Simple Regression Analysis

Parthasarathi Edupally[1]

[1]DataConscientious LLP, Mumbai

Elements of Econometrics,
Russell Square International College, Mumbai

Outline

1 Objectives

Outline

1 Objectives

2 Causal Relation
Problem Statement
Causation v/s Correlation
Simple Regression Model
Disturbance term

Outline

1. Objectives
2. Causal Relation
   - Problem Statement
   - Causation v/s Correlation
   - Simple Regression Model
   - Disturbance term
3. Least Squares method of Regression
   - Data Fitting
   - Derivation of Least Square Estimates

Outline

1 Objectives

2 Causal Relation
- Problem Statement
- Causation v/s Correlation
- Simple Regression Model
- Disturbance term

3 Least Squares method of Regression
- Data Fitting
- Derivation of Least Square Estimates

4 Thankyou

Objectives

- To understand motivation for Regression Model

Objectives

- To understand motivation for Regression Model
- Procedure for finding Regression Coefficients

Objectives

- To understand motivation for Regression Model
- Procedure for finding Regression Coefficients
- Interpretation of Regression Coefficients

Problem Statement

- Given a population of interest, What causes a particular characteristic of its people ?

Problem Statement

- Given a population of interest, What causes a particular characteristic of its people ?
- For example,
  - Population: All patients suffering with Diabetes in Mumbai

Problem Statement

- Given a population of interest, What causes a particular characteristic of its people ?
- For example,
  - Population: All patients suffering with Diabetes in Mumbai
  - Characteristic: Average sugar levels (say Y, during early morning, averaged over a month)

Problem Statement

- Given a population of interest, What causes a particular characteristic of its people ?
- For example,
  - Population: All patients suffering with Diabetes in Mumbai
  - Characteristic: Average sugar levels (say Y, during early morning, averaged over a month)
  - Possible causes: Lifestyle (say X_1, number of hours of walking, everyday averaged over a month), Diet (say X_2, percentage of fibre in diet in any day, averaged over a month) etc

Problem Statement

- Given a population of interest, What causes a particular characteristic of its people ?
- For example,
  - Population: All patients suffering with Diabetes in Mumbai
  - Characteristic: Average sugar levels (say Y, during early morning, averaged over a month)
  - Possible causes: Lifestyle (say X_1, number of hours of walking, everyday averaged over a month), Diet (say X_2, percentage of fibre in diet in any day, averaged over a month) etc
- How do we approach this problem ?

Problem Statement

- Given a population of interest, What causes a particular characteristic of its people ?
- For example,
  - Population: All patients suffering with Diabetes in Mumbai
  - Characteristic: Average sugar levels (say Y, during early morning, averaged over a month)
  - Possible causes: Lifestyle (say X_1, number of hours of walking, everyday averaged over a month), Diet (say X_2, percentage of fibre in diet in any day, averaged over a month) etc
- How do we approach this problem ?
  - Correlation: to measure possible linear relation between variables Y and X_1, X_2

Problem Statement

- Given a population of interest, What causes a particular characteristic of its people ?
- For example,
  - Population: All patients suffering with Diabetes in Mumbai
  - Characteristic: Average sugar levels (say Y, during early morning, averaged over a month)
  - Possible causes: Lifestyle (say X_1, number of hours of walking, everyday averaged over a month), Diet (say X_2, percentage of fibre in diet in any day, averaged over a month) etc
- How do we approach this problem ?
  - Correlation: to measure possible linear relation between variables Y and X_1, X_2
  - Regression: Assuming (Contemplating) that linear relation exists, we try to fit regression model to data

Causation v/s Correlation

- Correlation doesn't imply Causation, Why?

Causation v/s Correlation

- Correlation doesn't imply Causation, Why?
- Because there might be other variables, which we are not aware of, actually influencing both X and Y together

Causation v/s Correlation

- Correlation doesn't imply Causation, Why?
- Because there might be other variables, which we are not aware of, actually influencing both X and Y together
- Medical researchers solve this probelm with curated experiments to get data - control and treatment groups

Causation v/s Correlation

- Correlation doesn't imply Causation, Why?
- Because there might be other variables, which we are not aware of, actually influencing both X and Y together
- Medical researchers solve this probelm with curated experiments to get data - control and treatment groups
- In Economics, we have only observational data - so difficult to establish causation
- Ceteris Paribus assumption is basis of all regression modelling in Economics

<Scatterplot Demo>

Simple Regression Model

- Assumptions:

Simple Regression Model

- Assumptions:
  - There is linear relation between X and Y (Model is correctly specified)

Simple Regression Model

- Assumptions:
  - There is linear relation between X and Y (Model is correctly specified)
  - All the other variables influencing Y can be modelled with error term

Simple Regression Model

- Assumptions:
  - There is linear relation between X and Y (Model is correctly specified)
  - All the other variables influencing Y can be modelled with error term
- Model: $Y_i = \beta_1 + \beta_2 X_i + u_i$
- Where Y_i, X_i and u_i are Y, X and u values of observation i; β_1 and β_2 are parameters of model

Simple Regression Model

- Assumptions:
  - There is linear relation between X and Y (Model is correctly specified)
  - All the other variables influencing Y can be modelled with error term
- Model: $Y_i = \beta_1 + \beta_2 X_i + u_i$
- Where Y_i, X_i and u_i are Y, X and u values of observation i; β_1 and β_2 are parameters of model
- Y is called dependent variable
- X is called regressor or independent variable

Disturbance or Error Term

- It is used to capture, mathematically, everything else not explained by this simple regression model

Disturbance or Error Term

- It is used to capture, mathematically, everything else not explained by this simple regression model
- Following are reasons for its inclusion:
  - Omission of explanatory variables: there might be other variables influencing Y

Disturbance or Error Term

- It is used to capture, mathematically, everything else not explained by this simple regression model
- Following are reasons for its inclusion:
  - Omission of explanatory variables: there might be other variables influencing Y
  - Aggregation of variables: we are trying to find average relation

Disturbance or Error Term

- It is used to capture, mathematically, everything else not explained by this simple regression model
- Following are reasons for its inclusion:
  - Omission of explanatory variables: there might be other variables influencing Y
  - Aggregation of variables: we are trying to find average relation
  - Model misspecification: X chosen might be not appropriate

Disturbance or Error Term

- It is used to capture, mathematically, everything else not explained by this simple regression model
- Following are reasons for its inclusion:
  - Omission of explanatory variables: there might be other variables influencing Y
  - Aggregation of variables: we are trying to find average relation
  - Model misspecification: X chosen might be not appropriate
  - Functional misspecification: there might be nonlinear relation

Disturbance or Error Term

- It is used to capture, mathematically, everything else not explained by this simple regression model
- Following are reasons for its inclusion:
  - Omission of explanatory variables: there might be other variables influencing Y
  - Aggregation of variables: we are trying to find average relation
  - Model misspecification: X chosen might be not appropriate
  - Functional misspecification: there might be nonlinear relation
  - Measurement error: wouldnt have been possible to measure X and Y without errors

Data Fitting

- This whole business of finding a causal relation between X and Y boils down to fitting a line

Data Fitting

- This whole business of finding a causal relation between X and Y boils down to fitting a line
- Problem: We have to fit a "best" possible line (Model) to data

Data Fitting

- This whole business of finding a causal relation between X and Y boils down to fitting a line
- Problem: We have to fit a "best" possible line (Model) to data
- Means we have to find (estimate) best possible values of parameters β_1 and β_2

Data Fitting

- This whole business of finding a causal relation between X and Y boils down to fitting a line
- Problem: We have to fit a "best" possible line (Model) to data
- Means we have to find (estimate) best possible values of parameters β_1 and β_2
- Procedure to find "best" values:
  - Define best line: the one which has least "Residual Error"
  - Residual error: $e_i = Y_i - \hat{Y}_i$
  - Fitted Y: $\hat{Y}_i = b_1 + b_2 X_i$

Data Fitting

- This whole business of finding a causal relation between X and Y boils down to fitting a line
- Problem: We have to fit a "best" possible line (Model) to data
- Means we have to find (estimate) best possible values of parameters β_1 and β_2
- Procedure to find "best" values:
  - Define best line: the one which has least "Residual Error"
  - Residual error: $e_i = Y_i - \hat{Y}_i$
  - Fitted Y: $\hat{Y}_i = b_1 + b_2 X_i$
  - So we use standard calculus to minimize residual error to find parameters

Derivation of Least Square Estimates

Derive estimates of simple regression model using calculus

Interpretation of Coefficients

From textbook !!

"In God we trust, all others bring data."

William Edwards Deming (1900 - 1993).