Unit II: Multiple Linear Regression

CTE309 Machine Learning

AS2024: BE Information Technology

College of Science and Technology

Royal University of Bhutan

Overview

- Introduction
- Assumptions
- Dummy variable vs Dummy trap
- Significance level
- Building a model
- Demo

Introduction

- examine the relationship between a dependent (response) variable and two or more independent (predictor) variables.
- models the linear relationship between these variables
- predict the dependent variable based on the values of the independent variables.

Dependent Variable
(Response Variable)

Independent Variables
(Predictors)

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

Y intercept

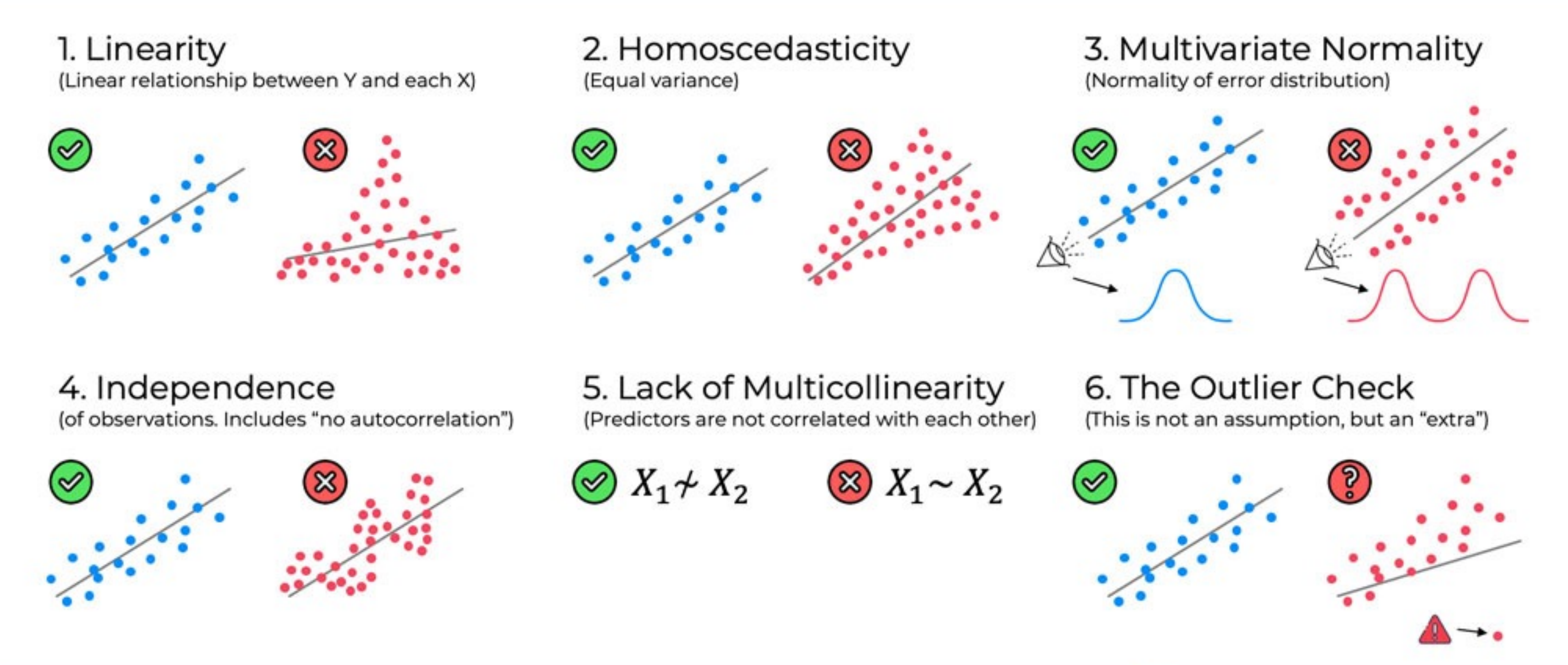
Slope
Coefficient

Error Term

Introduction

- Advantages:
  - More accurate predictions
  - insights into relationship
- Challenges:
  - Multicollinearity
  - Overfitting
- Examples:
  - Potato $=\beta 0+ \beta 1(\text{fertilizer}) - \beta 2(\text{sun}) + \beta 3(\text{rain})$

Assumptions of Linear Regression



Dummy Variable

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 + b_1{}^*x_1 + b_2{}^*x_2 + b_3{}^*x_3 + ???$$

Dummy Variable

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	~~0~~	1
191,050.39	153,441.51	101,145.55	407,934.54	California	~~0~~	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	~~0~~	1

$$y = b_0 + b_1 {}^* x_1 + b_2 {}^* x_2 + b_3 {}^* x_3 + ???$$

Dummy Variable

Profit	R&D Spend	Admin	Marketing	~~State~~
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

Dummy Variables

New York	~~California~~
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1{}^*x_1 + b_2{}^*x_2 + b_3{}^*x_3 + b_4{}^*D_1 + \cancel{b_5{}^*D_2}$$

Always omit one dummy variable

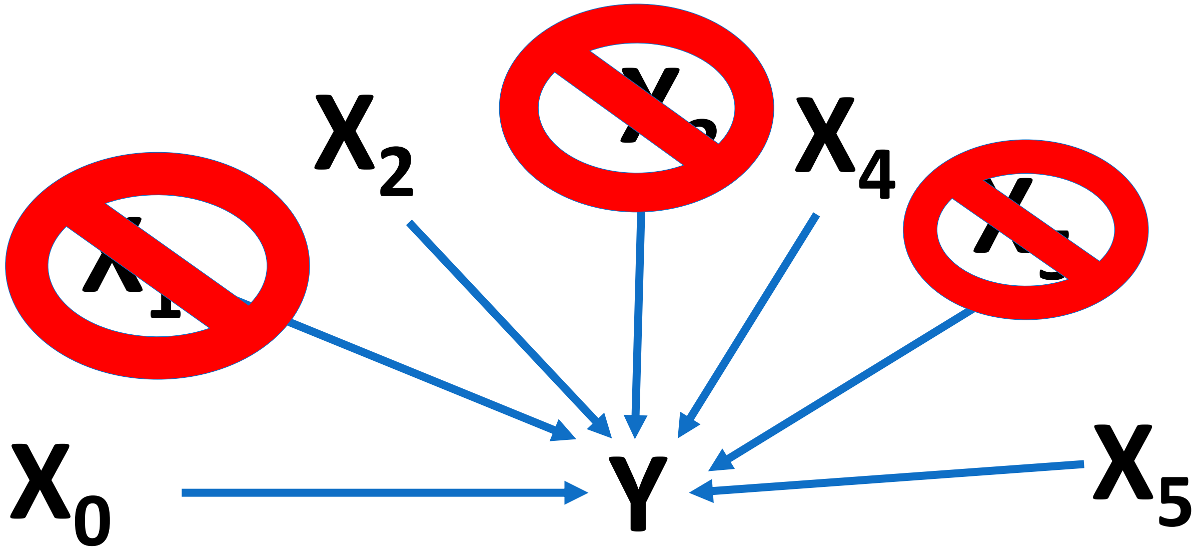
Statistical Significance

- Tail = 0.5
- Tail = 0.25
- Tail = 0.13
- Tail = 0.06
- Tail = 0.03
- Tail = 0.01

P Value

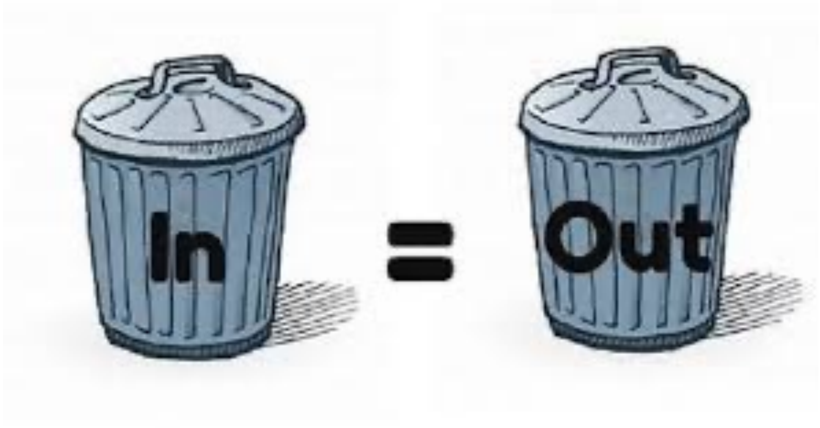
$\alpha = 0.05$

Building Model



Why can’t we use all of them?

Building model



Methods of building model

- All in
- Backward elimination
- Forward selection
- Bidirectional elimination
- Score comparison

Stepwsie regression

Method : All in

- Prior knowledge or
- You have to or
- Preparing for backward elimination

Building A Model

Backward Elimination

STEP 1: Select a significance level to stay in the model (e.g. SL = 0.05)

STEP 2: Fit the full model with all possible predictors

STEP 3: Consider the predictor with the <u>highest</u> P-value. If P > SL, go to STEP 4, otherwise go to FIN

STEP 4: Remove the predictor

STEP 5: Fit model without this variable*

FIN: Your Model Is Ready

Building A Model

Forward Selection

STEP 1: Select a significance level to enter the model (e.g. SL = 0.05)

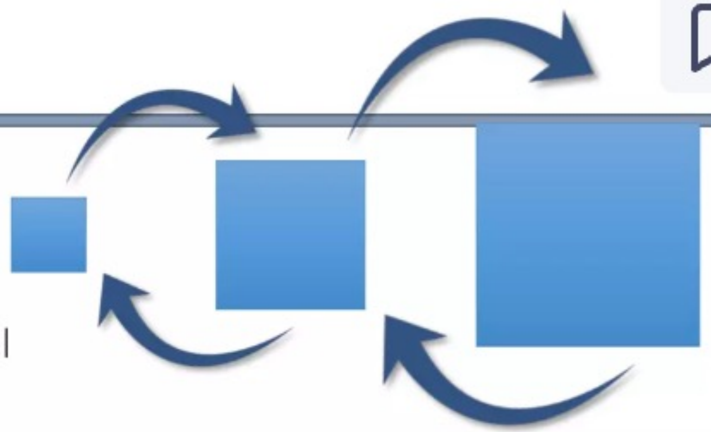
STEP 2: Fit all simple regression models $y \sim x_n$ Select the one with the lowest P-value

STEP 3: Keep this variable and fit all possible models with one extra predictor added to the one(s) you already have

STEP 4: Consider the predictor with the <u>lowest</u> P-value. If P < SL, go to STEP 3, otherwise go to FIN

FIN: Keep the previous model

Building A Model



Bidirectional Elimination

STEP 1: Select a significance level to enter and to stay in the model
e.g.: SLENTER = 0.05, SLSTAY = 0.05

STEP 2: Perform the next step of Forward Selection (new variables must have: P < SLENTER to enter)

STEP 3: Perform ALL steps of Backward Elimination (old variables must have P < SLSTAY to stay)

STEP 4: No new variables can enter and no old variables can exit

FIN: Your Model Is Ready

Building Model:

All Possible Models

STEP 1: Select a criterion of goodness of fit (e.g. Akaike criterion)

↓

STEP 2: Construct All Possible Regression Models: 2^N-1 total combinations

↓

STEP 3: Select the one with the best criterion

↓

FIN: Your Model Is Ready

Example:
10 columns means
1,023 models

Thank you