Categorical Explanatory Variables

INSR 260, Spring 2009
Bob Stine

Overview

- Review MRM
- Group identification, dummy variables
- Partial F test
- Interaction
- Prediction similar to SRM
- Example (from Bowerman, Ch 4)
  - Sales volume and location

Multiple Regression Model

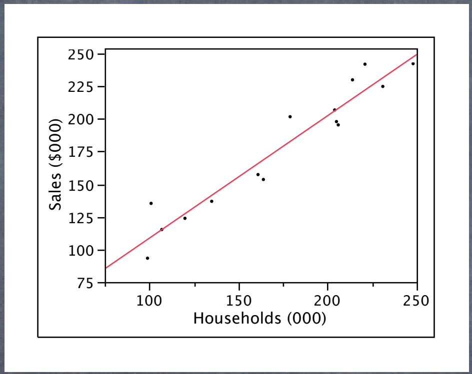
- Equation has **k** explanatory variables
  Mean $E\ Y|X = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k = \mu_{y|x}$
  Observations $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i$
- Assumptions
  - Independent observations
  - Equal variance σ^2
  - Normal distribution around "line"

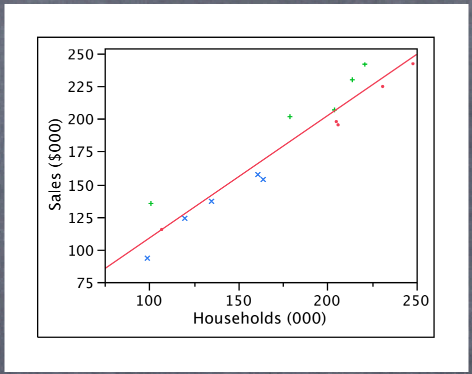
    $y_i \sim N(\mu_{y|x}, \sigma^2)$ $\varepsilon_i \sim N(0, \sigma^2)$
- Issue for this lecture
  How to incorporate categorical explanatory variables that measure group differences.

Example (Table 4.9)

- Context
  - Retailer is studying the relationship between
    - Y = Sales volume in franchise stores, in $1,000
    - X = Number of households near location, in thousands
  - Overall 15 locations, SRM gives

B&W





Color

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	14.867648	13.12805	1.13	0.2779
Households (000)	0.9371196	0.073045	12.83	<.0001*

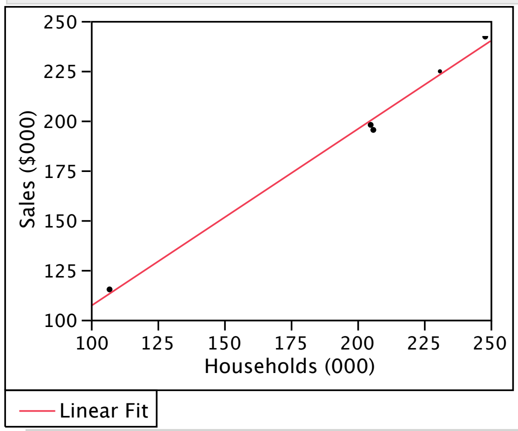
- Question
  - Does the type of location influence the relationship between sales volume and population near the location?
  - Three locations: in mall, suburban, or downtown

Separate Fits

- Question
  - Does the type of location influence the relationship between sales volume and population near the location?
    - Mall, suburban, downtown
    - Five stores from each type of location
  - Are differences important? Statistically significant?

Downtown

Bivariate Fit of Sales ($000) By Households (000)

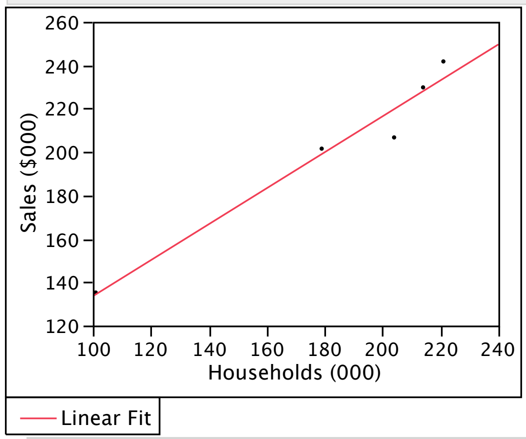


Linear Fit

Sales ($000) = 18.155451 + 0.887074*Households (000)

Mall

Bivariate Fit of Sales ($000) By Households (000)

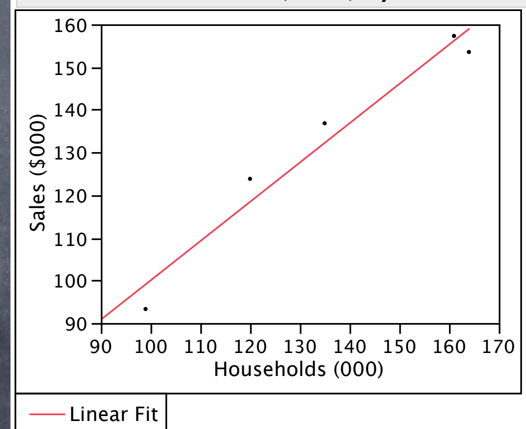


Linear Fit

Sales ($000) = 50.630163 + 0.8289871*Households (000)

Street

Bivariate Fit of Sales ($000) By Households (000)



Linear Fit

Sales ($000) = 7.9004191 + 0.9207038*Households (000)

SRM

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	14.867648	13.12805	1.13	0.2779
Households (000)	0.9371196	0.073045	12.83	<.0001*

Qualitative Variables

- Represent categories using "dummy variables"
  - A 0/1 indicator for each of the categories
  - Redundant: only need 2 dummies for the 3 categories
- Data table
  - JMP software makes the manual creation of dummy variables unnecessary.

Store	Households (000)	Sales ($000)	DM	DD	Location
1	161.00000	157.27000	0	0	street
2	99.000000	93.280000	0	0	street
3	135.00000	136.81000	0	0	street
4	120.00000	123.79000	0	0	street
5	164.00000	153.51000	0	0	street
6	221.00000	241.74000	1	0	mall
7	179.00000	201.54000	1	0	mall
8	204.00000	206.71000	1	0	mall
9	214.00000	229.78000	1	0	mall
10	101.00000	135.22000	1	0	mall
11	231.00000	224.71000	0	1	downtown
12	206.00000	195.29000	0	1	downtown
13	248.00000	242.16000	0	1	downtown
14	107.00000	115.21000	0	1	downtown
15	205.00000	197.82000	0	1	downtown

Regression with Categorical

- Add the dummy variables to the regression…

Summary of Fit

RSquare	0.986846
RSquare Adj	0.983258
Root Mean Square Error	6.349409
Mean of Response	176.9893
Observations (or Sum Wgts)	15

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	14.977693	6.188445	2.42	0.0340*
Households (000)	0.8685884	0.04049	21.45	<.0001*
DD	6.8637768	4.770477	1.44	0.1780
DM	28.373756	4.461307	6.36	<.0001*

- Or simply add the categorical variable itself…

Summary of Fit

RSquare	0.986846
RSquare Adj	0.983258
Root Mean Square Error	6.349409
Mean of Response	176.9893
Observations (or Sum Wgts)	15

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	26.723538	7.194046	3.71	0.0034*
Households (000)	0.8685884	0.04049	21.45	<.0001*
Location[downtown]	-4.882067	2.553028	-1.91	0.0822
Location[mall]	16.627912	2.359355	7.05	<.0001*

- Interpretation of fitted models?
  - By default, JMP handles a categorical explanatory variable differently than with dummy variables.
  - Same fit, but different slope estimates, interpretation.

JMP Fit with Dummy Vars

- Add the dummy variables to the regression...

Summary of Fit

RSquare	0.986846
RSquare Adj	0.983258
Root Mean Square Error	6.349409
Mean of Response	176.9893
Observations (or Sum Wgts)	15

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	14.977693	6.188445	2.42	0.0340*
Households (000)	0.8685884	0.04049	21.45	<.0001*
DD	6.8637768	4.770477	1.44	0.1780
DM	28.373756	4.461307	6.36	<.0001*

- Add categorical variable “indicator parameterization”

Summary of Fit

RSquare	0.986846
RSquare Adj	0.983258
Root Mean Square Error	6.349409
Mean of Response	176.9893
Observations (or Sum Wgts)	15

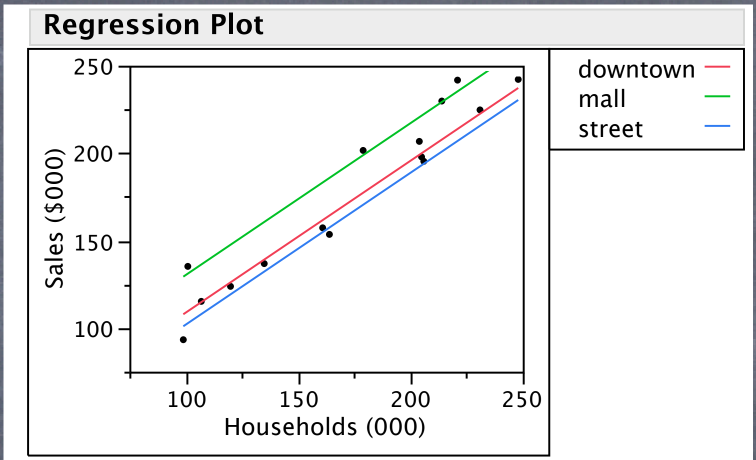
Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	14.977693	6.188445	11.00	2.42	0.0340*
Households (000)	0.8685884	0.04049	11.00	21.45	<.0001*
Location[downtown]	6.8637768	4.770477	11.00	1.44	0.1780
Location[mall]	28.373756	4.461307	11.00	6.36	<.0001*

- Interpretation of fitted models?
  - Slope estimates now match up
  - Still missing that other category

Interpretation

- Plot of fitted model (with categorical variable added) shows fit of the model as 3 parallel lines



- Slopes are shifts (changes in the intercept) relative to the excluded group (street locations)

Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	14.977693	6.188445	11.00	2.42	0.0340*
Households (000)	0.8685884	0.04049	11.00	21.45	<.0001*
Location[downtown]	6.8637768	4.770477	11.00	1.44	0.1780
Location[mall]	28.373756	4.461307	11.00	6.36	<.0001*

Partial F-Test

- Are the differences among intercepts for the locations statistically significant?
  - H_0: $\beta_{downtown} = \beta_{mall} = 0$
  - Test of two coefficient simultaneously
- Partial F-test considers the contribution to the fit obtained by 1 or more explanatory variables
- Two ways to compute test statistic
  - JMP provides "Effect Test" for categorical variable
  - Compare R^2 statistics between the models (then you'll need to obtain the p-value of the test)

$$F = \frac{(\text{Change in } R^2)/(\#\text{ added } x\text{'s})}{(1 - R_{all}^2)/(n-k-1)}$$

Example

- Test H_0: $\beta_{downtown} = \beta_{mall} = 0$
- JMP provides effect test, rejecting H_0

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Households (000)	1	1	18552.427	460.1867	<.0001*
Location	2	2	2024.342	25.1066	<.0001*

- Compare explained variation obtained by two regressions, with and without categorical terms

With

Summary of Fit

RSquare	0.926798
RSquare Adj	0.921167
Root Mean Square Error	13.77793
Mean of Response	176.9893
Observations (or Sum Wgts)	15

Without

Summary of Fit

RSquare	0.986846
RSquare Adj	0.983258
Root Mean Square Error	6.349409
Mean of Response	176.9893
Observations (or Sum Wgts)	15

$$F = \frac{(0.9868-0.9268)/2}{(1-0.9868)/(15-1-3)} \approx 25$$

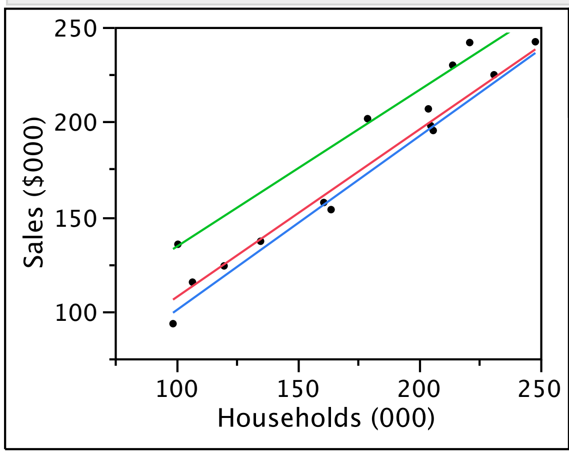
Interaction

- Why assume that the slopes parallel?
  - Why should the relationship between the number of households and sales be the same in the three locations?
- Interaction implies that the slope of an explanatory variable depends on the value of another explanatory variable.
  - Most common interaction: between a categorical and numerical variable. The slope depends upon the group. Slopes in the initial simple regressions are not identical.
  - Can also have interactions between other variables (text)
- An interaction is obtained by adding the product of two explanatory variables.

Fitting an Interaction

- Two approaches
  - Let JMP build the products for you
  - Build products of the dummy and numerical variables and add these to the regression model
- JMP builds this model by "crossing" the number of households with the location

Regression Plot



Summary of Fit

RSquare	0.987657
RSquare Adj	0.9808
Root Mean Square Error	6.799532
Mean of Response	176.9893
Observations (or Sum Wgts)	15

Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	7.9004191	17.03513	9.00	0.46	0.6538
Households (000)	0.9207038	0.123428	9.00	7.46	<.0001*
Location[downtown]	10.255032	21.28319	9.00	0.48	0.6414
Location[mall]	42.729744	21.5042	9.00	1.99	0.0782
Location[downtown]*Households (000)	-0.03363	0.138188	9.00	-0.24	0.8132
Location[mall]*Households (000)	-0.091717	0.14163	9.00	-0.65	0.5334

Mall: $\hat{y}$ = 7.90 + 0.921 Households + 42.73 - 0.092 Households
= 50.63 + 0.829 Households

Testing the Interaction

- Fitted equation with the interaction reproduces original simple regressions for each category:
  Are the slopes really so different?

- Partial F test
  - Test H_0: $\beta_{\text{interaction terms}} = 0$; not significant.

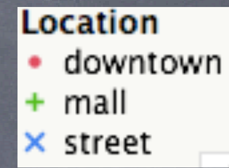
Effect Tests

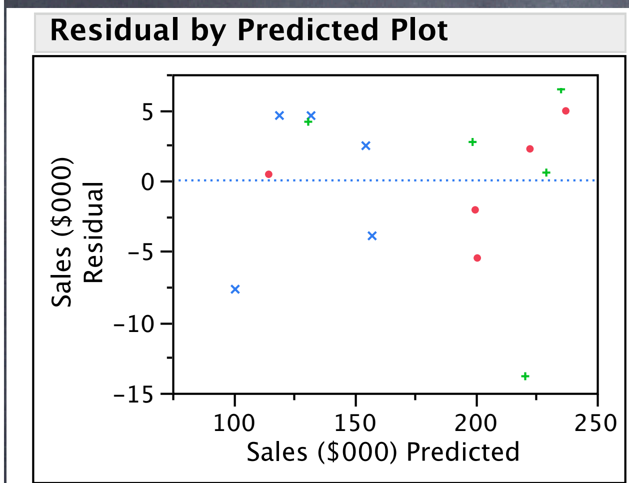
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Households (000)	1	1	13437.839	290.6507	<.0001*
Location	2	2	229.353	2.4804	0.1387
Location*Households (000)	2	2	27.362	0.2959	0.7508

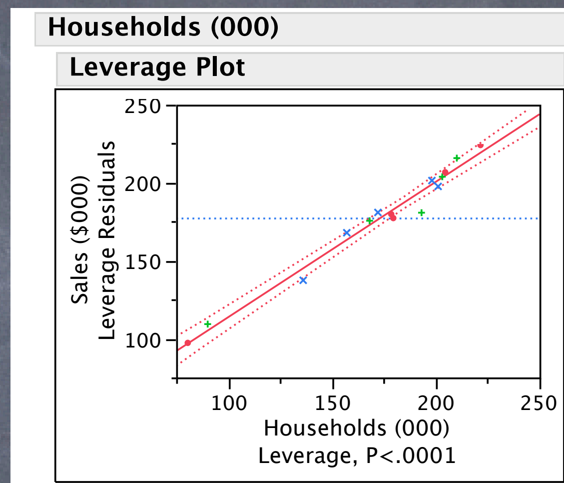
  - Location is not statistically significant when the interaction is present in the fitted model.
  - Typical advice: Remove an interaction that is not statistically significant.
  - Decide status of Location <u>after</u> simplifying model.

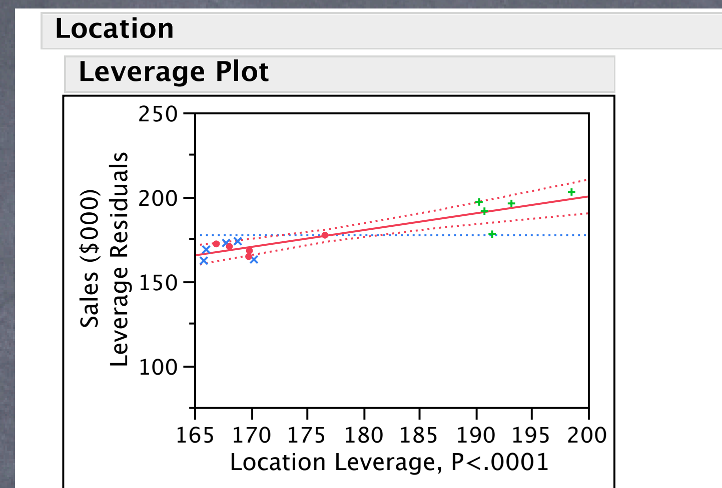
Checking Assumptions

- Usual diagnostic plots
  - Color-coding is very helpful









Least Squares Means Table

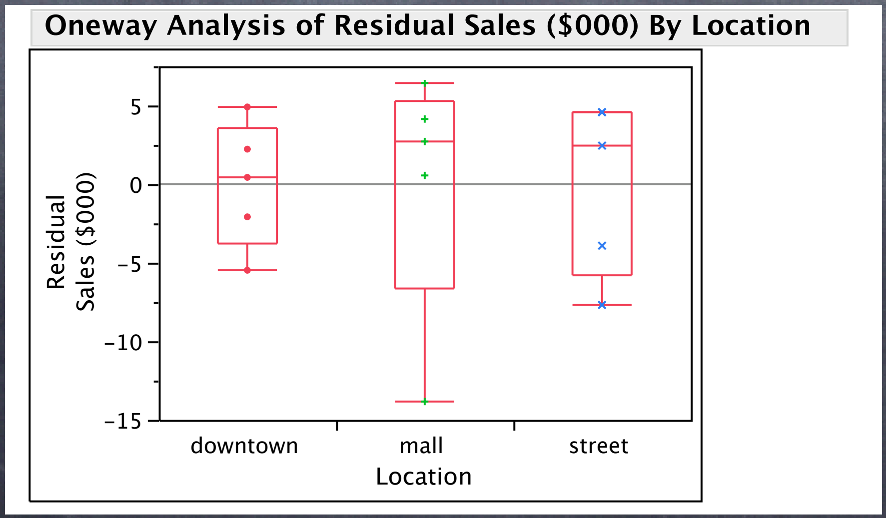
Level	Least Sq Mean	Std Error	Mean
downtown	172.10727	3.0340765	195.038
mall	193.61725	2.8730165	202.998
street	165.24349	3.2142985	132.932

- Least squares means
  - Average of response in each group at the average value of the explanatory variable
  - Handy comparison among groups at common value of explanatory variable

Another Diagnostic

- Why assume that variances of the errors are the same in each group?
  - Slopes, intercepts may be different
  - Why force all 3 groups to have the same RMSE?

- Plot residuals, grouped by category
  - Too few to be definitive in this example (5 in each), but seem similar



Prediction

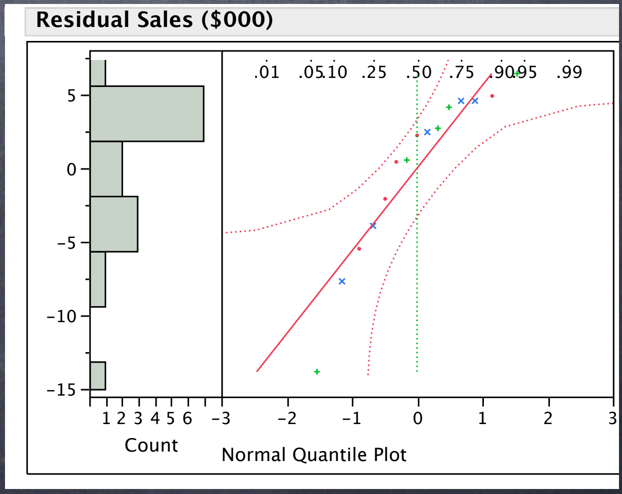
- Use fitted model with number of households, location to predict sales

Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	14.977693	6.188445	11.00	2.42	0.0340*
Households (000)	0.8685884	0.04049	11.00	21.45	<.0001*
Location[downtown]	6.8637768	4.770477	11.00	1.44	0.1780
Location[mall]	28.373756	4.461307	11.00	6.36	<.0001*

- Prediction interval determined by common estimate s^2 and any extrapolation.

- Check the normal quantile plot before rely on normality



Summary

- Distinguishing groups using dummy variables
  - Refer to JMP's "indicator parameterization"
- Partial F test
  - Test a subset of estimates, such as those associated with a categorical variable
- Interaction: slope depends on group
  - Other types of interaction, such as quadratic are described in the text
- Discussion
  Why not fit separate regressions for each group?