

# 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.

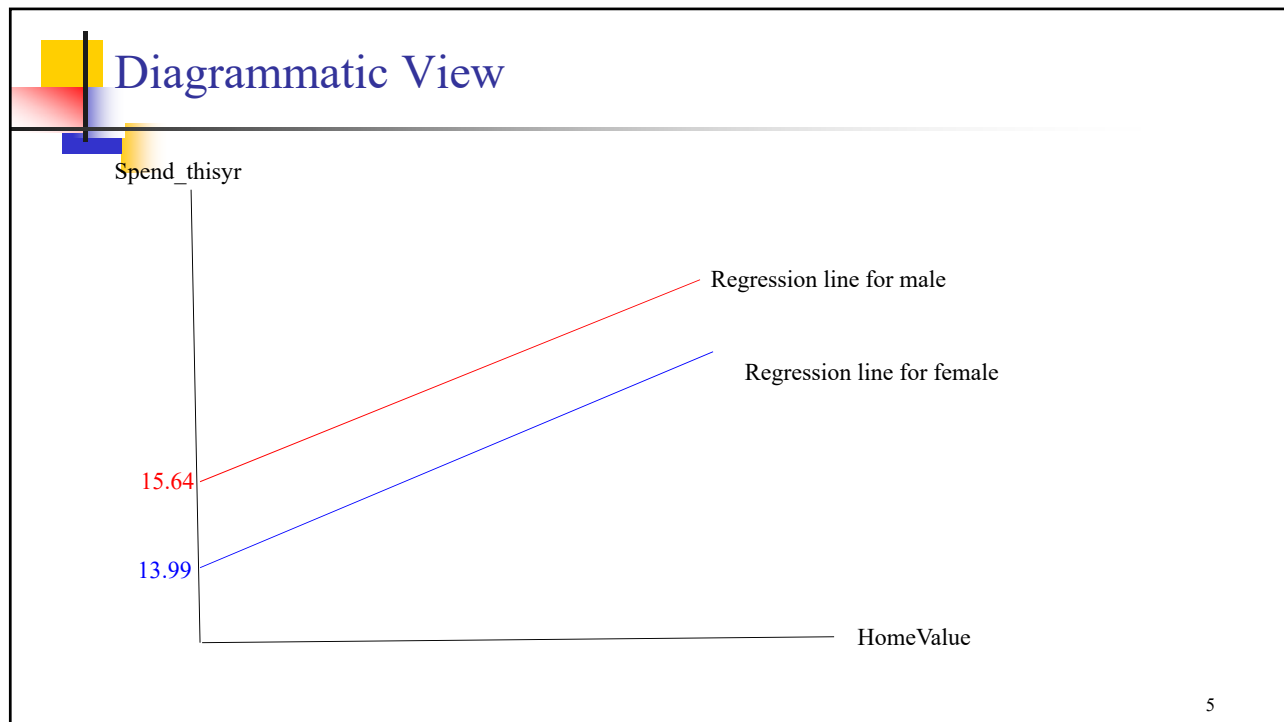
2

## 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 > Emphasis > Minimal report > Run
  - Click **Red** Triangle next to Response Spend\_thisyr > Estimates > Indicator Parametrization estimates

## Mathematics

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



### 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 or 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.

The screenshot displays the JMP Pro interface for fitting a model. The 'Model Specification' window shows 'Midrange Price (\$1000)' as the response variable and various X variables (Manufacturer, Model, Vehicle Category, etc.) as potential predictors. The 'Stepwise Regression Control' window shows the 'Stopping Rule' set to 'P-value Threshold' with 'Prob to Enter' and 'Prob to Leave' both at 0.05. The 'Direction' is set to 'Forward' and 'Rules' are set to 'Whole Effects'. A yellow callout box points to the 'Make Model' button, stating: 'After JMP stops adding variables, click Make Model'.

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
1856.3817	74	5.0086165	0.7689	0.7471	11.878086	8	509.0185	528.1789

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	56.0548447	1	0	0.000	1

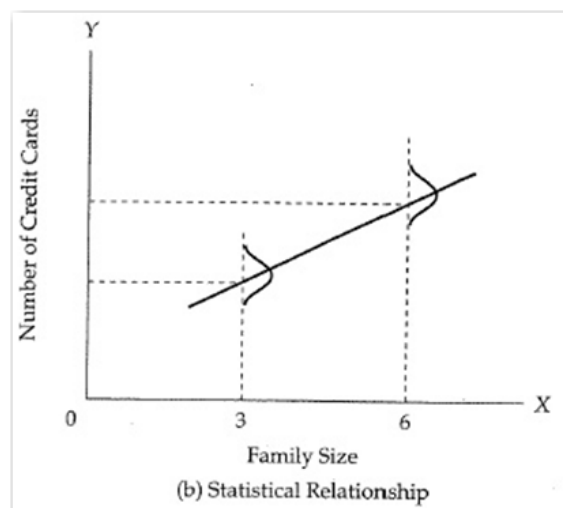
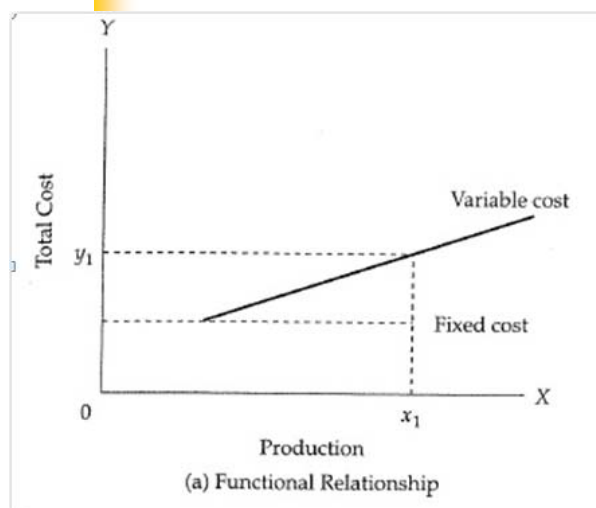
# Advanced Topics of MR

## Assumptions of Regression Lecture



13

## 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 Multicollinearity)
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.)

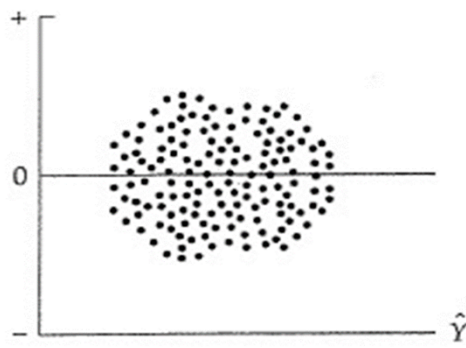
- Practical applications
  - Focus is on predicting Y by using X
    - So, the interpretation of coefficient is less critical
  - **Good news:** Multiple regression is fairly robust and may work well for *predictions* even when some assumptions are violated.

17

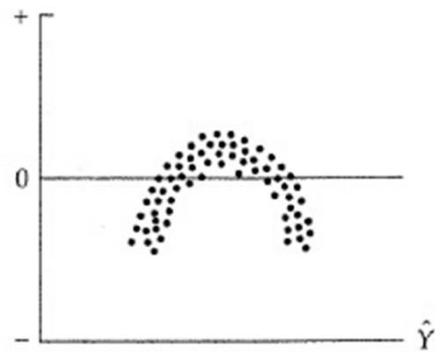
## How Do We Check MR Assumptions?

- 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



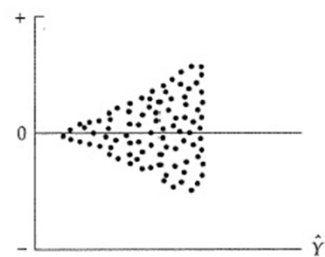
(a) Null plot



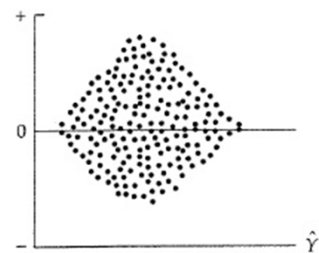
(b) Nonlinearity

19

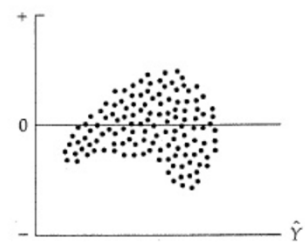
## Residual Plots (contd.)



(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,  $\log X$ , square-root of X, etc.
  - Y variable such as Y-square,  $\log Y$ , square root of Y, etc.



## Advanced Topics of MR

Testing Assumptions of Regression using JMP



## 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

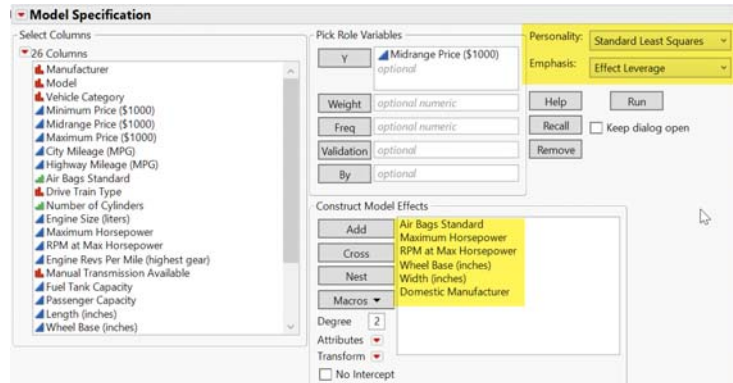
## Cars 1993 Data

- 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...

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

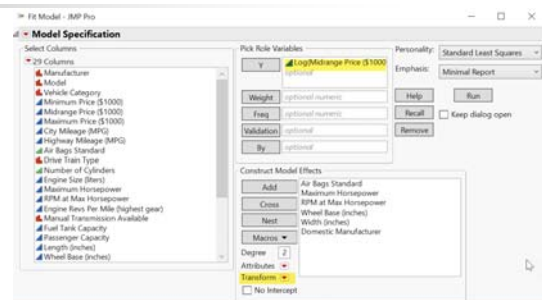
The VIF for the  $i^{th}$  term,  $x_i$ , is defined as follows:

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