Multiple Regression Analysis

Parthasarathi Edupally[1]

[1]DataConscientious LLP, Mumbai

Elements of Econometrics,
Russell Square International College, Mumbai

Outline

Outline

Outline

Outline

Outline

Outline

Causal Inference and Ceteris Paribus

- Important goal of regression framework is to find the causal relation

Causal Inference and Ceteris Paribus

- Important goal of regression framework is to find the causal relation
- We needed to have ‘ceteris paribus’ along with corrrelation between Y and X to ensure this

Causal Inference and Ceteris Paribus

- Important goal of regression framework is to find the causal relation
- We needed to have ‘ceteris paribus’ along with corrrelation between Y and X to ensure this
- How did we deal with factors other than X ?
    - Assumed that all other factors are randomly affecting Y (as a collection)

Causal Inference and Ceteris Paribus

- Important goal of regression framework is to find the causal relation
- We needed to have 'ceteris paribus' along with corrrelation between Y and X to ensure this
- How did we deal with factors other than X ?
  - Assumed that all other factors are randomly affecting Y (as a collection)
  - These were modelled using disturbance term

Causal Inference and Ceteris Paribus

- Important goal of regression framework is to find the causal relation
- We needed to have ‘ceteris paribus’ along with corrrelation between Y and X to ensure this
- How did we deal with factors other than X ?
  - Assumed that all other factors are randomly affecting Y (as a collection)
  - These were modelled using disturbance term
- Multiple regression analysis helps us to explicitly enforce ‘ceteris paribus’ condition
- So we can measure - How much are Y and X_1 correlated keeping other factors X_2, $X3$...constant?

Extension of Simple Linear Regression

- Our model now explicitly includes other relevant explanatory variables -

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i,$$

Extension of Simple Linear Regression

- Our model now explicitly includes other relevant explanatory variables -

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i,$$

- In SLR framework, we had to fit a line to data in two dimensional space

Extension of Simple Linear Regression

- Our model now explicitly includes other relevant explanatory variables -

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i,$$

- In SLR framework, we had to fit a line to data in two dimensional space
- Now it boils down to fitting plane in higher dimensional space

Extension of Simple Linear Regression

- Our model now explicitly includes other relevant explanatory variables -

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i,$$

- In SLR framework, we had to fit a line to data in two dimensional space
- Now it boils down to fitting plane in higher dimensional space
- This chapter extends ideas from simple linear regression to multiple variables case

Derivation of Regression Coefficients

- Idea is same as in SLR - except that now we have more unknowns and equations

Derivation of Regression Coefficients

- Idea is same as in SLR - except that now we have more unknowns and equations
- Steps involved in the OLS algorithm to derive coefficients are:
  - precisely define ‘good’ model - in OLS case the one with least residual sum of squares

Derivation of Regression Coefficients

- Idea is same as in SLR - except that now we have more unknowns and equations
- Steps involved in the OLS algorithm to derive coefficients are:
    - precisely define ‘good’ model - in OLS case the one with least residual sum of squares
    - derive residual sum of squares as a function of parameters

Derivation of Regression Coefficients

- Idea is same as in SLR - except that now we have more unknowns and equations
- Steps involved in the OLS algorithm to derive coefficients are:
  - precisely define ‘good’ model - in OLS case the one with least residual sum of squares
  - derive residual sum of squares as a function of parameters
  - minimize this function with respect to parameters using normal equations
- Solving for unknown parameters in normal equations is a problem of finding k (parameters) unknowns in N (sample size) equations
- Using linear Algebra we can get analytical solution for the parameter vector

Interpretation of Regression Coefficients

- Consider a two variable regression model,

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3, \quad \text{Fitted model}$$

Interpretation of Regression Coefficients

- Consider a two variable regression model,

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3, \quad \text{Fitted model}$$

- $\hat{\beta}_1$ - is the predicted value of Y fixing X_2 and X_3 to zero

Interpretation of Regression Coefficients

- Consider a two variable regression model,

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3, \quad \text{Fitted model}$$

- $\hat{\beta}_1$ - is the predicted value of Y fixing X_2 and X_3 to zero
- $\hat{\beta}_2, \hat{\beta}_3$ - are partial effects of X_2 and X_3

Interpretation of Regression Coefficients

- Consider a two variable regression model,

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3, \quad \text{Fitted model}$$

- $\hat{\beta}_1$ - is the predicted value of Y fixing X_2 and X_3 to zero
- $\hat{\beta}_2, \hat{\beta}_3$ - are partial effects of X_2 and X_3
- More precisely, we have difference equation given by

$$\triangle \hat{Y} = \hat{\beta}_2 \triangle X_2 + \hat{\beta}_3 \triangle X_3$$

- Multiple regression model allows us to model correlation between explanatory variables:
    - β_2 derived is equivalent to simple regression of residual of X_2 on residual of Y

Interpretation of Regression Coefficients

- Consider a two variable regression model,

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3, \quad \text{Fitted model}$$

- $\hat{\beta}_1$ - is the predicted value of Y fixing X_2 and X_3 to zero
- $\hat{\beta}_2, \hat{\beta}_3$ - are partial effects of X_2 and X_3
- More precisely, we have difference equation given by

$$\triangle \hat{Y} = \hat{\beta}_2 \triangle X_2 + \hat{\beta}_3 \triangle X_3$$

- Multiple regression model allows us to model correlation between explanatory variables:
  - β_2 derived is equivalent to simple regression of residual of X_2 on residual of Y
  - Where residual is obtained by removing the effect of X_3 on X_2 and Y

Assumptions of CLRM

- In order to talk about properties of coefficients we have to revisit underlying assumptions
- These are similar to the case of simple linear regression setup
- Assumption.1
  - The model is correctly specified and is linear in parameters

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u,$$

Assumptions of CLRM

- In order to talk about properties of coefficients we have to revisit underlying assumptions
- These are similar to the case of simple linear regression setup
- Assumption.1
  - The model is correctly specified and is linear in parameters

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u,$$

- Assumption.2
  - There doesnt exist an exact linear relationship among regressors in the sample
  - This we will deal seperately in Multicollinearity section

Assumptions of CLRM

- Assumption.3
  - The disturbance term has zero expectation
  - This is reasonable assumption when intercept term is used
  - It means inherent error in the model for generating Y is on an average zero
- Assumption.4
  - The disturbance term is Homoscedastic
  - Error in generating Y doesnt depend on X value

Assumptions of CLRM

- Assumption.3
  - The disturbance term has zero expectation
  - This is reasonable assumption when intercept term is used
  - It means inherent error in the model for generating Y is on an average zero
- Assumption.4
  - The disturbance term is Homoscedastic
  - Error in generating Y doesnt depend on X value
- Assumption.5
  - Disturbance terms have independent distributions
  - Error terms are independent across X values

Assumptions of CLRM

- Assumption.3
  - The disturbance term has zero expectation
  - This is reasonable assumption when intercept term is used
  - It means inherent error in the model for generating Y is on an average zero
- Assumption.4
  - The disturbance term is Homoscedastic
  - Error in generating Y doesnt depend on X value
- Assumption.5
  - Disturbance terms have independent distributions
  - Error terms are independent across X values
- Assumption.6
  - Disturbance term has normal distribution
  - Effect of all factors other than X can be modeled as following normal distribution

Unbiased Coefficients

- We can show this using an expression for estimate of coefficient

$$b_j = \beta_j + \sum_{i=1}^{n} a_{ij}^{\star} u_i$$

- This is similar to the case of SLR except that now $a_{ij}^{\star}$ is more complex

Unbiased Coefficients

- We can show this using an expression for estimate of coefficient

$$b_j = \beta_j + \sum_{i=1}^{n} a_{ij}^{\star} u_i$$

- This is similar to the case of SLR except that now $a_{ij}^{\star}$ is more complex
- Using above equation and assumptions of regression model we can deduce that

$$E(b_j) = \beta_j$$

Precision of Coefficients

- OLS regression coefficients are most efficient (least variance) among all the unbiased estimators: Gauss-Markov theorem
- Variance of estimators are now dependent on correlation between regressors along with MSD and n,

Precision of Coefficients

- OLS regression coefficients are most efficient (least variance) among all the unbiased estimators: Gauss-Markov theorem
- Variance of estimators are now dependent on correlation between regressors along with MSD and n,

$$\sigma^2_{b_2} = \frac{\sigma^2_u}{nMSD(X_2)} \star \frac{1}{1 - r^2_{X_2x_3}}$$

- Intuitively, greater the correlation, harder it is to discriminate between effects of explanatory variables on Y, and the less accurate will be the regression estimates

Precision of Coefficients

- Another important difference compared to SLR is estimate of σ_u^2
- Degrees of freedom becomes more important while deriving estimates of population variance of disturbace term

$$s_u^2 = \frac{1}{n-k} \sum_{i=1}^{n} e_i^2$$

- Finally standard error is given by,

$$s.e(b_2) = s_u \star \frac{1}{\sqrt{n}} \star \frac{1}{\sqrt{MSD(X_2)}} \star \frac{1}{\sqrt{1 - r_{X_2 X_3}^2}}$$

Multicollinearity

- It is a problem of correlation between explanatory variables and thus smearing the effect of individual variables on Y
- It makes variance of individual coefficients large and thus less reliable
- Multicollinearity in itself doesn't cause this issue, it has to be accompanied with atleast one of the other factors affecting variance
  - high variance of disturbance term
  - low sample size
  - low mean square deviation of X
- Note that there doesnt have to be exact linear relation, it is actually the degree of correlation that matters

What can we do about Multicollinearity

- Direct methods:
  - Reduce σ_u^2 by including ommited variable in regression equation
  - Increase sample size n, with bigger budget or clever sampling techniques
  - Make sure you have broad representation of population in survey, thus increasing MSD(X)
  - finally, keep an eye on variables which might be exactly linearly related during survey
- Indirect methods:
  - If two correlated variables are similar conceptually, combine them into a new index kind of variable
  - Drop the variable that is insignificant among correlated variables
  - Use extraneous information about variables of the population in different survey
  - Theoretical restriction of coefficients of variables

Evaluation of Estimates

- R2 cannot be used to take decision on inclusion of individual variables,
  - it just increases with arbitrary inclusion of variables in regression
  - and it cannot be decomposed to find contribution of individual variables
- F test is more useful in multiple regression setting,
  - It takes into account the number of explanatory variables in model(model complexity), hence statistic doesn't increase with arbitrary inclusion of variables
  - t-statistic tests individual significance of coefficient whereas f-statistic tests joint explainatory power of model

Model Selection

- Hypothesis testing can be thought of as a tool for model selection
- Analysis of variance with F test:
  - Given a model with k explanatory variables,

$$Y = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k + u$$

  - If we have to take decision on inclusion of $m - k$ more variables in the model
  - We can measure additional improvement in fit (in terms of increase in ESS)
  - Specifically,

$$F(m-k, n-m) = \frac{\frac{(RSS_k - RSS_m)}{(m-k)}}{\frac{RSS_m}{(n-m)}}$$

Model Selection

- When deciding on inclusion of only one variable to model, f-test is equivalent to t-test
- In other words, we derive same conclusion with either tests
- Adjusted R2:
  - R2, as it is now, increases with arbitrary inclusion of variables in model
  - Adjusted R2 takes care of this with inclusion of degrees of freedom

$$\bar{R}^2 = R^2 - \frac{k-1}{n-k}(1 - R^2)$$

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

William Edwards Deming (1900 - 1993).