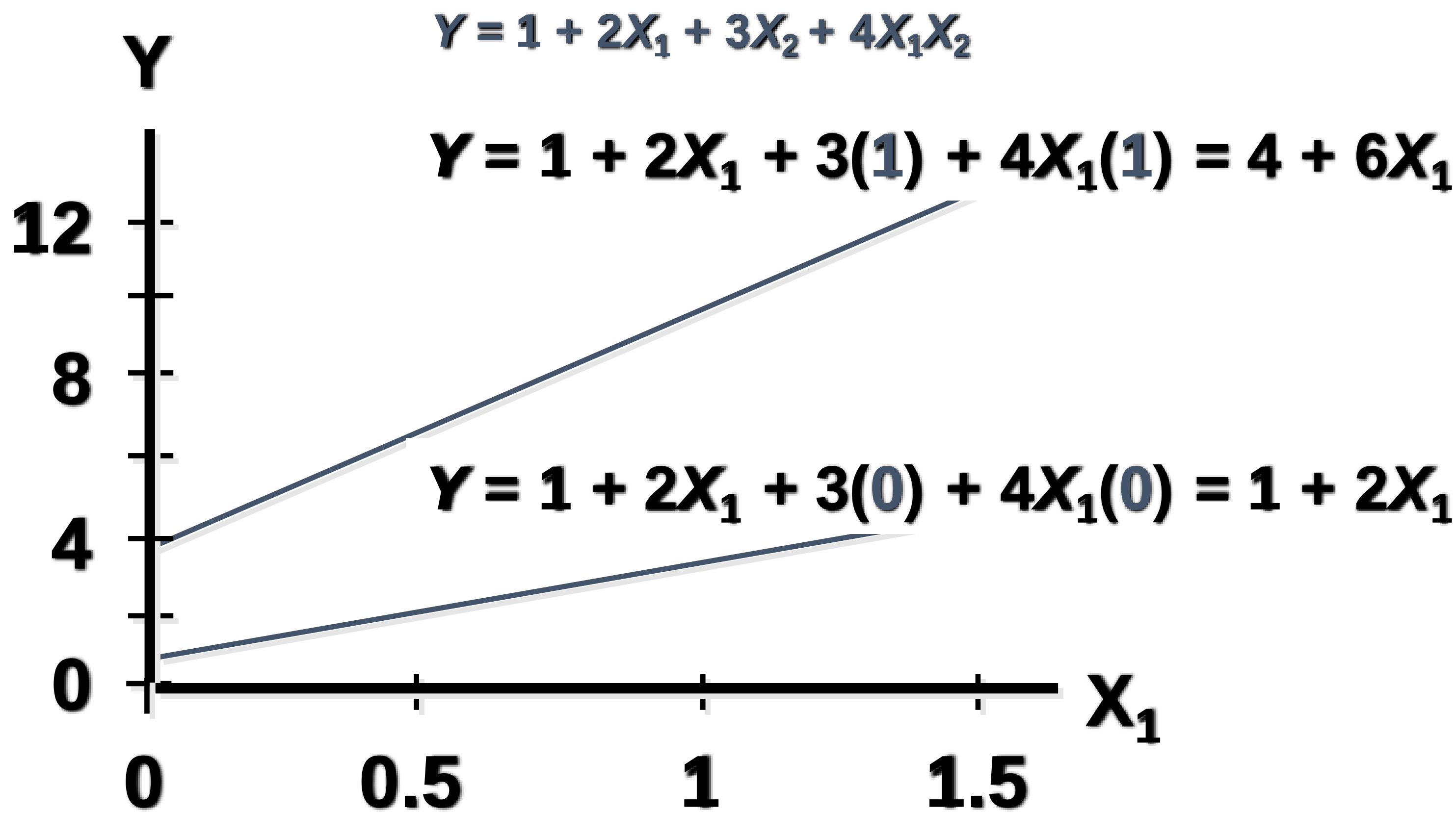
$Y = 1 + 2X_1 + 3X_2 + 4X_1X_2$



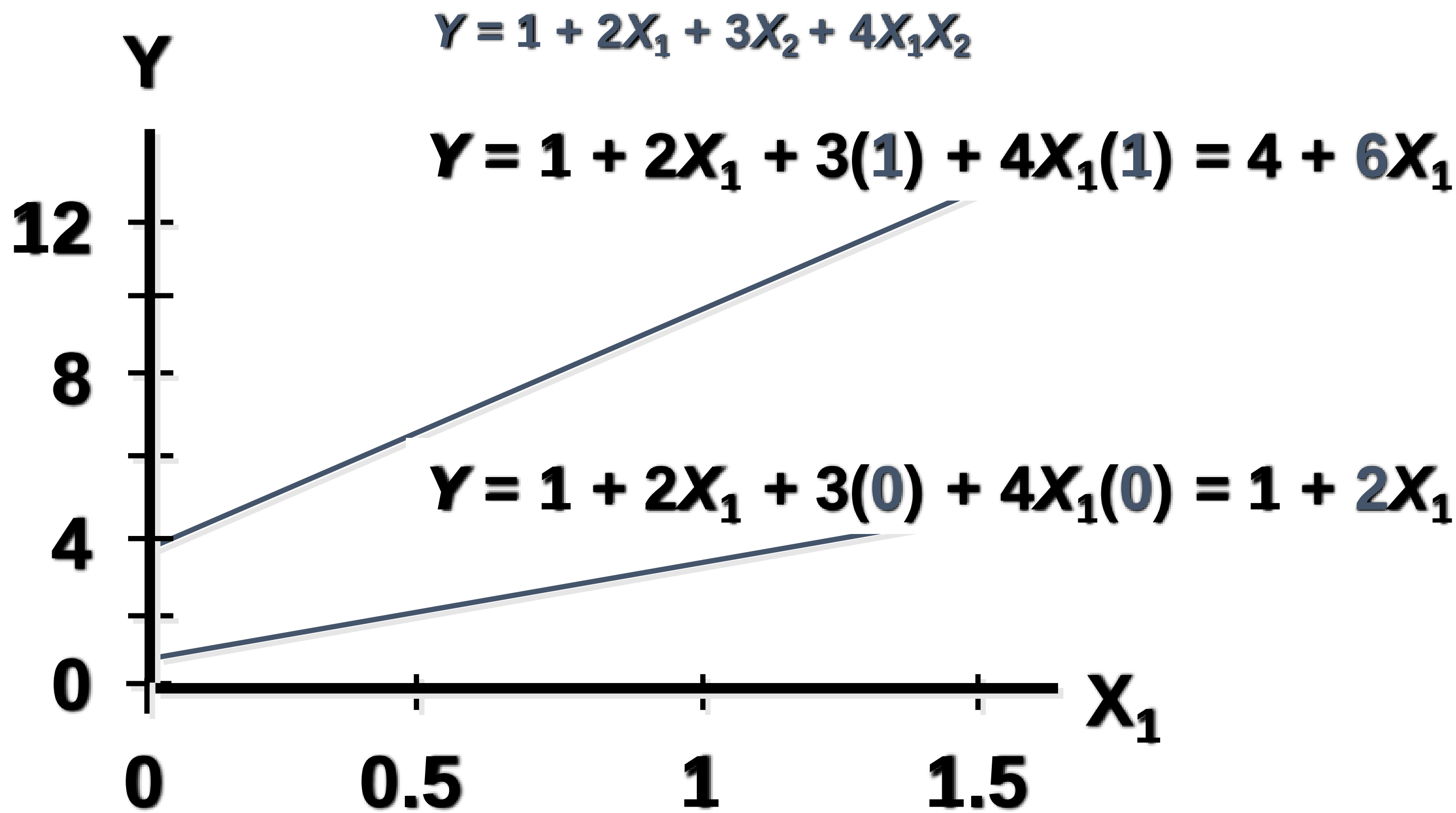
Interaction Example



Interaction Example

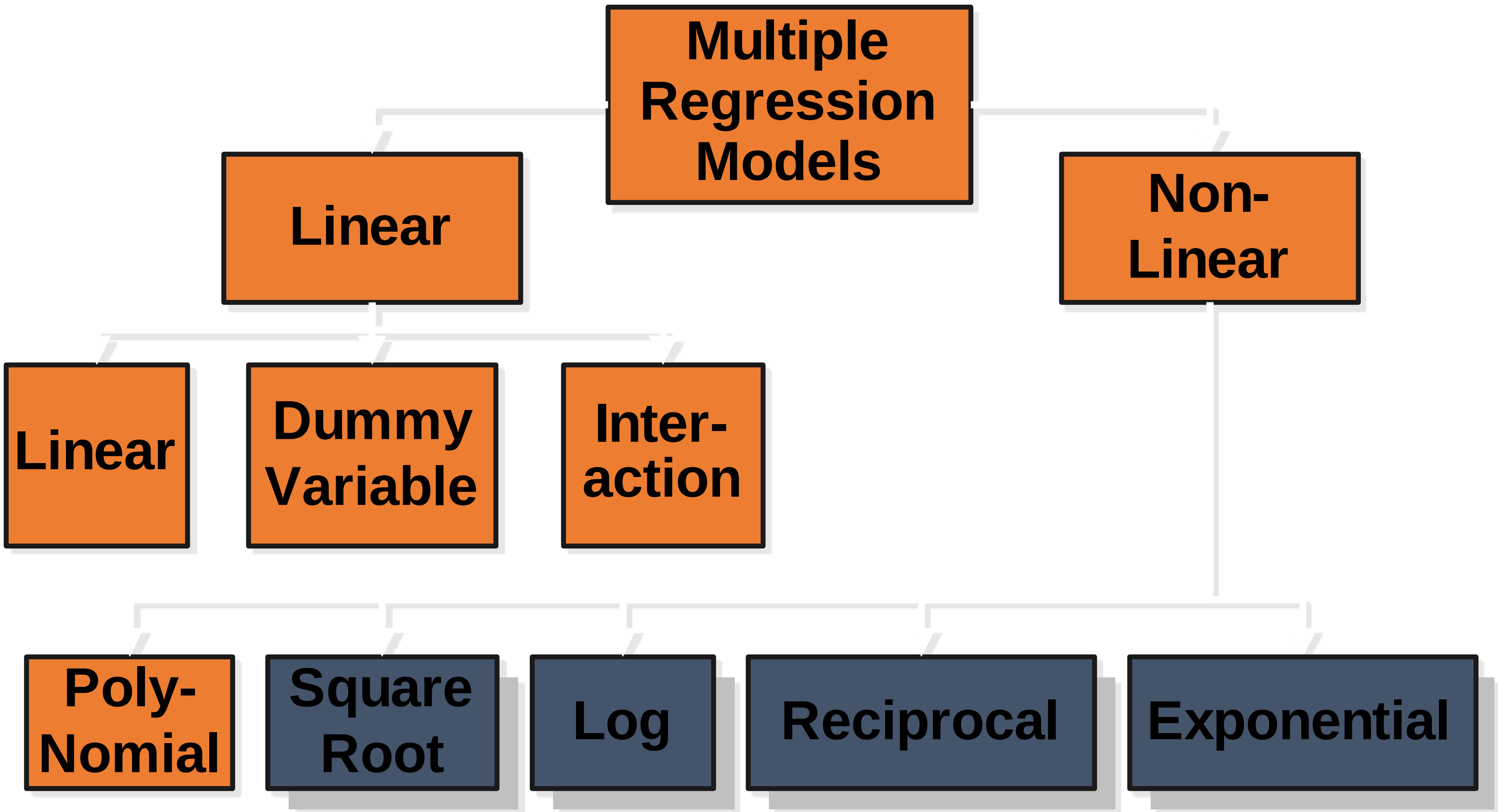


Interaction Example



Effect (slope) of X_1 on Y does depend on X_2 value

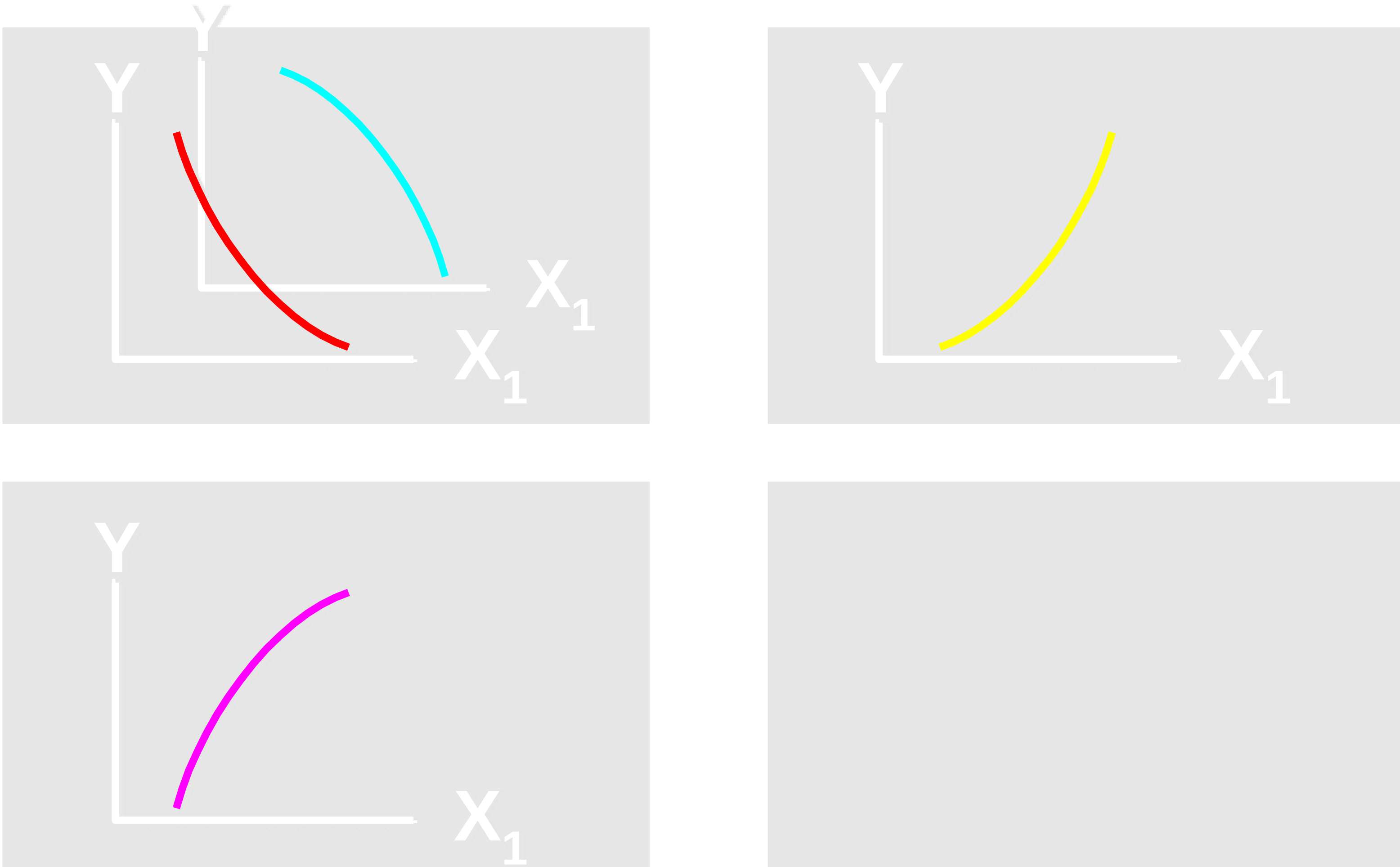
Multiple Regression Models



Inherently Linear Models

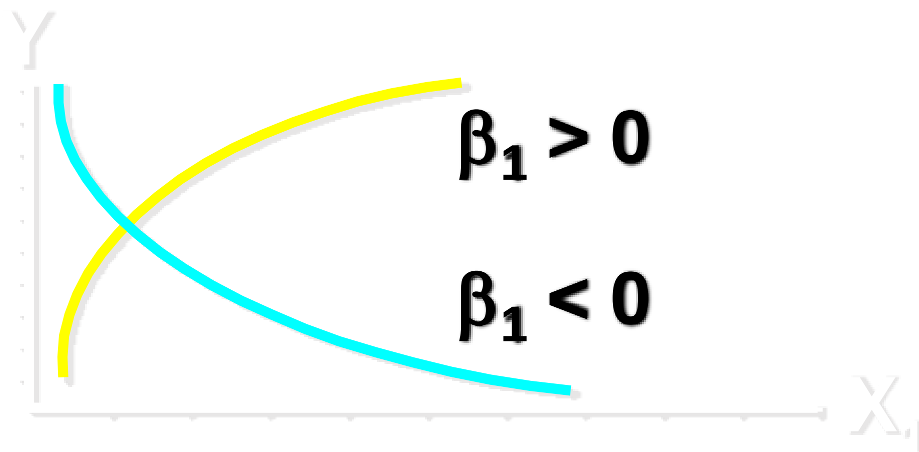
- Non-linear models that can be expressed in linear form
  - Can be estimated by least square in linear form
- Require data transformation

Curvilinear Model Relationships



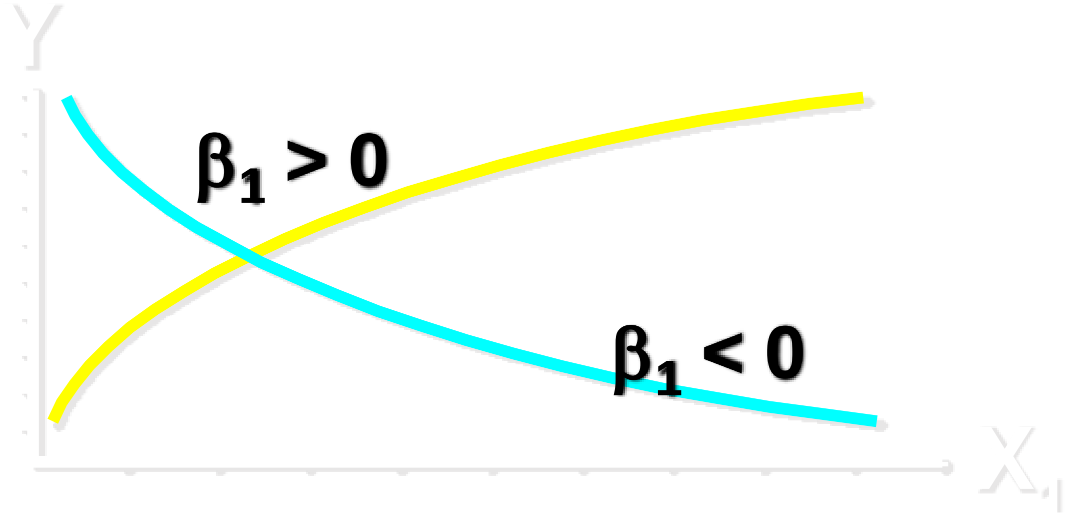
Logarithmic Transformation

$$Y = \beta \; + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \; \varepsilon$$



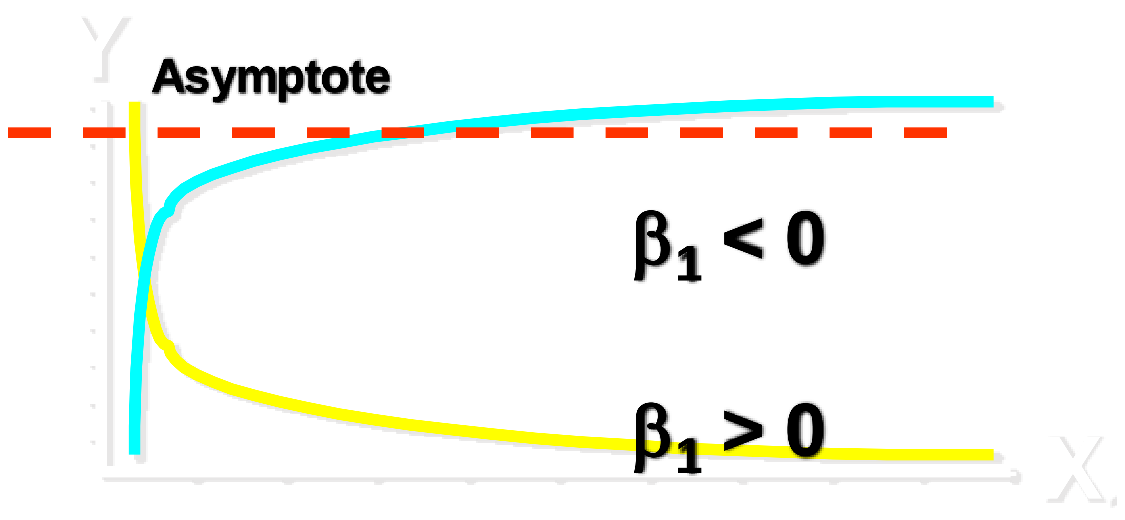
Square-Root Transformation

$$Y_i = \beta_0 + \beta_1 \sqrt{X_{1i}} + \beta_2 \sqrt{X_{2i}} + \varepsilon_i$$



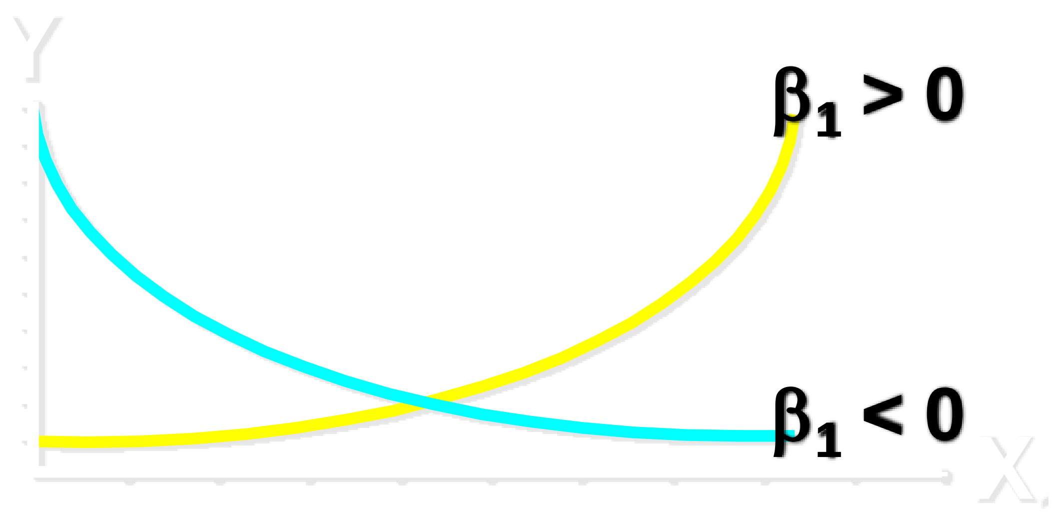
Reciprocal Transformation

$$Y_i = \beta_0 + \beta_1 \frac{1}{X_{1i}} + \beta_2 \frac{1}{X_{2i}} + \varepsilon_i$$



Exponential Transformation

$$Y_i = e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}} \varepsilon_i$$



Overview

- Explained the linear multiple regression model
- Interpreted linear multiple regression computer output
- Explained multicollinearity
- Described the types of multiple regression models

Source of Elaborate Slides

Prentice Hall, Inc

Levine, et. all, First Edition