St. John's University
Center for Educational Research
Leadership and Accountability

Educational Research and Data Analysis II
EDU 7211

Dr. Francesco Ianni

Regression Analysis

We will learn….

- A little Algebra Review
- The Regression Equation
- Linear Regression
  - The variables are both non nominal
  - One variable is nominal
- SPSS application

Algebra Review

Algebra Review

- The Equation of a line
- The Slope
- The Y-Intercept
  - Examples
  - Practice
- The Correct Window

The Equation of a line from Algebra

- From the algebra books

$y = m(x) + b$

The Slope

The Y - Intercept

The Equation of a line from Statistics

- From the stat books

$y = b(x) + a$

The Slope

The Y - Intercept

The X and the Y variable

- The X variable
  - *Independent variable (Used the most)*
- The Y variable
  - Dependent variable

Example

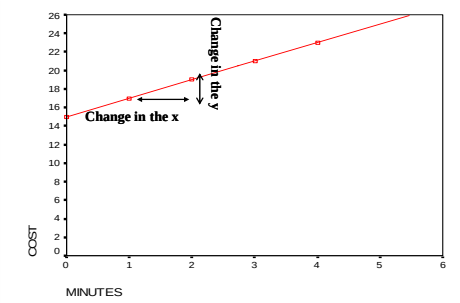
- **Cell phone problem**
  - **2 dollars per minute**
  - **15 dollars monthly fee**
- **The equation will be equal to**

  y = 2x + 15

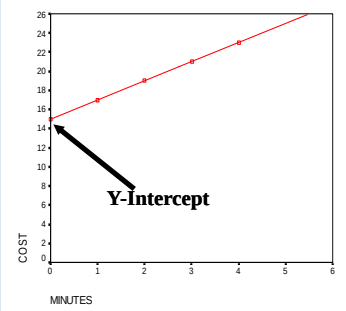
  x = number of minutes

  y = monthly cost

The Graph



The Y-Intercept



How do we use the equation of the line?

- **Given the equation**

  y = 2x + 15
- **We can find:**
  - **The monthly cost given a specific number of minutes.**
  - **The number of minutes we can talk given a specific amount of money we are allowed to spend.**

A numerical example

- Given the equation y = 2x + 15
  - Find the monthly cost if we use 300 minutes per month on average
  - y = 2 (300) + 15 = 615
  - We will spend $615

A numerical example

- Given the equation y = 2x + 15
  - Find the number of minutes we will be talking if the monthly cost has to be $515

    515 = 2x + 15

    515 – 15 = 2x

    500/2 = x

    X = 250

    We will be talking for 250 minutes

Practice

- Given the equation y = 3.6X + 17.2
  1. Find the value of y if x = 13.6
  2. Find the value of x if y = 234.2
  3. Determine the value of the slope and the value of the y intercept

Let's check our answers

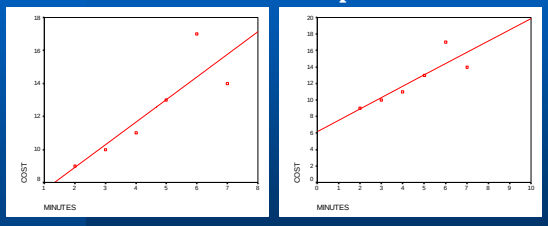
1. y = 3.6 (13.6) + 17.2

   y = 66.16
2. 234.2 = 3.6 x + 17.2

   x = 60.27777
3. The slope is equal to 3.6
4. The y-intercept is equal to 17.2

The Graphs

The Same Graph



SPSS original output **Window Modified**

A quick review

- **The Equation of a line**
- **The Slope**
- **The Y-Intercept**
- **The Correct Window**

The Regression Equation

The Regression Equation

- **Based on the algebra review**

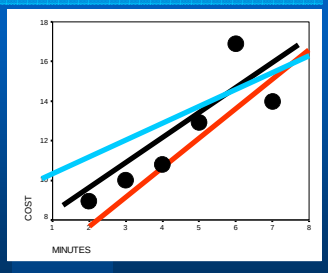
The Regression Equation

- What's the "best fit" line?
- How do we get the Regression Equation?
- The Regression Equation on SPSS
  - Both variables non nominal
- The Regression Equation on SPSS
  - One variable nominal

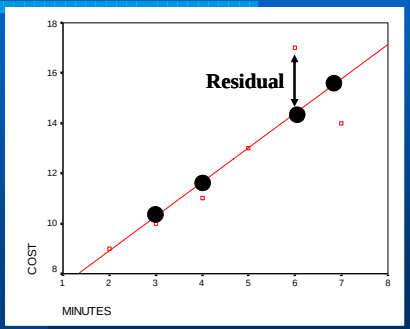
The Best Fit Line

- Given the following data our goal is to find a line that best fits the points
  - X = 2,3,4,5,6,7
  - Y = 9,9,15,9,19,14
- Let's plot the points

The Graph



How do we decide?



The Residual is the difference between the actual value and the predicted value

How do we decide?

- ***Y*** **= actual value**
- $\hat{Y}$ **= Expected value**
- **Residual =** ***Y*** **-** $\hat{Y}$
- **The Best Fitting Line is the line that minimizes the sum of the squared differences (residuals)**

The Regression Equation

$$\hat{Y} = bx + a$$

SLOPE

Y-INTERCEPT

The Linear Regression

The Linear Regression

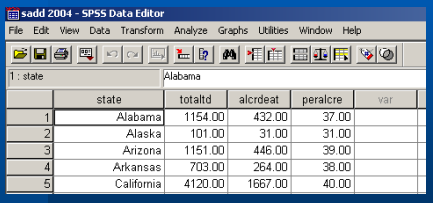
- **Both variables are non nominal**
- **One variable is nominal**

Linear Regression when both variables are non nominal

- **When Both Variable are Non-Nominal**
  - **X is the independent**
  - **Y is the dependent**
- **Create a scatter plot to look at the data**
- **Write the equation using SPSS tables**

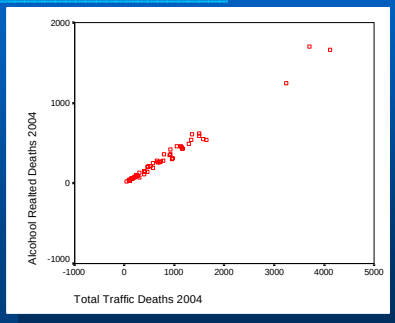
The Linear Regression on SPSS

- **Using the data SADD [on blackboard]**

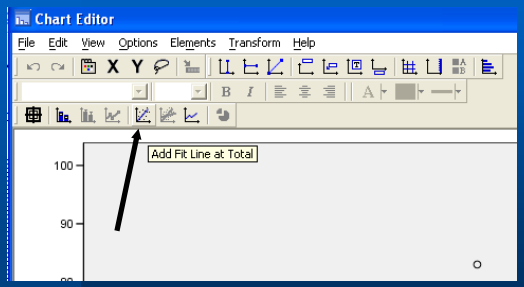


	state	totaltd	alcrdeat	peralcre	var
1	Alabama	1154.00	432.00	37.00	
2	Alaska	101.00	31.00	31.00	
3	Arizona	1151.00	446.00	39.00	
4	Arkansas	703.00	264.00	38.00	
5	California	4120.00	1667.00	40.00	

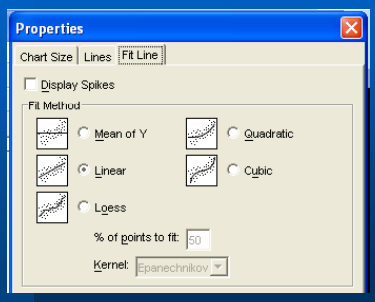
The Linear Regression on SPSS



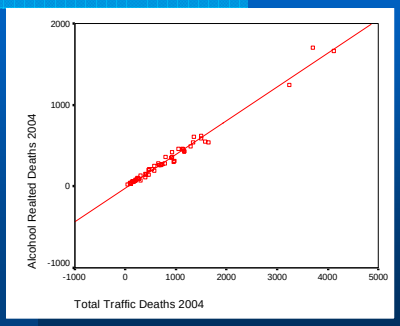
The Linear Regression on SPSS



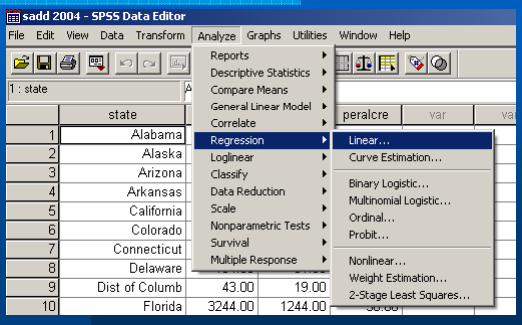
The Linear Regression on SPSS



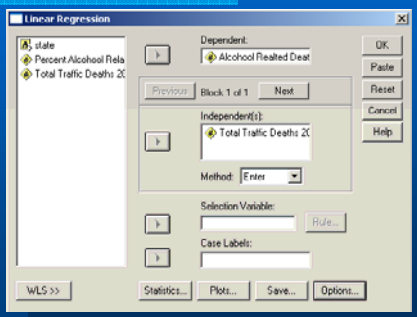
Linear Regression on SPSS



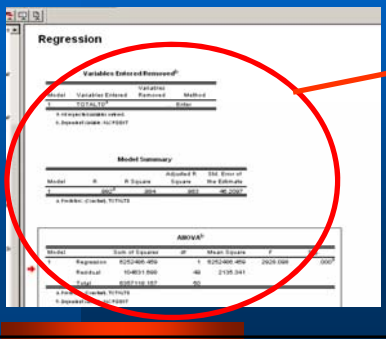
The Linear Regression on SPSS



The Linear Regression on SPSS



The Linear Regression on SPSS



We will keep some tables and get rid of others

The Linear Regression on SPSS

- We will use the following tables
  - Model Summary table

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.992[a]	.984	.983	46.2097

a. Predictors: (Constant), TOTALTD

  - Coefficients table

Coefficients [a]

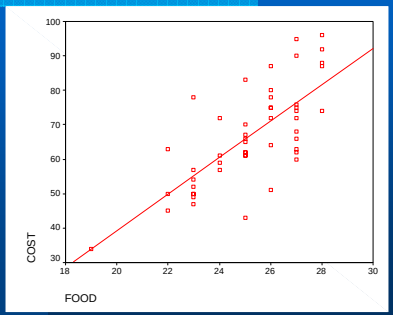
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-15.887	9.118		-1.742	.088
	TOTALTD	.414	.008	.992	54.112	.000

a. Dependent Variable: ALCRDEAT

The Linear Regression on SPSS

- Using Zagat let's look at the relationship between food and cost
  - Create a scatter Plot
  - Create Regression Analysis Tables
  - Interpret the results

Create a scatter plot



Create Regression Analysis Tables

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.734[a]	.539	.529	9.83

a. Predictors: (Constant), FOOD

Coefficients [a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-66.566	17.848		-3.730	.001
	FOOD	5.291	.706	.734	7.493	.000

a. Dependent Variable: COST

The Linear Regression on SPSS

- **The Model Summary**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.734[a]	.539	.529	9.83

a. Predictors: (Constant), FOOD

The X variable or Independent

Absolute value of r

The Linear Regression on SPSS

- **R value in the Model Summary Table**
  - **R is the absolute value of r (correlation coefficient) and represents the simple correlation between Food and Cost.**
  - **In this case .734 is considered to be very strong.**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.734[a]	.539	.529	9.83

a. Predictors: (Constant), FOOD

The Linear Regression on SPSS

- **R Square in the Model Summary Table**
  - **The R Square tells us what percentage of the variation in the dependent variable is explained by the independent variable.**
  - **In this case R Square is .539 which tell us that 53.9% of the variation in Cost can be explained by the variation in Food Rating**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.734[a]	.539	.529	9.83

a. Predictors: (Constant), FOOD

The Linear Regression on SPSS

Correlations

		FOOD	COST
FOOD	Pearson Correlation	1.000	.734**
	Sig. (2-tailed)	.	.000
	N	50	50
COST	Pearson Correlation	.734**	1.000
	Sig. (2-tailed)	.000	.
	N	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

The Values are exactly the same

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.734[a]	.539	.529	9.83

a. Predictors: (Constant), FOOD

The Linear Regression on SPSS

Coefficients [a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-66.566	17.848		-3.730	.001
	FOOD	5.291	.706	.734	7.493	.000

a. Dependent Variable: COST

$$\hat{Y} = bx + a$$

The Linear Regression on SPSS

- **The Equation is**

$\hat{Y}$= 5.291 X – 66.566

- **The Slope value is 5.291**
- **The Y – Intercept is – 66.566**

The Linear Regression on SPSS

- **Interpreting the Slope (b) when both variables are Non Nominal**
  - **Positive Slope**
    - **For every unit increase in X there is a b amount increase in the Y**
  - **Negative Slope**
    - **For every unit increase in the X, there is a b amount decrease in the Y**

The Linear Regression on SPSS

- **Interpreting the Y-Intercept when both variables are Non Nominal**
  - **The Y intercept represents the value of the Y variable when the X = 0**
  - **It is extremely important to put the significance of the Y- Intercept in the context of the problem.**

The Linear Regression on SPSS

- **Let's interpret the Slope and the Y-Intercept**
  $\hat{Y}$ **= 5.291 X – 66.566**
- **Interpretation of the Slope:**
  - **For every unit increase in Food Rating there is an increase in Cost of $5.291**
- **Interpretation of the Y – Intercept:**
  - **The Y – Intercept doesn't make sense in the context of the problem. It is not realistic that the restaurant will give you back $66.566 if the food rating is 0**

The Linear Regression

- **Please Remember that:**
  - **Determining the Value of the Slope or Y- Intercept**
  - **Interpreting the Slope and the Y- Intercept**

ARE TWO DIFFERENT QUESTIONS

The Linear Regression on SPSS

Coefficients [a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-66.566	17.848		-3.730	.001
	FOOD	5.291	.706	.734	7.493	.000

[a]. Dependent Variable: COST

This is the correlation coefficient r

R = absolute value of r

The Linear Regression on SPSS

Coefficients [a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-66.566	17.848		-3.730	.001
	FOOD	5.291	.706	.734	7.493	.000

[a] Dependent Variable: COST

- The Significance Test is the same as the previous chapters
- In this case since the value is <.01 the result is Statically Significant
- The probability that these results would happen by chance is less than 1/1000

Linear Regression when one variable is nominal

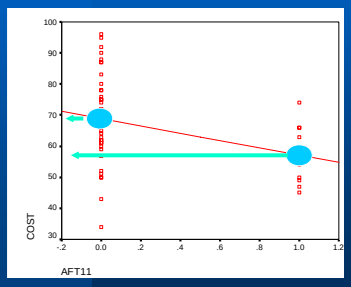
- Determine the values of the categories
- Create a scatter plot to look at the data
- Write the equation using SPSS tables
- Interpret the results

The Linear Regression on SPSS

- Using the Zagat data let's analyze the relationship between
  - Aft11 = Open after 11 o'clock
    - 0 = Restaurant is Closed after 11
    - 1 = Restaurant is Open after 11
      - You obtain this info from ***Variable View Window***
  - Cost = Cost of Dinner

The Linear Regression on SPSS

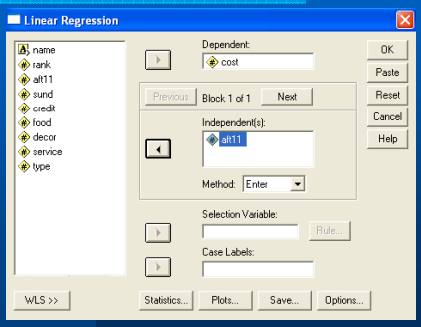
- Scatter Plot



From the scatter plot we noticed that we only have 2 categories.

The restaurants that close after 11 (0) have a higher Cost compared to the ones open.

The Linear Regression on SPSS



We create the regression equation the same way we did before.

The Linear Regression on SPSS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.319[a]	.101	.083	13.73

a. Predictors: (Constant), AFT11

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	68.878	2.145		32.117	.000
	AFT11	-11.767	5.055	-.319	-2.328	.024

a. Dependent Variable: COST

Interpreting the Linear Regression

- **The Summary Table**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.319[a]	.101	.083	13.73

a. Predictors: (Constant), AFT11

r = .319 this is the correlation

This tells us that the correlation is moderate

R Square interpreted the same way

10% of the variation in Cost is explained by the fact that the restaurant is open or closed after 11

Interpreting the Linear Regression

- **The Coefficients Table**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	68.878	2.145		32.117	.000
	AFT11	-11.767	5.055	-.319	-2.328	.024

a. Dependent Variable: COST

$\hat{Y}$ **= -11.767 X + 68.878**

Slope Y - Intercept

Interpreting the Linear Regression

$\hat{Y}$ = -11.767 X + 68.878

Interpreting the Linear Regression

- In general for the interpretation of the slope when the independent variable is Nominal

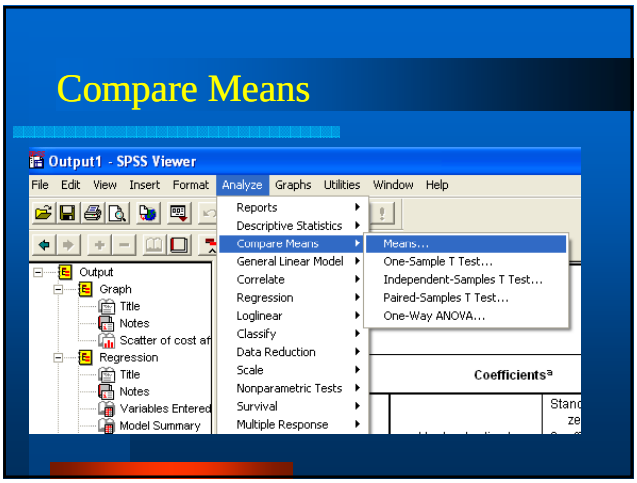
Interpreting the Linear Regression

- In general for the interpretation of the Y-Intercept

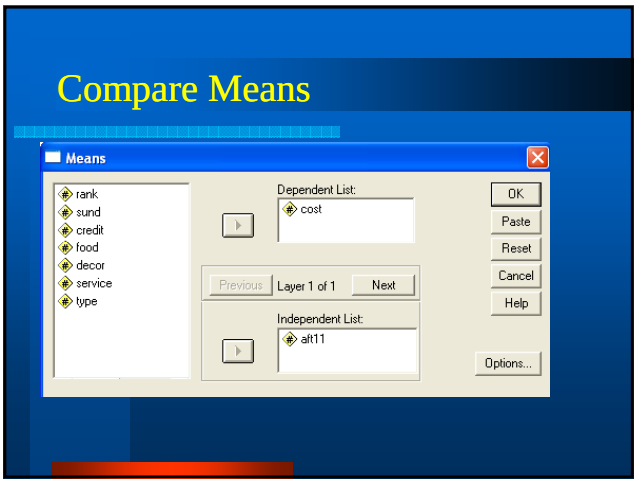
Interpreting the Linear Regression

- Another way to look at the regression
  - Create a Regression analysis
    - Summary Table
    - Coefficient Table
  - Create Compare Means table
  - Compare the above tables

Compare Means



Compare Means



To Verify

Report

COST

AFT11	Mean	N	Std. Deviation
0	68.88	41	14.32
1	57.11	9	10.33
Total	66.76	50	14.34

As we noticed before the Y – Intercept 68.878 is the Cost (on average) for the restaurants that are closed after 11

Coefficients a

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	68.878	2.145		32.117	.000
	AFT11	-11.767	5.055	-.319	-2.328	.024

a. Dependent Variable: COST

To Verify

Report

COST

AFT11	Mean	N	Std. Deviation
0	68.88	41	14.32
1	57.11	9	10.33
Total	66.76	50	14.34

The Slope 11.767 is the difference between the two categories 0 and 1.

68.88 – 57.11 = 11.7

Coefficients a

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	68.878	2.145		32.117	.000
	AFT11	-11.767	5.055	-.319	-2.328	.024

a. Dependent Variable: COST

Conclusion

- **Tables are always created the same way regardless of the fact that:**
  - **Independent Nominal**
  - **Dependent Non Nominal**
  - **Independent Non-Nominal**
  - **Dependent Non-Nominal**

Slope Interpretation

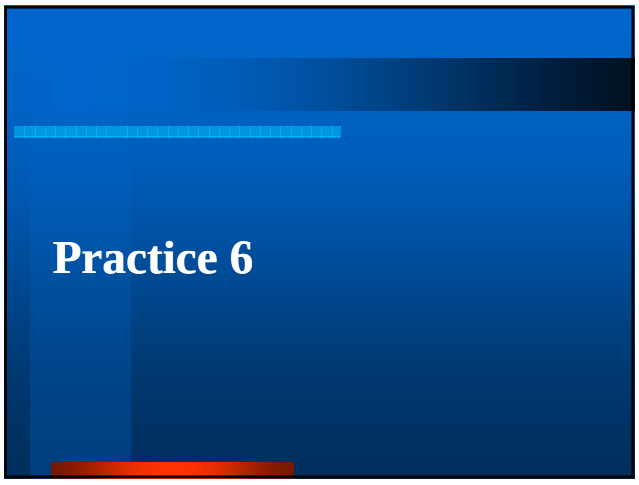
- **If Independent is Non – Nominal**
  - **The Slope represents the rate of change.**
  - **For every unit increase in the independent variable there is b amount increase or decrease (depending if b is positive or negative) in the dependent variable.**
- **If Independent is Nominal**
  - **The Slope represents the difference on average between the categories**

Y-Intercept Interpretation

- **If Independent is Non - Nominal**
  - **The Y- Intercept represents the value of the dependent variable when the independent is 0**
- **If Independent is Nominal**
  - **The Y intercept doesn't have any meaning in the problem if there is no category labeled X = 0**
  - **If there is a category labeled 0 than the Y intercept represents the value of that category**

SPSS Applications

- **The analysis**
  - **Create a scatter plot**
  - **Do a mental estimation of slope and Y Intercept**
  - **Run a linear regression analysis**
  - **Interpret the tables**
  - **Make predictions if the model is reliable**
  - **Write conclusions**


Practice 6