Multiple Regression Lecture

Theory and Mechanics

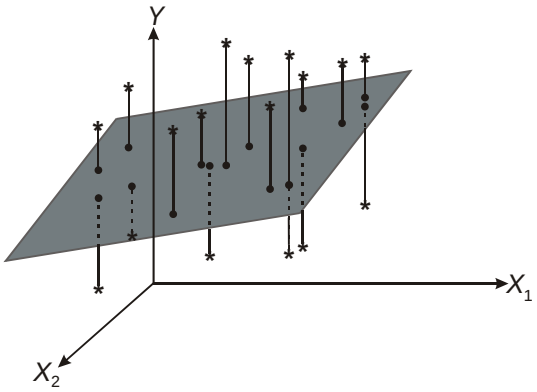
39

From Simple to Multiple Regression (MR)

- We will generalize most of the results of simple regression (one dependent, one independent variables) to multiple regression (one dependent, multiple independent variables)
- Why should we complicate our lives by using multiple regression?
  - Get better prediction
  - Understand the complex relationship among different variables

40

Pictorial Representation of Mechanics of MR



MR tries to FIT the best (in least squares sense) plane through all points in a multi-dimensional space.
MR model is simply the equation of the best fitting plane.

41

Multiple Regression Model

- Simple Regression Model (Equation): $Y_i = \beta_0 + \beta_1 X_i + e_i$
- MR Model or equation :

$$Y_i = \beta_0 + \beta_1 X_{1i} + +\beta_2 X_{2i} + ... + \beta_p X_{pi} + e_i$$

In the model, Y_i is the "ith value" of the dependent variable, X_{1i},X_{2i} etc. are the "ith value" of the independent variables, β_0 is intercept, and β_1, β_2 etc. are the coefficients for the independent variables and e_i is the error for the "ith value"

- We will ignore how to calculate these coefficients (leave it for the programs to do) and instead focus on understanding what these mean for managers

42

Multiple Regression Interpretation

- As in simple regression, multiple regression output also provides two sets of statistical tests :
  - Test for overall model (the Analysis of Variance table)
  - Test for each coefficient (the Parameter Estimates table)
- Hypothesis for overall model test:
  - H_0: The MR model does not explain the relationship between the dependent and the independent variable in the population any better than the baseline model.
  - H_1: H_0 is not true.
- Think of what it means if you **fail to reject** this null hypothesis

43

Multiple Regression Interpretation (contd.)

- Hypothesis test for each coefficients in the regression model:
  - H_0: The value of regression coefficient for the independent variable equals 0 in the population.
  - H_1: H_0 is not true.

44

Multiple Regression Interpretation (contd.)

- Regression Prediction Equation:

$\hat{Y}_i = \beta_0 + \beta_1 X_{1i} + +\beta_2 X_{2i} + ... + \beta_p X_{pi}$

- How do we interpret the coefficients?
  - Statistician's interpretation of the regression coefficient β_1 is: If X_1 changes by 1 unit, then Y changes by β_1 units, provided all other independent variables (X_2 , X_3 .. X_p) are held constant

45

Which Independent Variable is More important?

- It is a crucial issue because managers are often interested about relative importance of each coefficient – such as which is more important, X_1 or X_2 .. X_p in predicting Y?
- Many people will intuitively and incorrectly infer the most important variable is the one for which the coefficient (β) is the largest!
- But, the values of regression coefficients depend on the scale/unit of measurement for independent variables.
  - Thus, coefficients are not directly comparable.
  - Standardized coefficients (or other similar metrics) take the effect of scale/unit of measurement out and these are comparable.

46

Prediction Using MR

Regression equation, $\hat{Y}_i = \beta_0 + \beta_1 X_{1i} + +\beta_2 X_{2i} + ... + \beta_p X_{pi}$

- Once the values of (β_0 , β_1 , .. β_p) are known, it is simple to calculate predicted value of Y (dependent or target variable) for any given value of X (independent variable). Some issues to keep in mind are:
  - Be careful about going beyond the range of X-values observed in the data that ere used to calculate the values of intercept and slope (β_0 , β_1 ,.. β_p).
    - Rule of thumb: 25% beyond observed range may be OK.
  - However, note that we are working with sample numbers and hence the values of the regression parameters will change from sample to sample!
    - So, use Confidence Intervals in your predictions

47

Diagnosing MR Model Performance

- Once a model is run, we get the regression equation.
- Then, this equation is used to predict Y for each observation by plugging-in the x-values for each observation.
- The difference between the actual Y-value and the predicted Y-value is called the residual.
  - The residual is calculated for each observation.
- These residuals are the primary tools for diagnosing model performance.
  - As before large studentized residuals indicate potential problems

48

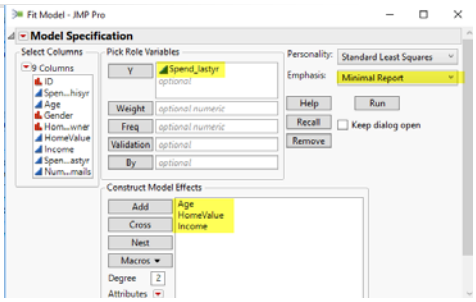
Multiple Regression Demo

Basics using JMP

49

JMP Demonstration

- Use E**commerce** data
  - Predict and explain **amount spend last year** by using multiple variables such as age, income and value of home
- JMP: Analyze > Fit Model > Move Spend_lastyr to Y > Select Age, HomeValue and Income and Click Add > Click Drop-down for Emphasis and change to Minimal report > Run



Rerunning Fit Model with only Age and HomeValue

- JMP: >Analyze > Fit Model > Recall > remove Income > Run
  - Red triangle options next to Response > Estimates > Show prediction expression
  - Right-click in the middle of the Parameter Estimates table > Columns > Std. Betas
  - Red triangle options next to Response > Save Columns> Prediction formula (repeat for mean confidence interval formula, residuals, studentized residuals, etc.).
  - Red triangle options next to Response > Factor Profiling > Profiler

51

Advanced Topics of MR

Handling Categorical Independent (X) Variables Lecture and Demo

1

Categorical Independent (X) Variables

- We can use categorical independent (X) variables (such as Gender: Male, Female) easily in building a MR model.
- Different software use different type of coding for categorical variables
  - Dummy Variable Coding for Gender (such as F=1, M=0)
  - Effects Coding for Gender (such as F= +1, M= -1)
- Model fit, significance etc., do not change due to different coding patterns mentioned above.
  - Only the intercept and the interpretation of the coefficient changes.

2

JMP Demonstration

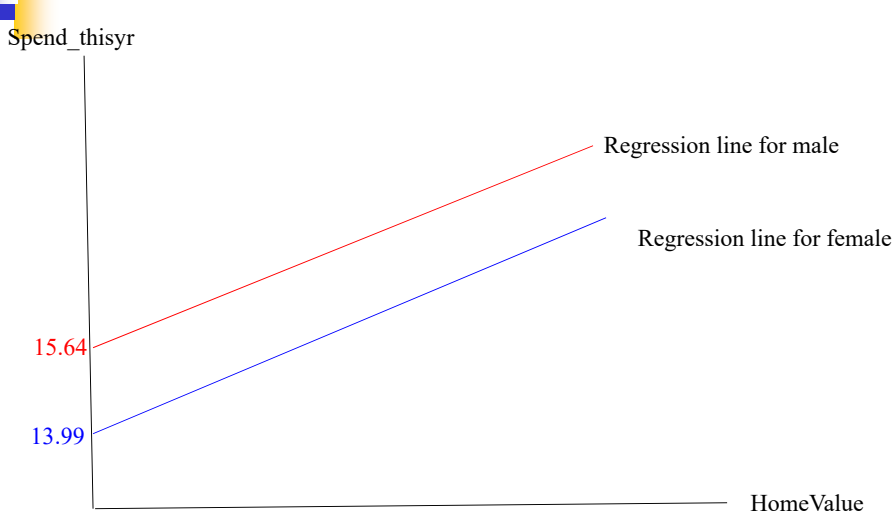
- Use **ecommerce** data to demonstrate and explain:
  - How Gender (male/female) can be used in JMP to model Spend_thisyr
  - What does it mean to have both Gender and Homeval in a MR model to predict/explain spend_thisyr
- JMP: Analyze > Fit Model > Spend_thisyr as Y > Add Gender and HomeValue as Construct Model Effects > Empahsis > Minimal report > Run
  - Click Red Triangle next to Response Spend_thisyr > Estimates > Indicator Parametrization estimates

Mathematics

- Dummy Coding: M=0, F=1 in JMP
- Spend_thisyr= 15.64 -1.65*Gender+ 0.0000093*HomeValue
- So, for males (M=**0**):
  - Spend_thisyr= 15.64 -1.65***0**+ 0.0000093*HomeValue
  - *Spend_thisyr= 15.64 + 0.0000093*HomeValue*
- So, for females (F=**1**):
  - Spend_thisyr= 15.64 -1.65***1**+ 0.0000093*HomeValue
  - *Spend_thisyr= 13.99+ 0.0000093*HomeValue*
- *The difference between the two equations is in the intercept values only!*

4

Diagrammatic View



5

How to Handle Independent Categorical Variables (X) With More than 2 levels?

- Data: Cars 1993
- We want to predict "City Mileage"
- We have an independent categorical variable "Vehicle Category" coded as Compact, Large, Midsize, Small, Sporty and Van.
  - For dummy variable, software will create 1/0 indicator for each level except one
  - Think why?
- JMP: Analyze > Fit Model > City Mileage as Y > Add Vehicle Category as Construct Model Effects > Emphasis > Minimal report > Run
  - Click Red Triangle next to Response City Mileage> Estimates > Indicator Parametrization Estimates

6

Advanced Topics of MR

Handling Variables (X) Selection Lecture

7

Variable Selection Methods

- Basic idea: if we have a very large number of X (i.e., input or independent) variables to choose from, how should we proceed?
- Two naïve approaches:
  - Just run a model with all the X variables, then check which ones are NOT significant and then throw all of those non-significant variables out of the model.
  - Run a model with all the X variables

8

Variable Selection Methods

- Theoretical and managerial relevance/interpretation
  - Control variables (as dictated by theory or managerial relevance) are often included in the model whether of not it is statistically significant!
- Practical issues:
  - Avoid redundancy
  - Avoid irrelevancy

9

A Few Statistical Variable Selection Methods

- Three common statistical variable selection methods:
  - **Forward** selection
  - **Backward** selection
  - **Stepwise** selection
- Some Caveats to keep in mind when using above methods:
  - Warning signs: If coefficients for variables have unexpected signs, or if the signs of coefficients change from one step of selection method to the next step etc.
  - Best practice recommendation: Keep all managerially and/or theoretically important variables even if these are statistically non-significant

10

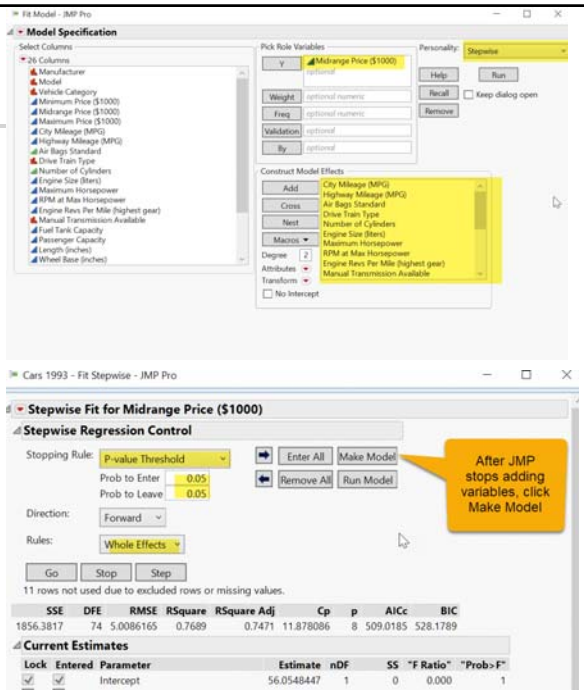
Advanced Topics of MR

Handling Variables (X) Selection Demo using JMP

11

JMP Demonstration

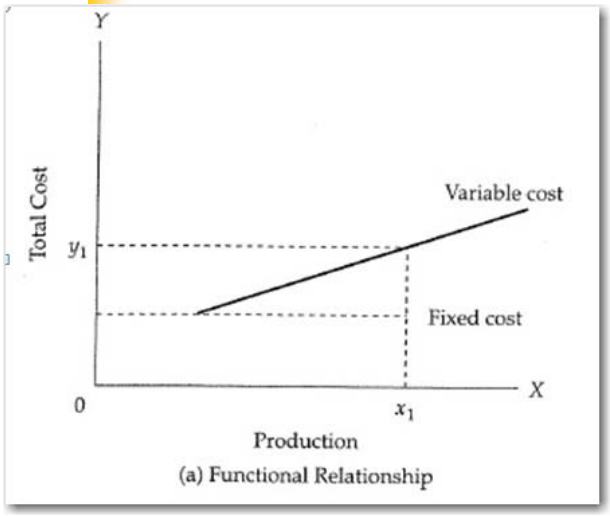
- Use cars 1993 data to show:
  - Correlations among X variables (redundancy in shared information among variables)
  - Building model for Mid-Range Price (Y) with **all** X variables
  - Select X variables using: Forward, Backward and Stepwise
- JMP: Analyze > Fit Model > Mid-range Price as Y > Add all X variables from City Mileage onwards as shown in screenshot as Construct Model Effects > Personality > Stepwise > Run>
- Next screen > select p-value threshold and set both Prob to enter and leave as 0.05> set rules as whole effects > Click step and watch for a few time> then Go > then Make model > Run
- For backward selection in JMP, before you hit Step or Go, first enter all variables in the model.

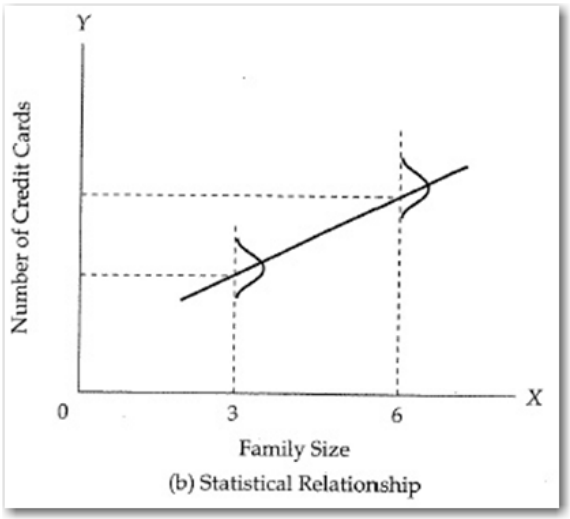


Advanced Topics of MR

Assumptions of Regression Lecture

Functional Versus Statistical Relationships





Classical Regression Assumptions

1. The independent (X) variables are measured **with no error**.
2. The error is a random variable with a mean of zero conditional on the independent (X) variables.
3. The errors are uncorrelated, that is, the correlations between errors is zero.
4. The variance of the error is constant across observations (homoscedasticity).
5. The error is Normally distributed (**needed** for F-test and t-tests)
6. The sample is representative of the population (needed for the **inference prediction**.)
7. The X variables are linearly independent, i.e., it is not possible to express any X as a linear combination of the other X's (No Multicollienarity)
8. X variables **are linearly related** to the Y variable

15

How Important are MR Assumptions?

- Theoretical research
  - Focus is on explaining Y by using X
    - So, the interpretation of coefficient is critical
  - If MR assumptions are satisfied, then the regression coefficients are BLUE (**B**est **L**inear **U**nbiased **E**stimate)
  - If assumptions are not satisfied, then ….
    - Concerns are generally less if only one assumption is violated than if multiple assumptions are violated

16

How Important are MR Assumptions? (contd.)

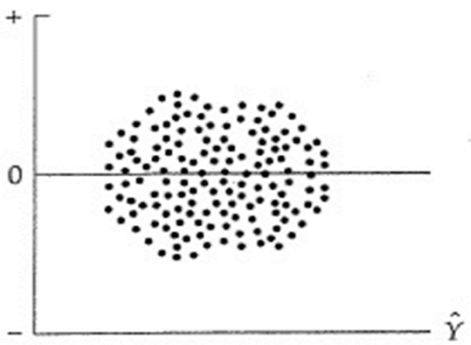
- Practical applications
  - Focus is on predicting Y by using X
    - So, the interpretation of coefficient is less critical
  - **Good news**: Multiple regression is fairly robust and may work well for *predictions* even when some assumptions are violated.

17

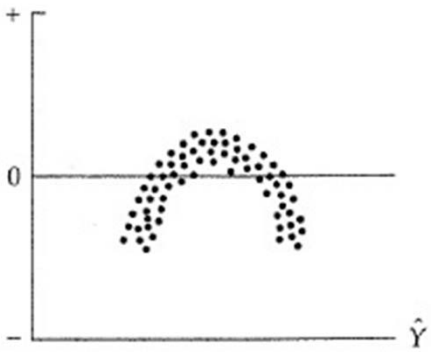
How Do We Check MR Assumptions?

- Assumptions checking starts with the residual (error).
- First, check the plot of residuals against predicted values of Y
  - If you find **no pattern** in the plot, chances are most regression assumptions are satisfied
  - If you find patterns (such as curves, funnel, diamonds, etc.) then may be some of the assumptions are violated. At that point more diagnostics work is needed to figure what exactly is the violation.
  - Some examples follow

Residual Plots



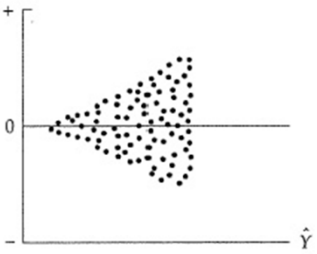
(a) Null plot



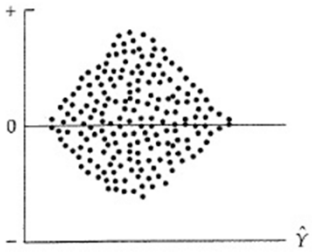
(b) Nonlinearity

19

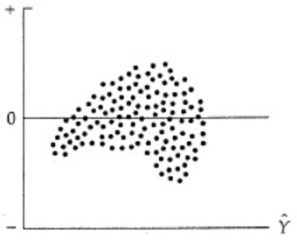
Residual Plots (contd.)



(c) Heteroscedasticity



(d) Heteroscedasticity



(h) Nonlinearity and heteroscedasticity

20

How to Fix violations of Assumptions

- Remedy is do transformations of:
  - X variables such as X-square, logX, square-root of X, etc.
  - Y variable such as Y-square, LogY, square toot of Y, etc.

Advanced Topics of MR

Testing Assumptions of Regression using JMP

22

Which Assumptions Are Tested using Data?

Usually not Tested

1. The independent (X) variables are measured **with no error**.
2. The error is a random variable with a mean of zero conditional on the independent (X) variables.
3. The sample is representative of the population (needed for the **inference prediction**.)

Tested with data

1. The errors are uncorrelated.
2. The variance of the error is constant (homoscedasticity).
3. The error is Normally distributed (**needed** for F-test and t-tests)
4. The X variables are linearly independent, i.e., it is not possible to express any X as a linear combination of the other X's (No Multicollienarity)
5. X variables are linearly related to the Y variable

23

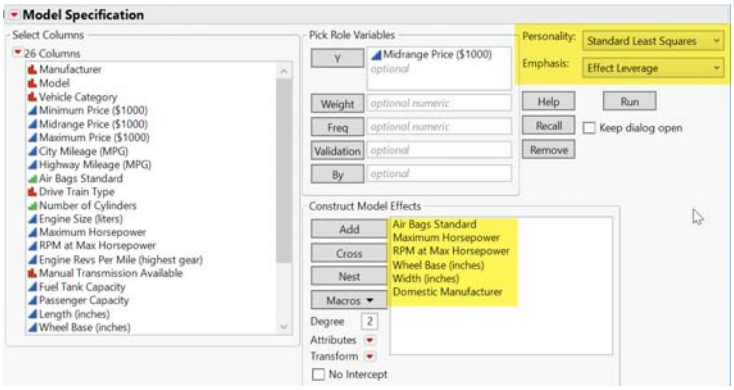
Cars 1993 Data

- Residual plot against predicted may be used to test:
  - The variance of the error is constant (homoscedasticity)
  - **All** X variables are linearly related to the Y variable
- Partial regression plots may be used to test:
  - **Each** X variable is linearly related to the Y variable
- Normal probability plots for residual may be used to test for:
  - The error is Normally distributed
- Sequence plot of residuals (and Durbin-Watson statistic) may be used for:
  - The errors are uncorrelated.
- VIF <10 criteria may be used to test:
  - There is no serious sign of multicollinearity (i.e., The X variables are linearly independent)

24

JMP: Midrange Price as Y, Air Bags, Max HP, RPM at Max HP, Wheelbase, Width and Domestic Manufacturer as X

- Red triangle > Row Diagnostics > Plot studentized residuals for checking heteroscedasticity and linearity of **all** X variables
- Leverage plots are the same as partial regression plots for checking linearity of **each** X variable
- Red triangle > Row Diagnostics > Plot residuals by row for checking autocorrelation
- Red triangle > Row Diagnostics > Durbin Watson statistic for checking autocorrelation



25

JMP for Checking MR Assumptions

- Normality of errors can be checked in two ways:

1. Red triangle > Row Diagnostics > Plot residual by normal quantile for checking normality visually
2. Red triangle > Save Columns> Studentized Residuals
   - Analyze > Distribution > Studentized Residuals as Y
   - Red triangle next to Studentized residuals > Continuous Fit > Normal
   - Red triangle next to Fitted Normal > Diagnostic plot (for checking **normality** visually)
   - Red triangle next to Fitted Normal > Goodness of Fit (for checking **normality** via significance testing, Wilk's Shapiro test)

26

Checking for Multicollinearity

- Right-click in the parameter estimates table > columns > VIF
- VIF < 10 means no serious Multicollinearity because…

The VIF for the i^{th} term, *xi*, is defined as follows:

$$VIF_i = \frac{1}{1 - R_i^2}$$

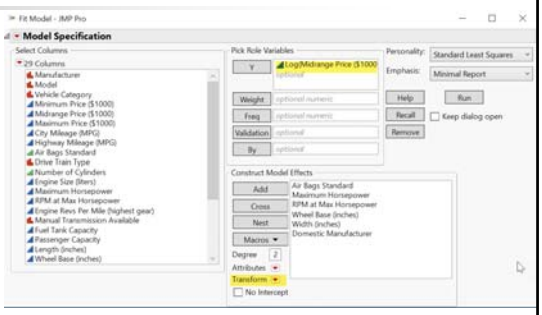
where R_i^2 is the RSquare, or *coefficient of multiple determination*, for the regression of X_i as a function of the other explanatory (X) variables.

27

Transformations of Y: One Example

- JMP: Fit Model > Recall > select Y variable> Transform > Log > run

1. Red triangle > Save Columns> Studentized Residuals
   - Analyze > Distribution > Studentized Residuals as Y
   - Red triangle next to Studentized residuals > Continuous Fit > Normal
   - Red triangle next to Fitted Normal > Diagnostic plot (for checking **normality** visually)
   - Red triangle next to Fitted Normal > Goodness of Fit (for checking **normality** via significance testing, Wilk's Shapiro test)



28