Management Science

Regression

Dan Brandon, Ph.D., PMP

Session Objectives

- Identify variables and use them in a regression model
- Develop simple linear regression equations from sample data and interpret the slope and intercept
- Compute the coefficient of determination and the coefficient of correlation and interpret their meanings
- List the assumptions used in regression and use residual plots to identify problems
- Develop a multiple regression model and use it to predict
- Use dummy variables to model categorical data
- Determine which variables should be included in a multiple regression model
- Transform a nonlinear function into a linear one
- Understand and avoid common mistakes made in the use of regression analysis

Copyright – Dan Brandon

Goals Of Regression Analysis

- To determine if two or more variables are related and to what degree
- In situations where one variable precedes another in time, we maybe able to use regression analysis to predict the value of one variable by knowing the value of the other variable
- **What might be some business variables that are related ?**

Regression and Decision Making

- Business (and personal) decisions are often based on the perceived relationship between two variables
- For example, a manager might decide how much to spend on advertising based on his believe about the relationship between advertising and sales revenue
- You might decide how much time to spend studying based on your believe about the relationship between study time and test grade
- You might spend the time and money to get a college degree based on your believe about the relationship between education and lifetime income

C

Regression Analysis Concepts

- Regression analysis can be used to quantify how variables are related or associated (or "covary" or are "correlated")
- It may be the case that one variable's (x) occurrence precedes the other variable's occurrence (y) in time
- Regression analysis may allow us to *predict* the value of one variable based on the observed value of another
- But we must be careful to note that **'correlation does not imply causality'**
- Therefore, regression analysis *alone* cannot be used to conclude causality – that is, we cannot conclude from regression analysis *alone* that one variable's outcome *caused* another variable's outcome !

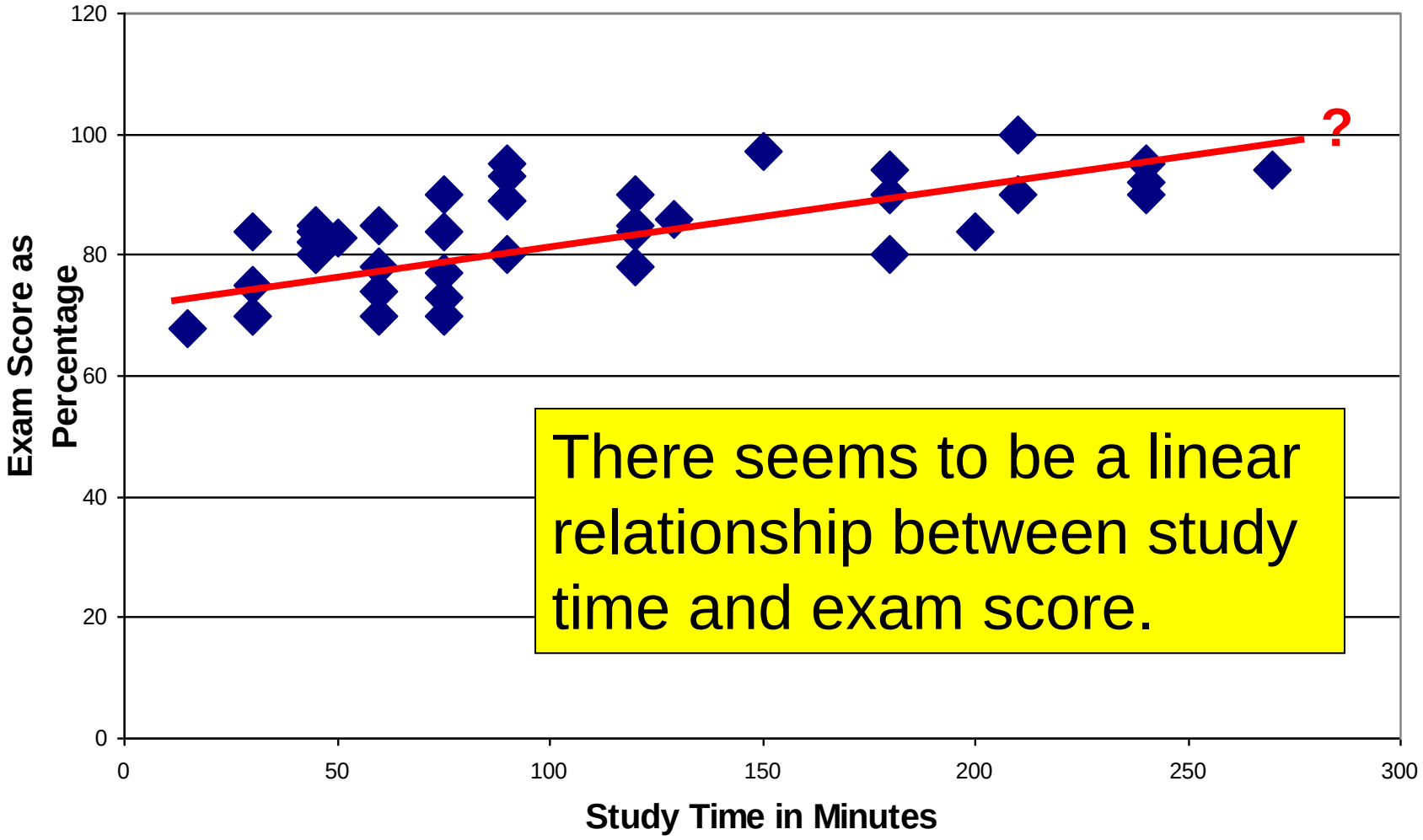
A

Relationship Between Variables

- Let's say that a professor at a college believes there is a *relationship* between the time that a student studies for an exam and that student's exam score
- The first step that should be made is to develop a scatter plot - plotting the values of students' reported study times and their exam scores to see if there appears to be a linear* relationship

* We may ultimately discover that the relationship is not linear but *curvilinear*

Study Time vs. Exam Score



There seems to be a linear relationship between study time and exam score.

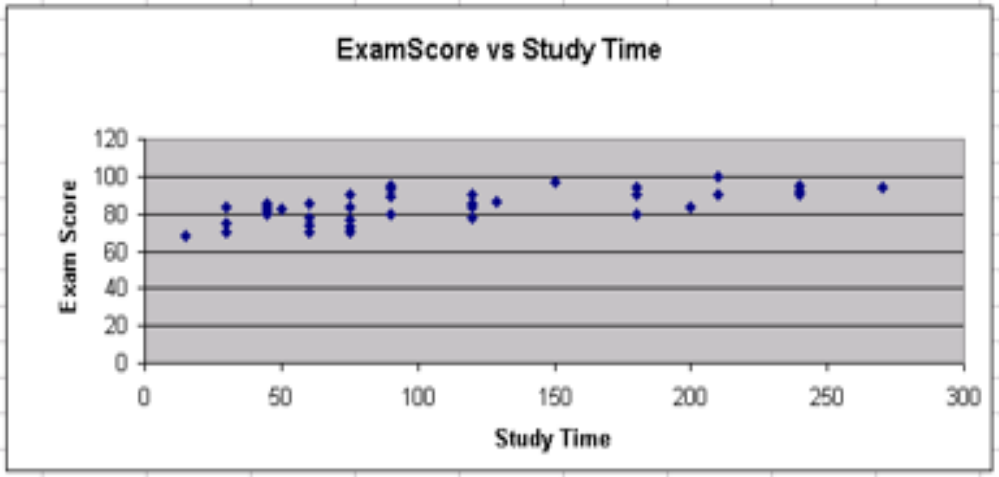
Scatter Plots in Excel

	A	B	C	D	E	F	G	H	I	J
1	Student	StudyTimeMin	ExamScore							
2	1	60	85							
3	2	45	80							
4	3	50	83							
5	4	75	90							
6	5	120	78							
7	6	180	94							
8	7	240	90							
9	8	45	84							
10	9	30	70							
11	10	90	93							
12	11	120	90							
13	12	180	90							
14	13	210	90							
15	14	240	92							
16	15	90	89							
17	16	75	77							
18	17	90	80							
19	18	120	85							
20	19	240	95							
21	20	270	94							
22	21	15	68							
23	22	30	75							
24	23	45	85							
25	24	90	89							
26	25	75	73							

Completed Excel Scatter Chart

[insert -> charts -> scatter (x,y)]

	A	B	C
1	Student	StudyTimeMin	ExamScore
2	1	60	85
3	2	45	80
4	3	50	83
5	4	75	90
6	5	120	78
7	6	180	94
8	7	240	90
9	8	45	84
10	9	30	70
11	10	90	93
12	11	120	90
13	12	180	90
14	13	210	90
15	14	240	92
16	15	90	89
17	16	75	77
18	17	90	80
19	18	120	85
20	19	240	95
21	20	270	94
22	21	15	68



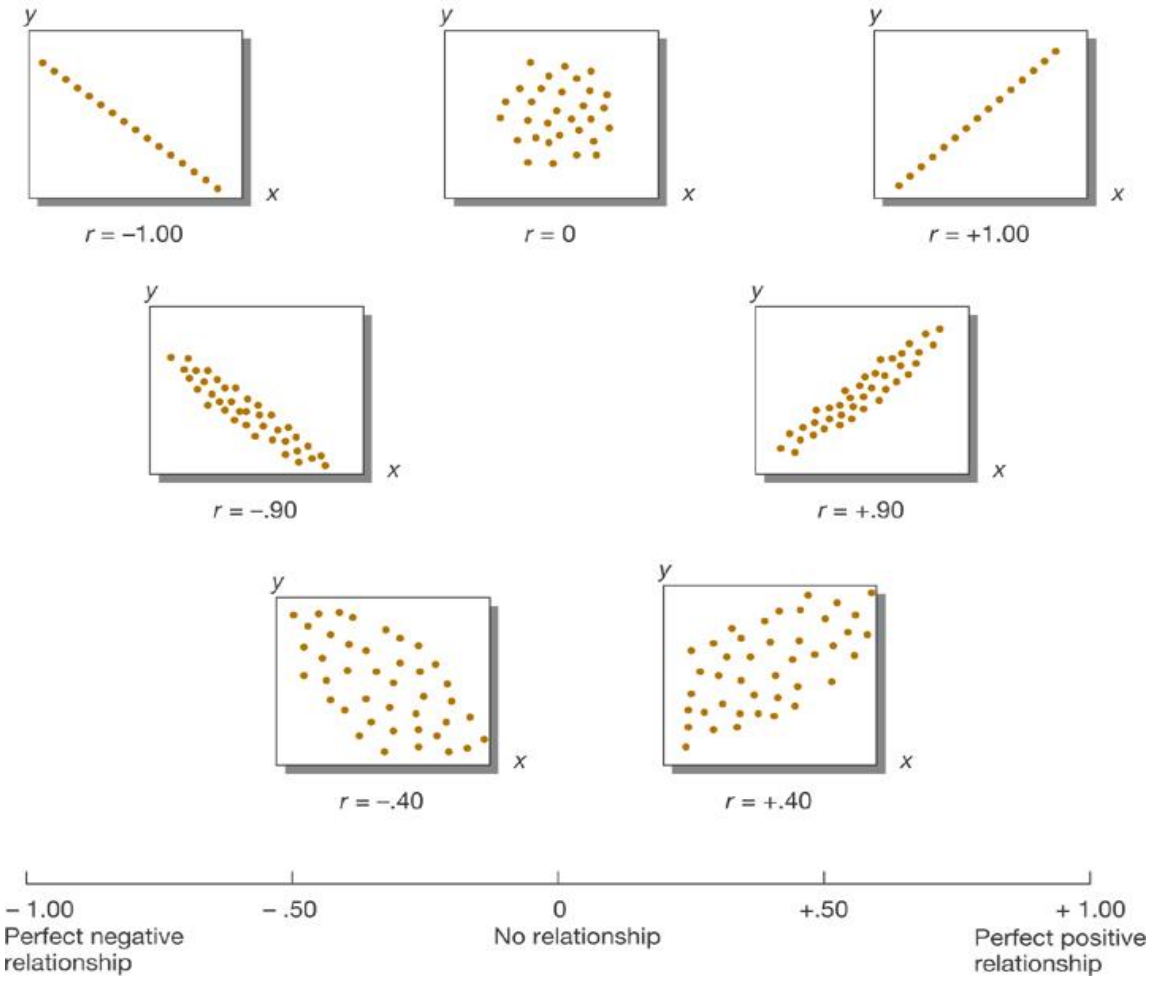
The Linear Correlation Coefficient

- Although the scatter plot indicates that there is somewhat of a linear relationship between study time and exam score, we can use a more precise measure to quantify the degree of association
- The linear correlation coefficient (r) measures the strength of the linear relationship between the metric values in the sample
- The linear correlation coefficient is also referred to as the Pearson product moment correlation coefficient after Karl Pearson, the person who developed this measure - Some just call it "*Pearson's r*"

Pearson's r

- Pearson's r measures the strength of the association between two variables
- Its value can be any number from -1 (indicating a perfect, negative correlation) to +1 (indicating a perfect, positive correlation)
  - See graphical examples on a following slide →

Scatterplots of Relationships and r Values



An example:

	Student	StudyTime Min	ExamScore
1	1	60	85
2	2	45	80
3	3	50	83
4	4	75	90
5	5	120	78
6	6	180	94
7	7	240	90
8	8	45	84
9	9	30	70
10	10	90	93
11	11	120	90
12	12	180	90
13	13	210	90
14	14	240	92
15	15	90	89
16	16	75	77
17	17	90	80
18	18	120	85
19	19	240	95
20	20	270	94
21	21	15	68
22	22	30	75
23	23	45	85
24	24	90	89
25	25	75	73
26	26	60	78
27	27	60	74
28	28	120	84
29	29	45	80

Data View | Variable View

start | Untitled - SP

Let's say this is the data set that contains 40 students' reported study times and exam scores

Let's refer to study time as 'x' and exam score as 'y'

The Calculation of r (Part 1)

First, we calculate $\Sigma(x)(y)$, Σx^2, and Σy^2

Measuring the Association between study time and exam score

StudyTime in Minutes (x)	Exam Score as Percentage (y)	(x) * (y)	x^2	y^2
60	85	5,100	3,600	7,225
45	80	3,600	2,025	6,400
200	84	16,800	40,000	7,056
210	100	21,000	44,100	10,000
4,354	3,367	380,809	666,416	286,001

Skipped rows

The Calculation of r (con't)

Next, we calculate r - the (standardized) correlation coefficient

$$r = \frac{n\,(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{n\,(\Sigma x^2) - (\Sigma x)^2} * \sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

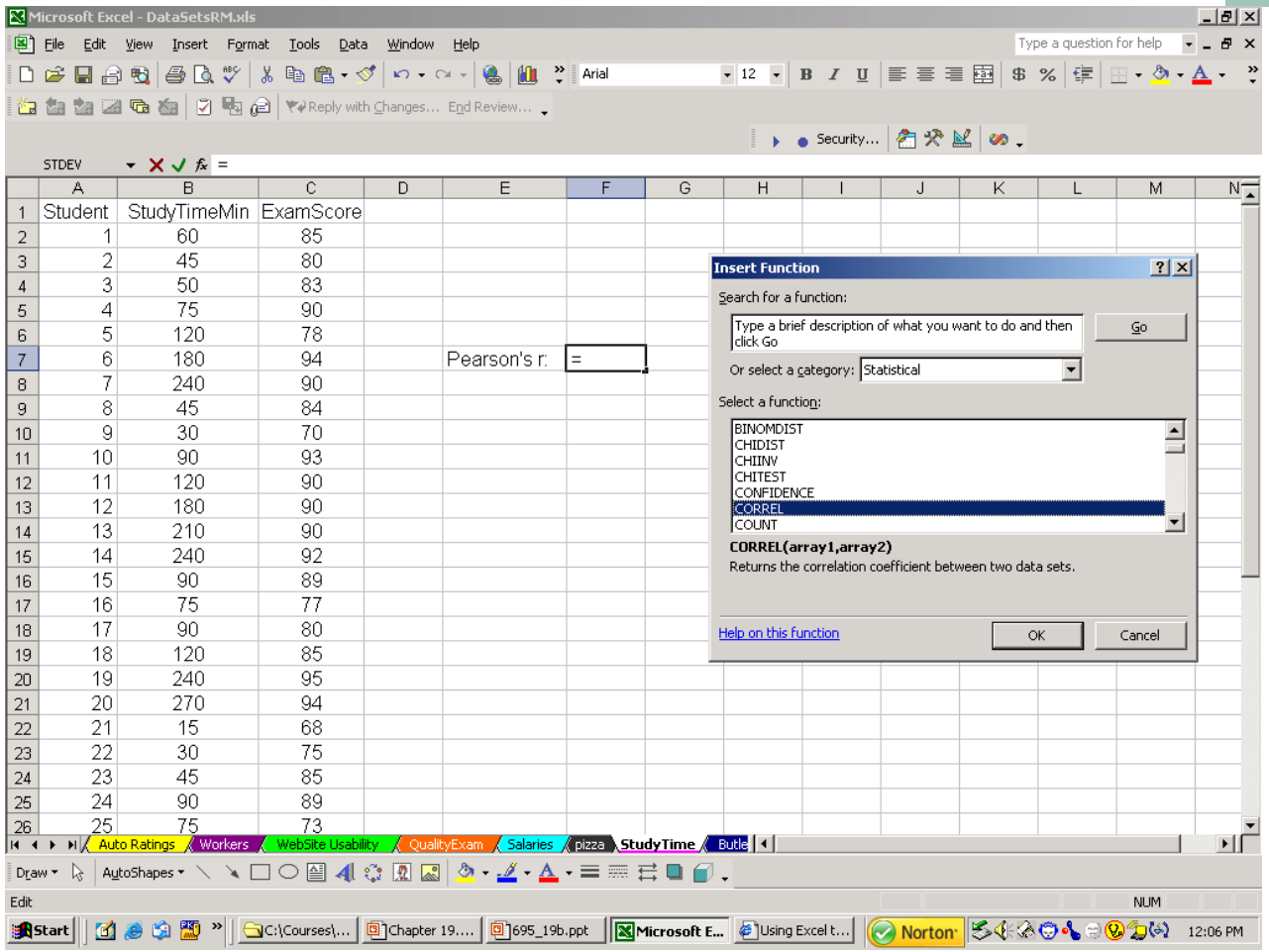
$$r = \frac{40\,(380{,}809) - (4{,}354)(3{,}367)}{\sqrt{40\,(666{,}416) - (4{,}354)^2} * \sqrt{n(286{,}001) - (3{,}367)^2}}$$

r = + .642 r measures the strength of the association

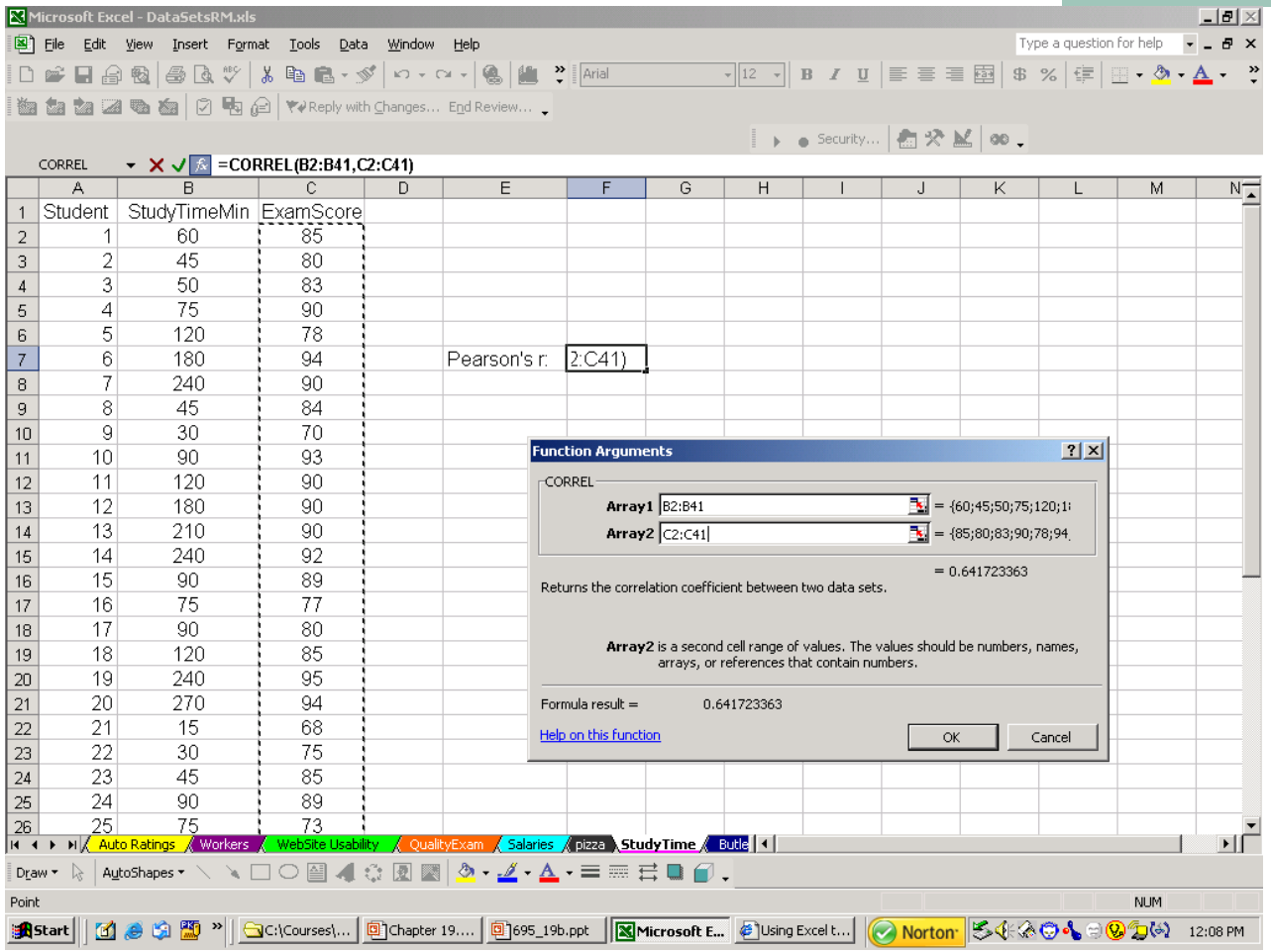
Interpreting r

- The general rule is that the absolute value of r should be > .7 if it is to be considered significant
- But the specific ‘significance’ criteria is related to sample size
  - The smaller the sample size, the larger r must be to be considered significant

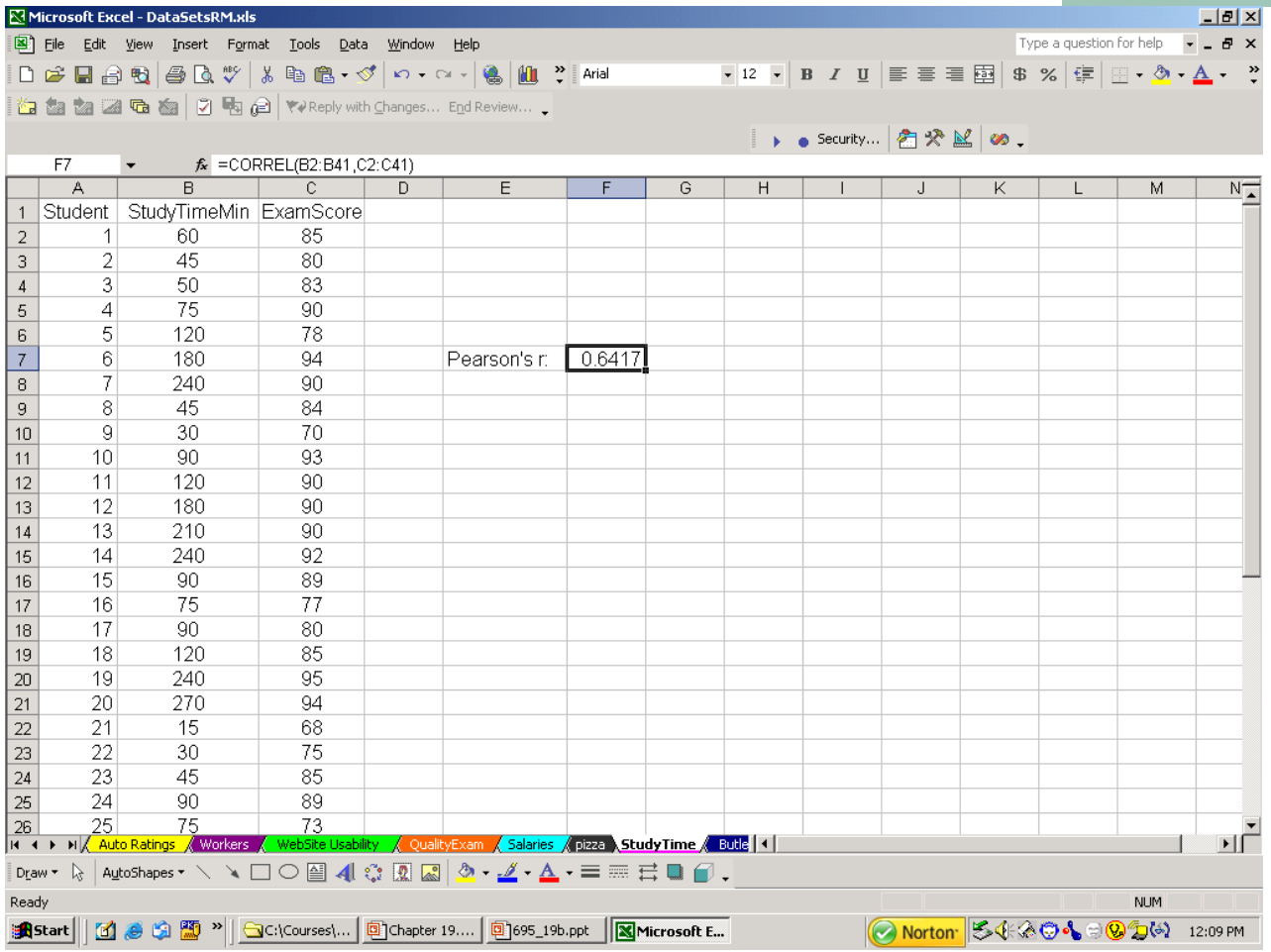
Pearson's r (correlation) in Excel



Using correlation function:

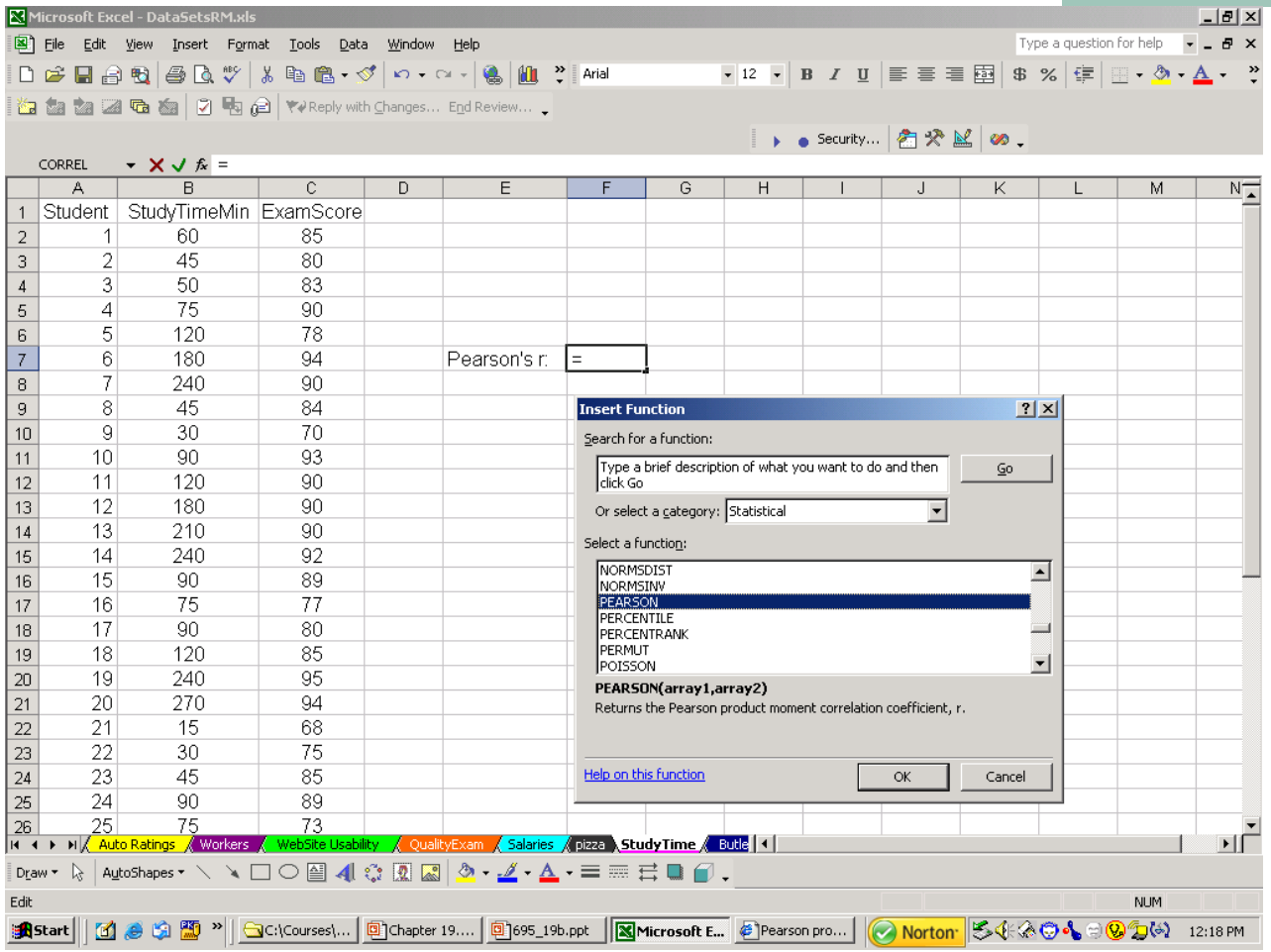


Result of function calculation:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Student	StudyTimeMin	ExamScore											
2	1	60	85											
3	2	45	80											
4	3	50	83											
5	4	75	90											
6	5	120	78											
7	6	180	94		Pearson's r:	0.6417								
8	7	240	90											
9	8	45	84											
10	9	30	70											
11	10	90	93											
12	11	120	90											
13	12	180	90											
14	13	210	90											
15	14	240	92											
16	15	90	89											
17	16	75	77											
18	17	90	80											
19	18	120	85											
20	19	240	95											
21	20	270	94											
22	21	15	68											
23	22	30	75											
24	23	45	85											
25	24	90	89											
26	25	75	73											

Or Using the Pearson function:



Pizza Sales – What is the R value ?

	A	B	C
1	PizzaStore	CampusSize	Sales
2	1	2	58
3	2	6	105
4	3	8	88
5	4	8	118
6	5	12	117
7	6	16	137
8	7	20	157
9	8	20	169
10	9	22	149
11	10	26	202

Independent variable ?

Draw a scatter plot also…

Copyright – Dan Brandon

Wait….

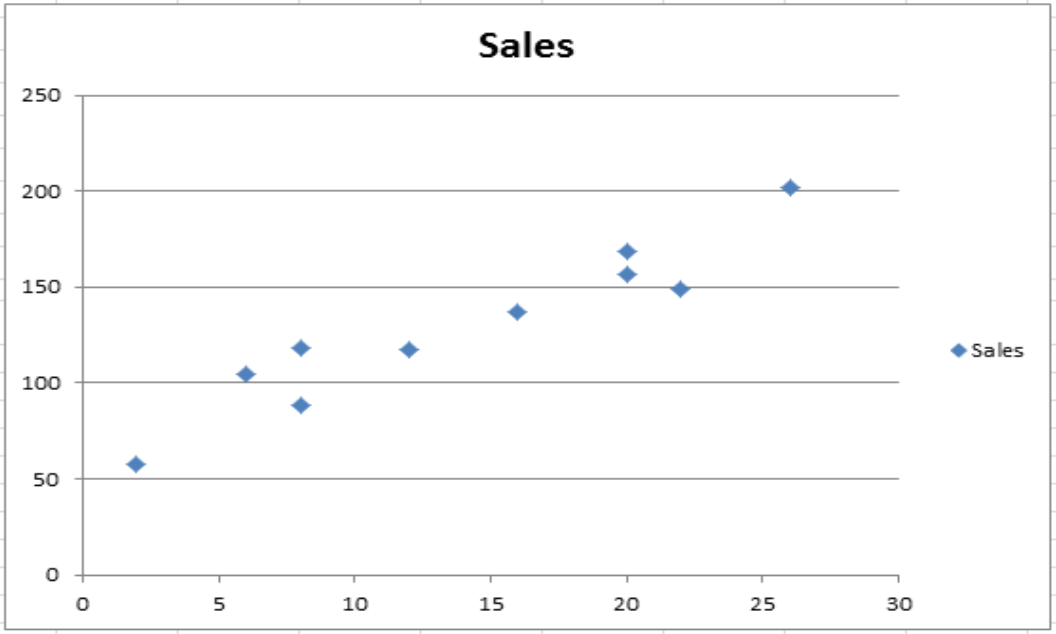
?

Don’t look ahead, until you have your answer !

Pizza Sales (con't)

B17 =CORREL(B2:B11,C2:C11)

	A	B	C
1	PizzaStore	CampusSize	Sales
2	1	2	58
3	2	6	105
4	3	8	88
5	4	8	118
6	5	12	117
7	6	16	137
8	7	20	157
9	8	20	169
10	9	22	149
11	10	26	202
17	R=	0.950122955	



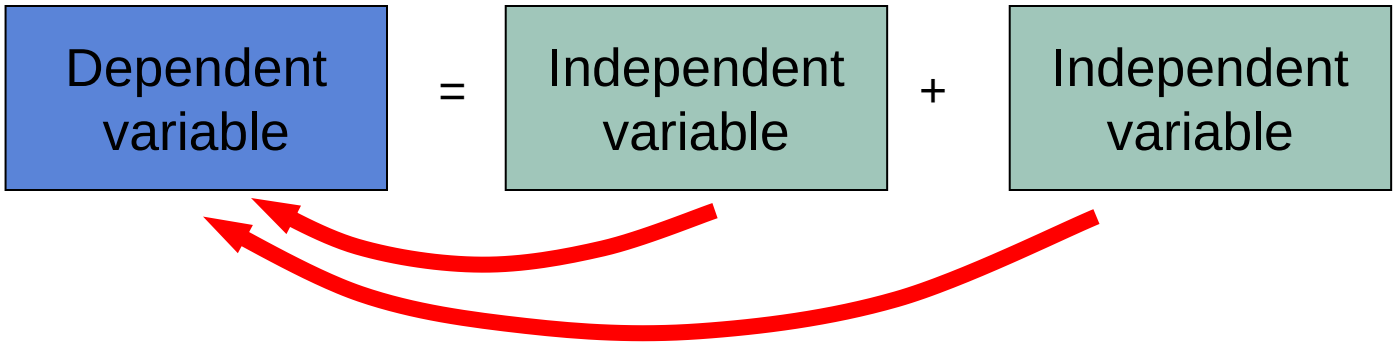
Copyright – Dan Brandon

Simple Linear Regression

- Simple Linear Regression (SLR) allows us to explain and possibly predict the effect of one variable upon another
- For example one might want to analyze the effect of local unemployment rate on the local crime rate
- In this example the local crime rate is called the “dependent” variable (it depends on the other), and the unemployment rate is called the independent variable
- “Simple” means there is only one independent variable that will be considered

Independent & Dependent Variables

- The variable to be predicted is called the *dependent variable*
  - Sometimes called the *response variable*
- The value of this variable depends on the value of the *independent variable*
  - Sometimes called the *explanatory* or *predictor variable*



Sales and Advertising Data for Appeglo

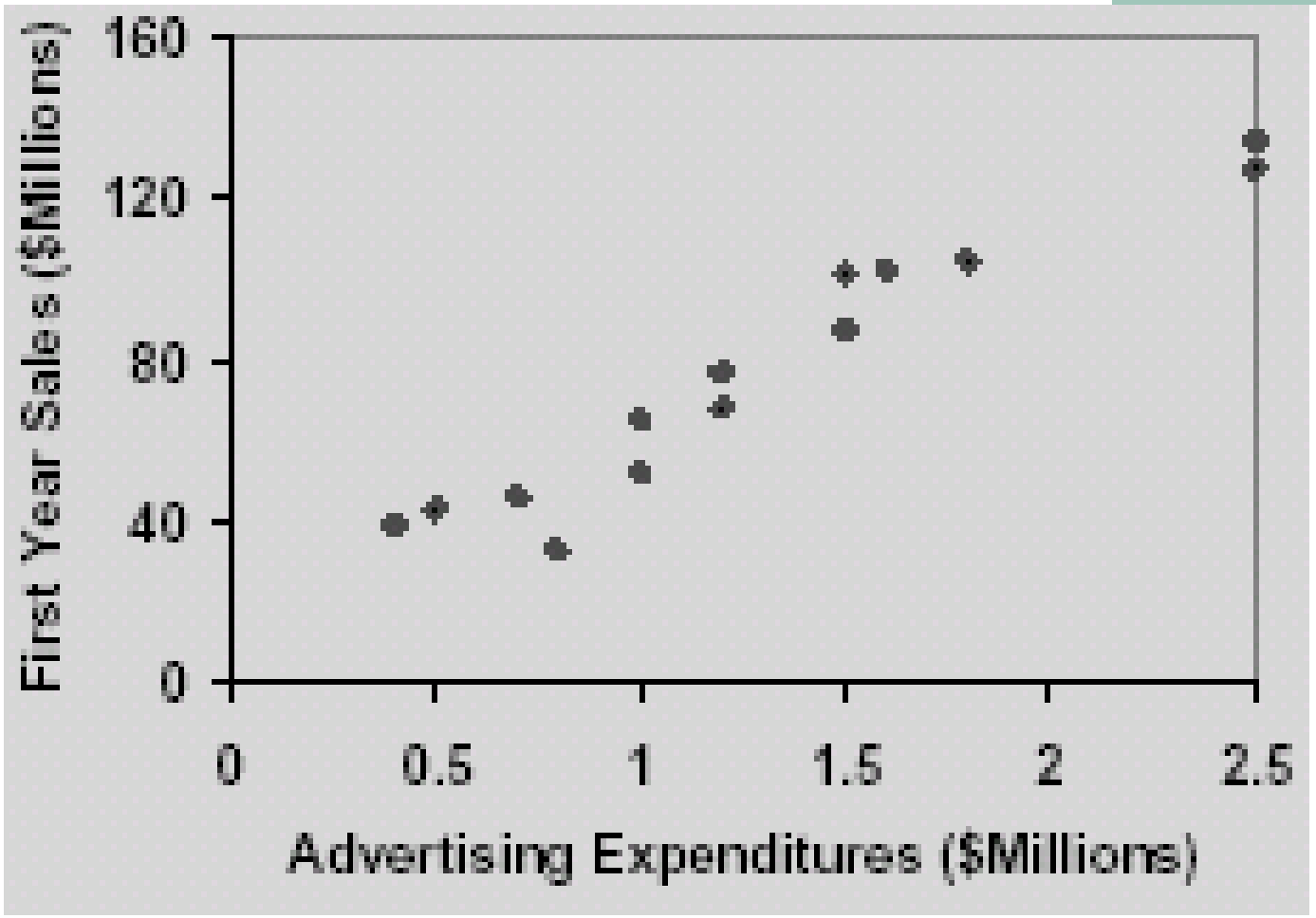
Appleglo	First-Year Advertising Expenditures ($ millions)	First-Year Sales ($ millions)
Region	x	y
Maine	1.8	104
New Hampshire	1.2	68
Vermont	0.4	39
Massachusetts	0.5	43
Connecticut	2.5	127
Rhode Island	2.5	134
New York	1.5	87
New Jersey	1.2	77
Pennsylvania	1.6	102
Delaware	1.0	65
Maryland	1.5	101
West Virginia	0.7	46
Virginia	1.0	52
Ohio	0.8	33

What is the independent variable, the dependent variable?

SLR Questions

- Business questions:
  - What is the relationship between advertising expenditure and sales ?
  - How much can we increase sales by a certain increase in advertising budget ?
  - How confident are we in our above estimate ?
- Here the dependent variable is sales and the independent variable is the advertising budget (sales depends upon advertising)

“Scatter Plot” of Sales versus Advertising



Simple Linear Regression

- True values for the slope and intercept are not known so they are estimated using sample data

$$\hat{Y} = b_0 + b_1 X$$

where

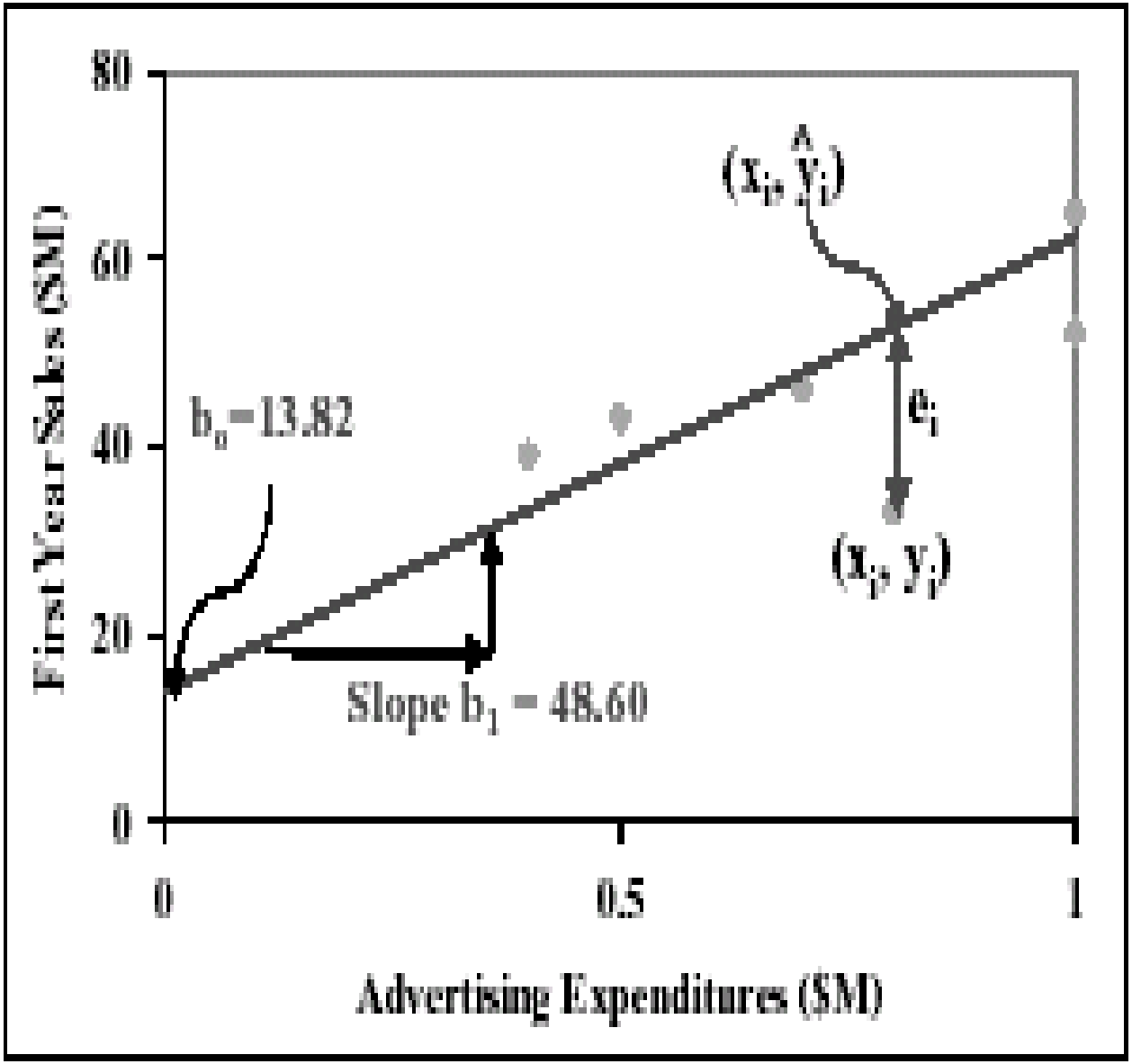
$\hat{Y}$ = dependent variable (response)

X = independent variable (predictor or explanatory)

b_0 = intercept (value of Y when X = 0)

b_1 = slope of the regression line

How do we choose the line that "best" fits the data?



Best choices:

$b_0 = 13.82$

$b_1 = 48.60$

Prediction Formula

- The “prediction formula” here is:
  - Sales = 13.82 + 48.6 * Advertising
- Thus a dollar of advertising adds $48.60 of sales
- We would have 13.82 million in sales even if we did not advertise
- The best fit regression line is the one that minimizes the sum of the distances of the data points from that line
- Or minimizes the sum of the squares of the difference between the dependent variable sample values (y_i) and the predicted value using the equation of the line
  - This formula is called the SS_{error} or “SSE”

Example: Study Time

Recall our data on the ‘studyTime’ worksheet

	A	B	C	D	E
1	Student	StudyTimeMin	ExamScore		
2	1	60	85		
3	2	45	80		
4	3	50	83		
5	4	75	90		
6	5	120	78		
7	6	180	94		
8	7	240	90		
9	8	45	84		
10	9	30	70		
11	10	90	93		
12	11	120	90		
13	12	180	90		
14	13	210	90		
15	14	240	92		
16	15	90	89		
17	16	75	77		
18	17	90	80		
19	18	120	85		
20	19	240	95		
21	20	270	94		
22	21	15	68		
23	22	30	75		
24	23	45	85		
25	24	90	89		
26	25	75	73		
27	26	60	78		
28	27	60	74		
29	28	120	84		
30	29	45	80		
31	30	30	84		
32	31	75	70		
33	32	90	95		
34	33	150	97		
35	34	129	86		
36	35	75	84		
37	36	60	70		
38	37	45	82		
39	38	180	80		
40	39	200	84		
41	40	210	100		
42					

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Coefficients of the Regression Line

- The regression formulas for the coefficients are shown below, and we use the sample data to derive estimates of β_0 and β_1 (see appendix for full derivation):

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Calculating the coefficients of the regression line

Microsoft Excel - studytime

File Edit View Insert Format Tools Data Window Help

Arial 12 B I U 100%

B55

	A	B	C	D	E	F	G	H
1	Using Regression to explain and predict exam score from study time							
2	Simple Linear Regression							
3		Study Time	Exam Score					
4	Student $_i$	x_i	y_i	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$	
5	1	60	85					
6	2	45	80					
7	3	50	83					
8	4	75	90					
9	5	120	78					
12	38	180	80					
13	39	200	84					
14	40	210	100					
15	Averages:				Totals:			
17	$\beta_1 = \Sigma(x_i - \overline{x})(y_i - \overline{y}) / \Sigma(x_i - \overline{x})^2 =$							
19	$\beta_0 = \overline{y} - b_1\overline{x} =$							
51	The regression equation ($\hat{y} = \beta_0 + \beta_1 * x$):							
52								

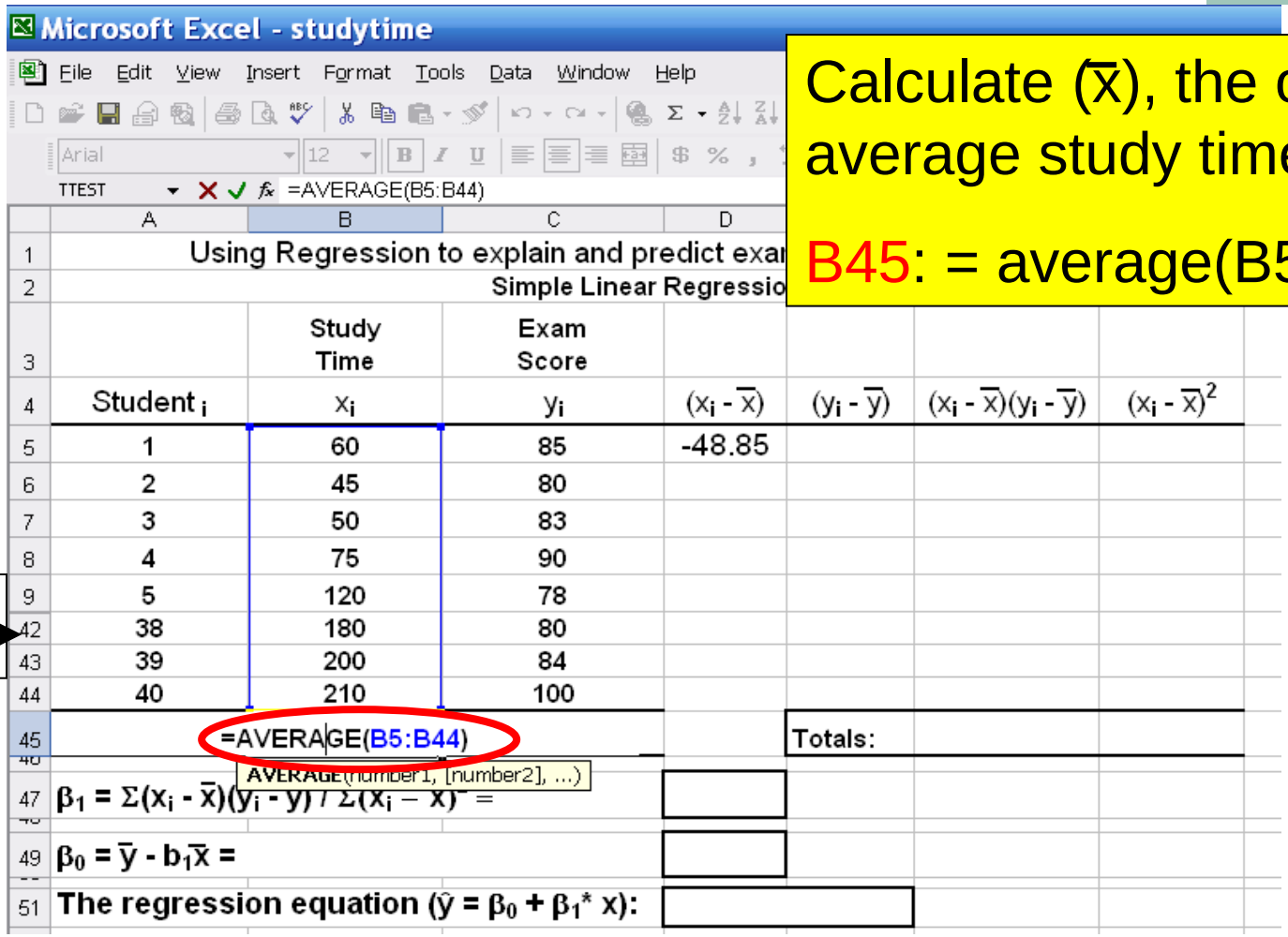
The coefficients can be calculated 'manually' in Excel

Follow along these next slides to see

G

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Calculating the coefficients of the regression line



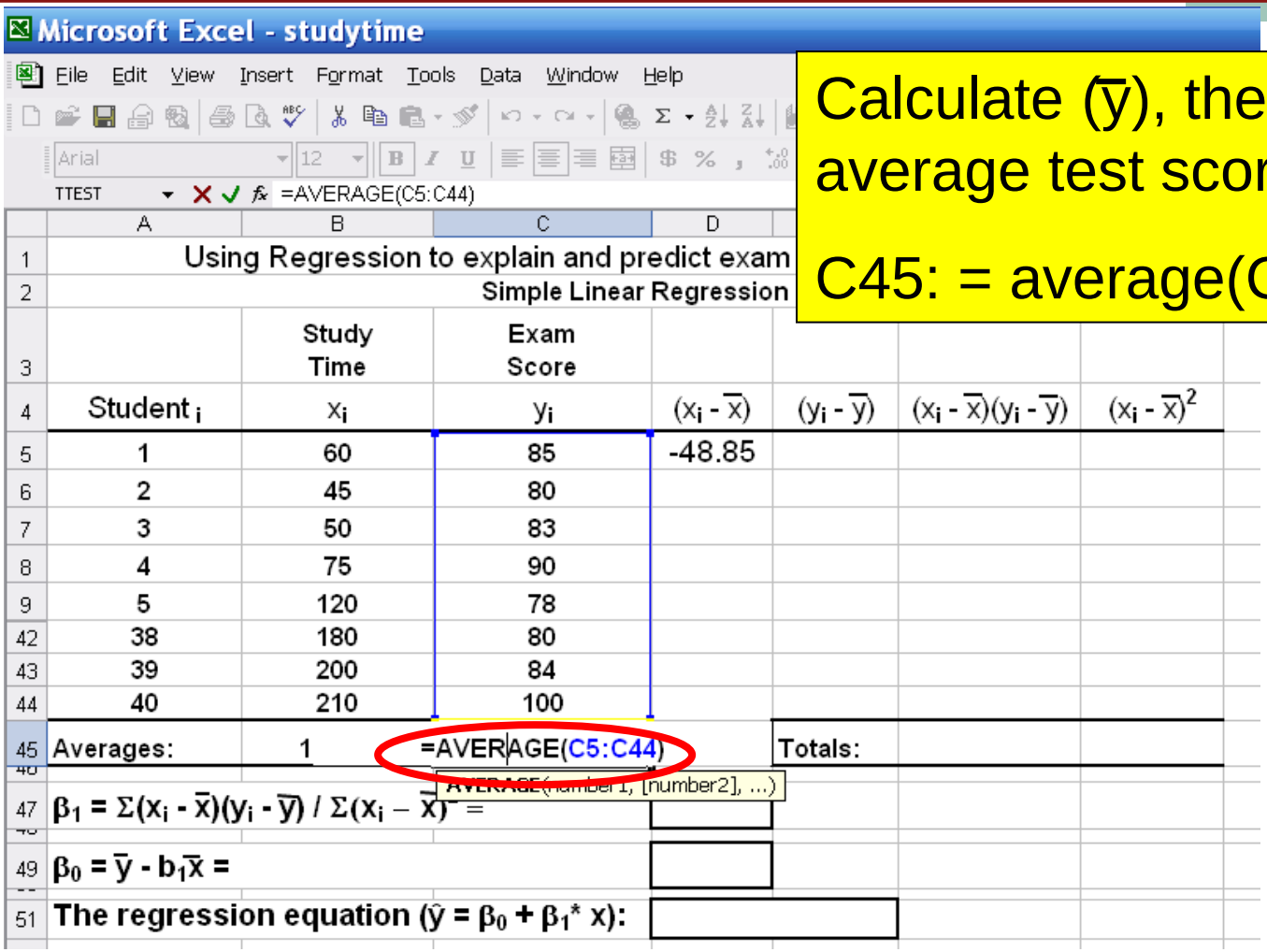
Calculate ($\bar{x}$), the overall average study time:

B45: = average(B5:B44)

Rows hidden

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Calculating the coefficients of the regression line

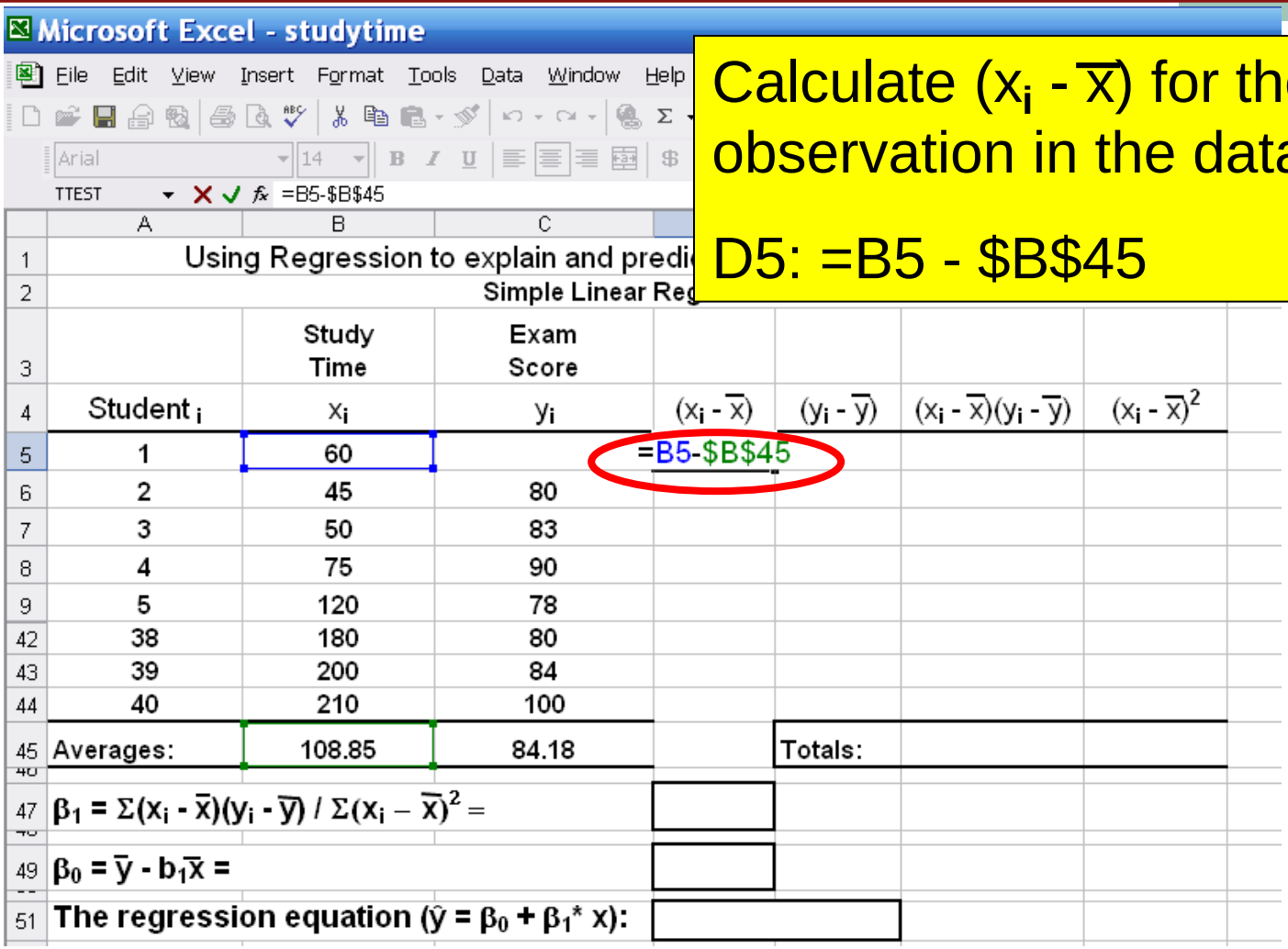


Calculate ($\bar{y}$), the overall average test score:

C45: = average(C5:C44)

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Calculating the coefficients of the regression line



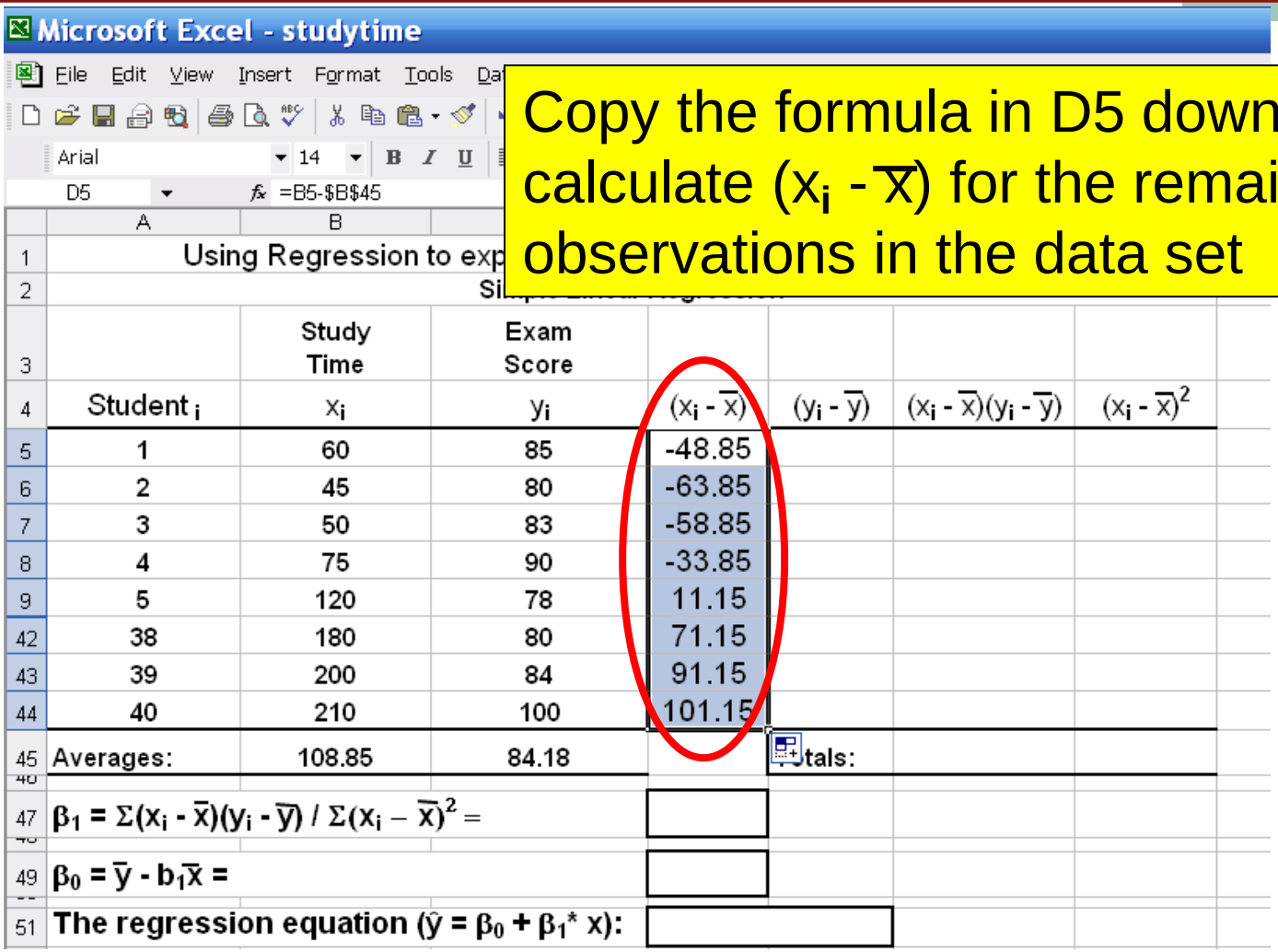
	A	B	C	D	E	F	G
1	Using Regression to explain and predi						
2	Simple Linear Reg						
3		Study Time	Exam Score				
4	Student $_i$	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
5	1	60		=B5-B45			
6	2	45	80				
7	3	50	83				
8	4	75	90				
9	5	120	78				
42	38	180	80				
43	39	200	84				
44	40	210	100				
45	Averages:	108.85	84.18		Totals:		
47	$\beta_1 = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \Sigma(x_i - \bar{x})^2 =$						
49	$\beta_0 = \bar{y} - b_1\bar{x} =$						
51	The regression equation ($\hat{y} = \beta_0 + \beta_1 * x$):						

Calculate $(x_i - \bar{x})$ for the first observation in the data set:

D5: =B5 - B45

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Calculating the coefficients of the regression line

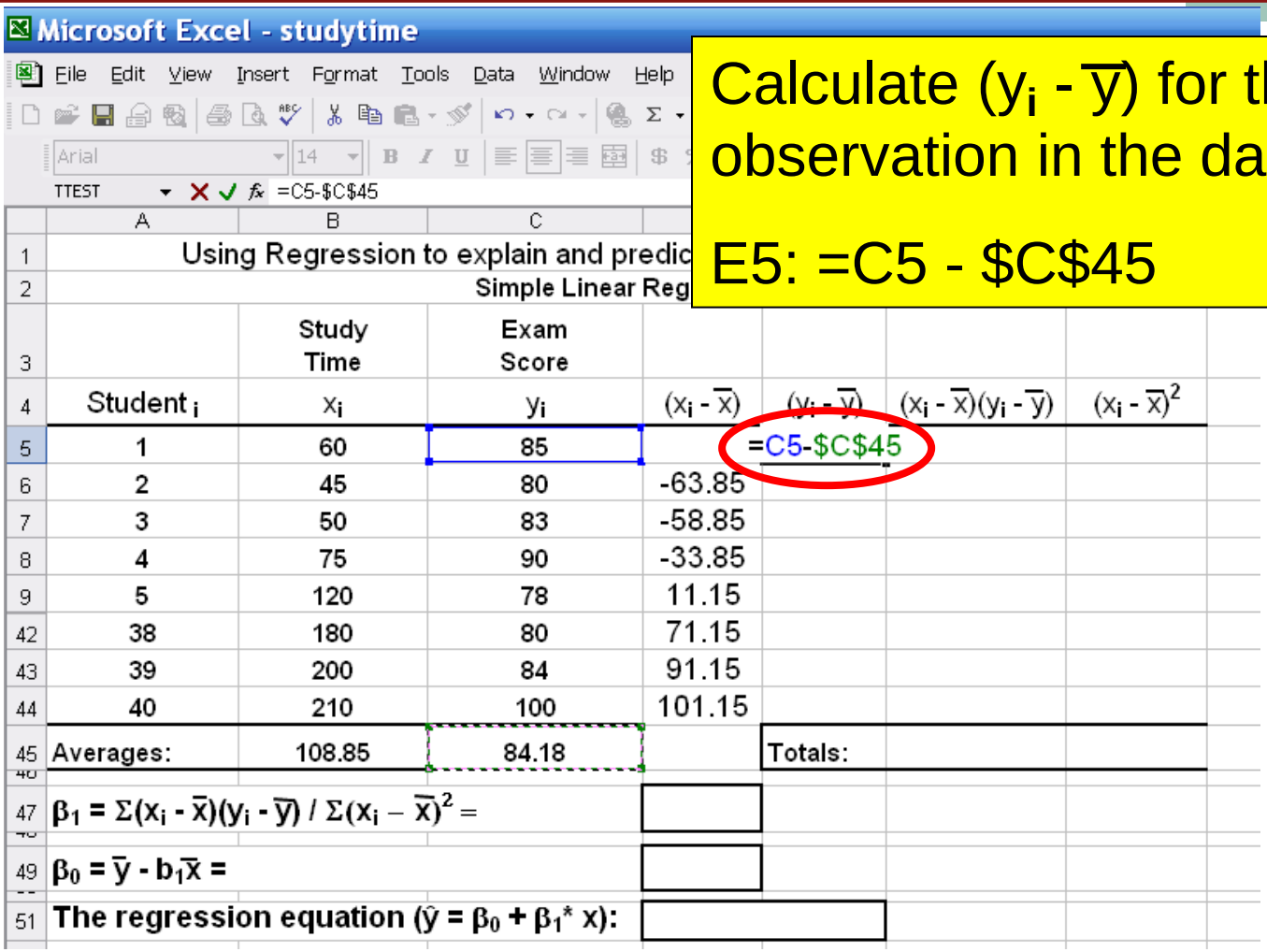


	A	B	C	D	E	F	G
1	Using Regression to exp						
2	Si						
3		Study Time	Exam Score				
4	Student $_i$	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
5	1	60	85	-48.85			
6	2	45	80	-63.85			
7	3	50	83	-58.85			
8	4	75	90	-33.85			
9	5	120	78	11.15			
42	38	180	80	71.15			
43	39	200	84	91.15			
44	40	210	100	101.15			
45	Averages:	108.85	84.18		tals:		
47	$\beta_1 = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \Sigma(x_i - \bar{x})^2 =$						
49	$\beta_0 = \bar{y} - b_1\bar{x} =$						
51	The regression equation ($\hat{y} = \beta_0 + \beta_1^* x$):						

Copy the formula in D5 down to D44 to calculate $(x_i - \bar{x})$ for the remaining observations in the data set

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Calculating the coefficients of the regression line



Calculate $(y_i - \bar{y})$ for the first observation in the data set:

E5: =C5 - C45

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Calculating the coefficients of the regression line

Copy the formula in E5 down to E44 to calculate $(y_i - \bar{y})$ for the remaining observations in the data set

Microsoft Excel - studytime

File Edit View Insert Format Tools Da

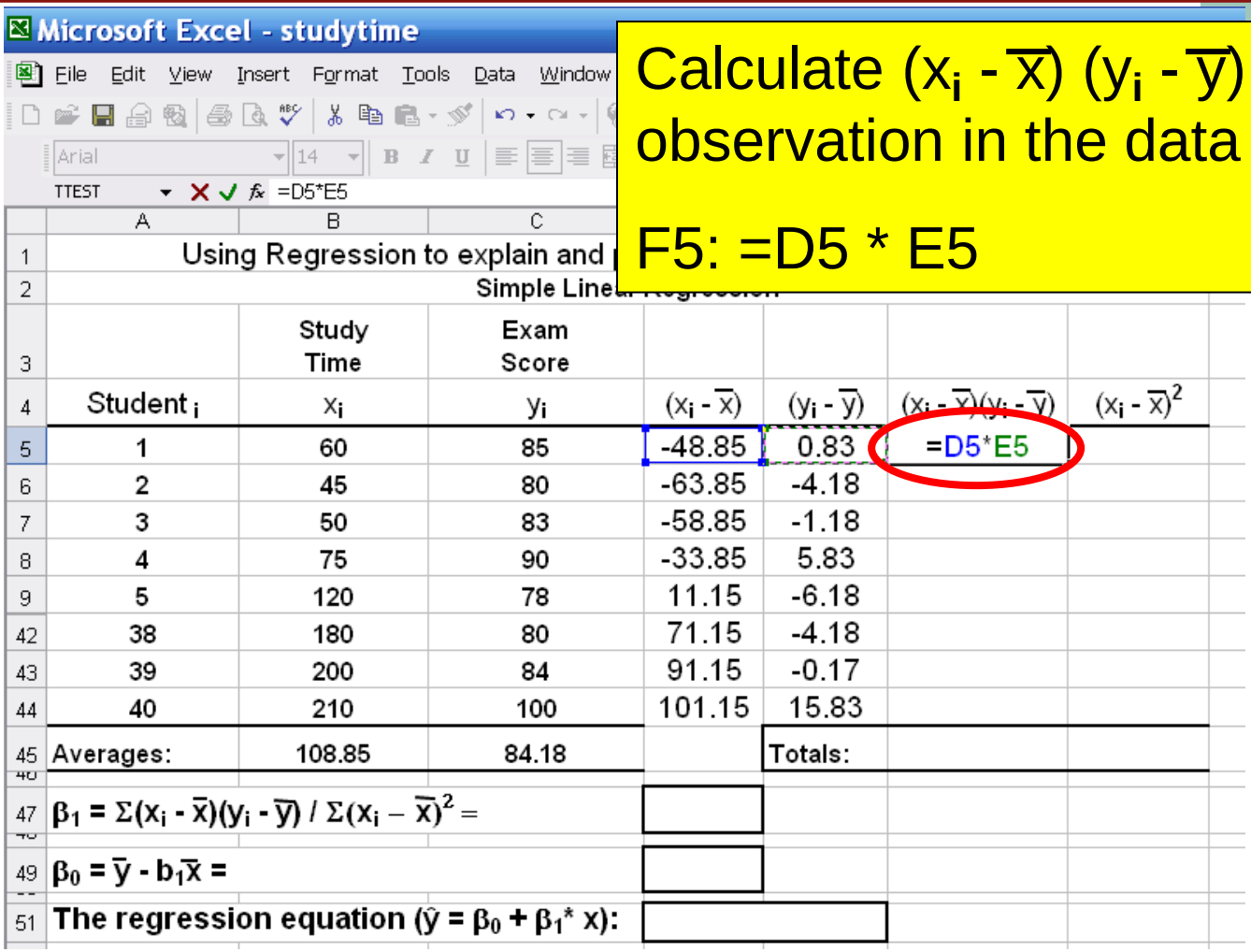
Arial 14 B I U

E5 =C5-C45

	A	B	C	D	E	F	G	H
1	Using Regression to explain and predict exam score from study time							
2	Simple Linear Regression							
3		Study Time	Exam Score					
4	Student $_i$	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	
5	1	60	85	-48.85	0.83			
6	2	45	80	-63.85	-4.18			
7	3	50	83	-58.85	-1.18			
8	4	75	90	-33.85	5.83			
9	5	120	78	11.15	-6.18			
42	38	180	80	71.15	-4.18			
43	39	200	84	91.15	-0.17			
44	40	210	100	101.15	15.83			
45	Averages:	108.85	84.18		Totals:			
47	$\beta_1 = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \Sigma(x_i - \bar{x})^2 =$							
49	$\beta_0 = \bar{y} - b_1\bar{x} =$							
51	The regression equation ($\hat{y} = \beta_0 + \beta_1 * x$):							

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Calculating the coefficients of the regression line



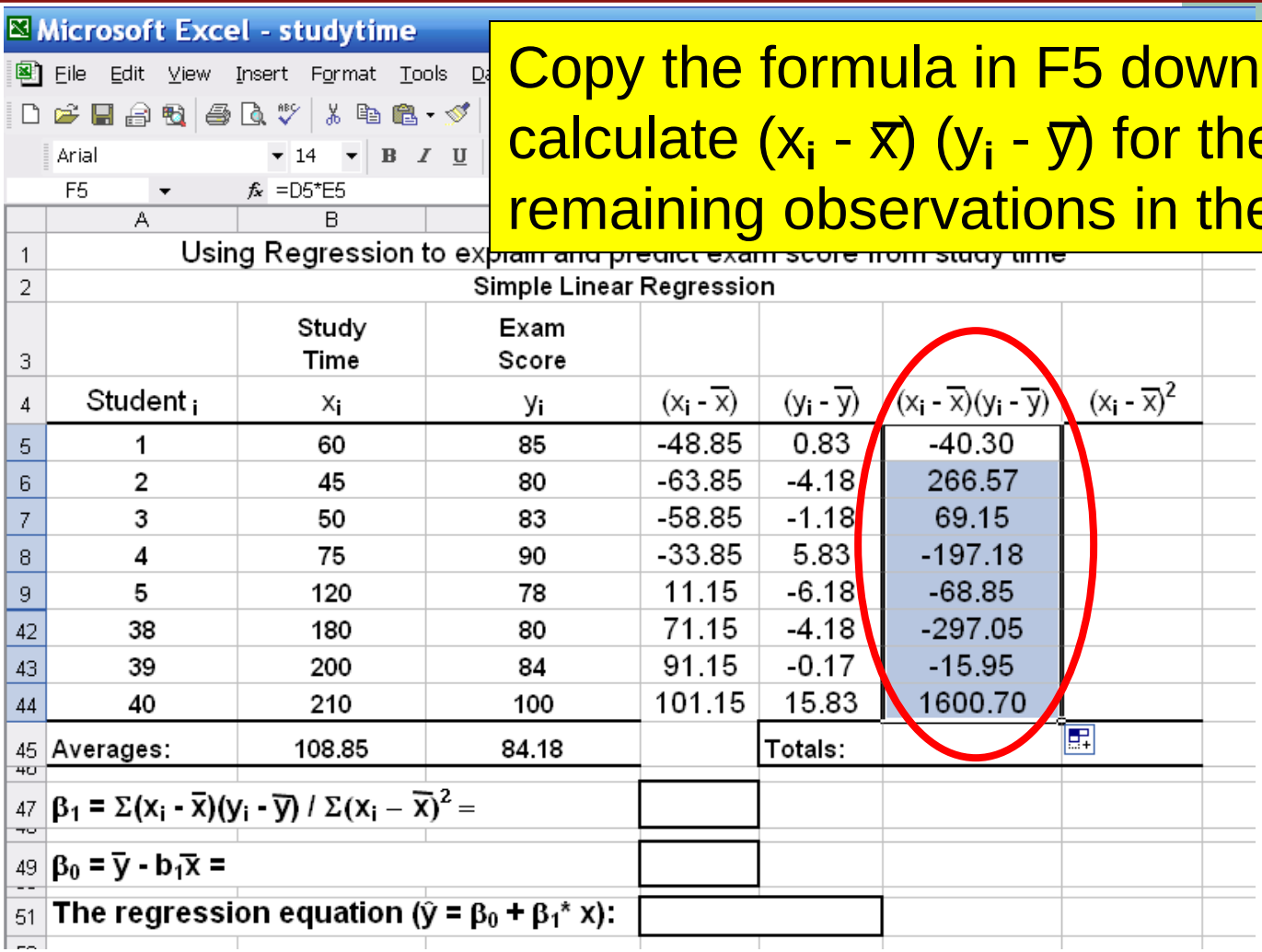
	A	B	C	D	E	F	G
1	Using Regression to explain and						
2	Simple Linear Regression						
3		Study Time	Exam Score				
4	Student $_i$	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
5	1	60	85	-48.85	0.83	=D5*E5	
6	2	45	80	-63.85	-4.18		
7	3	50	83	-58.85	-1.18		
8	4	75	90	-33.85	5.83		
9	5	120	78	11.15	-6.18		
42	38	180	80	71.15	-4.18		
43	39	200	84	91.15	-0.17		
44	40	210	100	101.15	15.83		
45	Averages:	108.85	84.18		Totals:		
47	$\beta_1 = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \Sigma(x_i - \bar{x})^2 =$						
49	$\beta_0 = \bar{y} - b_1\bar{x} =$						
51	The regression equation ($\hat{y} = \beta_0 + \beta_1 * x$):						

Calculate $(x_i - \bar{x})$ $(y_i - \bar{y})$ for the first observation in the data set:

F5: =D5 * E5

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Calculating the coefficients of the regression line



Copy the formula in F5 down to F44 to calculate $(x_i - \bar{x})(y_i - \bar{y})$ for the remaining observations in the data set

Using Regression to explain and predict exam score from study time

Simple Linear Regression

	Student $_i$	Study Time x_i	Exam Score y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
5	1	60	85	-48.85	0.83	-40.30	
6	2	45	80	-63.85	-4.18	266.57	
7	3	50	83	-58.85	-1.18	69.15	
8	4	75	90	-33.85	5.83	-197.18	
9	5	120	78	11.15	-6.18	-68.85	
42	38	180	80	71.15	-4.18	-297.05	
43	39	200	84	91.15	-0.17	-15.95	
44	40	210	100	101.15	15.83	1600.70	
45	Averages:	108.85	84.18		Totals:		

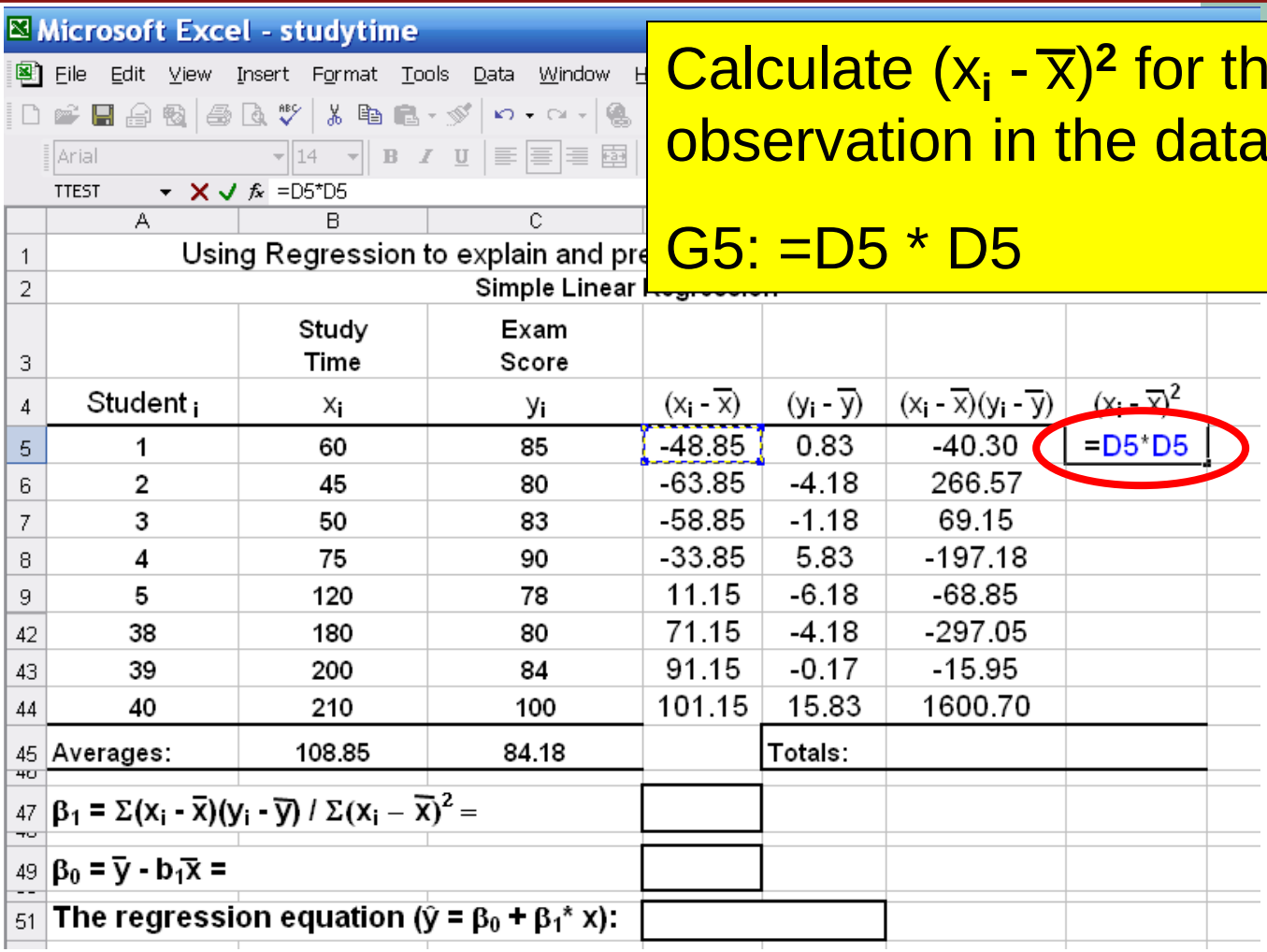
47 $\beta_1 = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \Sigma(x_i - \bar{x})^2 =$

49 $\beta_0 = \bar{y} - b_1\bar{x} =$

51 The regression equation ($\hat{y} = \beta_0 + \beta_1 * x$):

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Calculating the coefficients of the regression line



	A	B	C	D	E	F	G
1	Using Regression to explain and pre						
2	Simple Linear						
3		Study Time	Exam Score				
4	Student $_i$	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
5	1	60	85	-48.85	0.83	-40.30	=D5*D5
6	2	45	80	-63.85	-4.18	266.57	
7	3	50	83	-58.85	-1.18	69.15	
8	4	75	90	-33.85	5.83	-197.18	
9	5	120	78	11.15	-6.18	-68.85	
42	38	180	80	71.15	-4.18	-297.05	
43	39	200	84	91.15	-0.17	-15.95	
44	40	210	100	101.15	15.83	1600.70	
45	Averages:	108.85	84.18		Totals:		
47	$\beta_1 = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \Sigma(x_i - \bar{x})^2 =$						
49	$\beta_0 = \bar{y} - b_1\bar{x} =$						
51	The regression equation ($\hat{y} = \beta_0 + \beta_1 * x$):						

Calculate $(x_i - \bar{x})^2$ for the first observation in the data set:

G5: =D5 * D5

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Calculating the coefficients of the regression line

Microsoft Excel - studytime

File Edit View Insert Format Tool

Arial 14 B I

G5 =D5*D5

Copy the formula in G5 down to G44 to calculate $(x_i - \bar{x})^2$ for the remaining observations in the data set

	A	B	C	D	E	F	G
1	Using Regression to						
2	Simple Linear Regression						
3		Study Time	Exam Score				
4	Student $_i$	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
5	1	60	85	-48.85	0.83	-40.30	2386.32
6	2	45	80	-63.85	-4.18	266.57	4076.82
7	3	50	83	-58.85	-1.18	69.15	3463.32
8	4	75	90	-33.85	5.83	-197.18	1145.82
9	5	120	78	11.15	-6.18	-68.85	124.323
42	38	180	80	71.15	-4.18	-297.05	5062.32
43	39	200	84	91.15	-0.17	-15.95	8308.32
44	40	210	100	101.15	15.83	1600.70	10231.3
45	Averages:	108.85	84.18		Totals:		
47	$\beta_1 = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \Sigma(x_i - \bar{x})^2 =$						
49	$\beta_0 = \bar{y} - b_1\bar{x} =$						
51	The regression equation ($\hat{y} = \beta_0 + \beta_1^* x$):						

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Calculating the coefficients of the regression line

Microsoft Excel - studytime

File Edit View Insert Format Tools Data Window Help

Arial 12 B I U AutoSum

TTEST =SUM(F5:F44)

	A	B	C	D	E	F	G
1	Using Regression to explain and predict exam score from study time						
2	Simple Linear Regression						
3		Study Time	Exam Score				
4	Student i	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
5	1	60	85	-48.85	0.83	-40.30	2386.32
6	2	45	80	-63.85	-4.18	266.57	4076.82
7	3	50	83	-58.85	-1.18	69.15	3463.32
8	4	75	90	-33.85	5.83	-197.18	1145.82
9	5	120	78	11.15	-6.18	-68.85	124.323
42	38	180	80	71.15	-4.18	-297.05	5062.32
43	39	200	84	91.15	-0.17	-15.95	8308.32
44	40	210	100	101.15	15.83	1600.70	10231.3
45	Averages:	108.85	84.18		Totals:	=SUM(F5:F44)	
47	$\beta_1 = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \Sigma(x_i - \bar{x})^2 =$						
49	$\beta_0 = \bar{y} - b_1\bar{x} =$						
51	The regression equation ($\hat{y} = \beta_0 + \beta_1 * x$):						

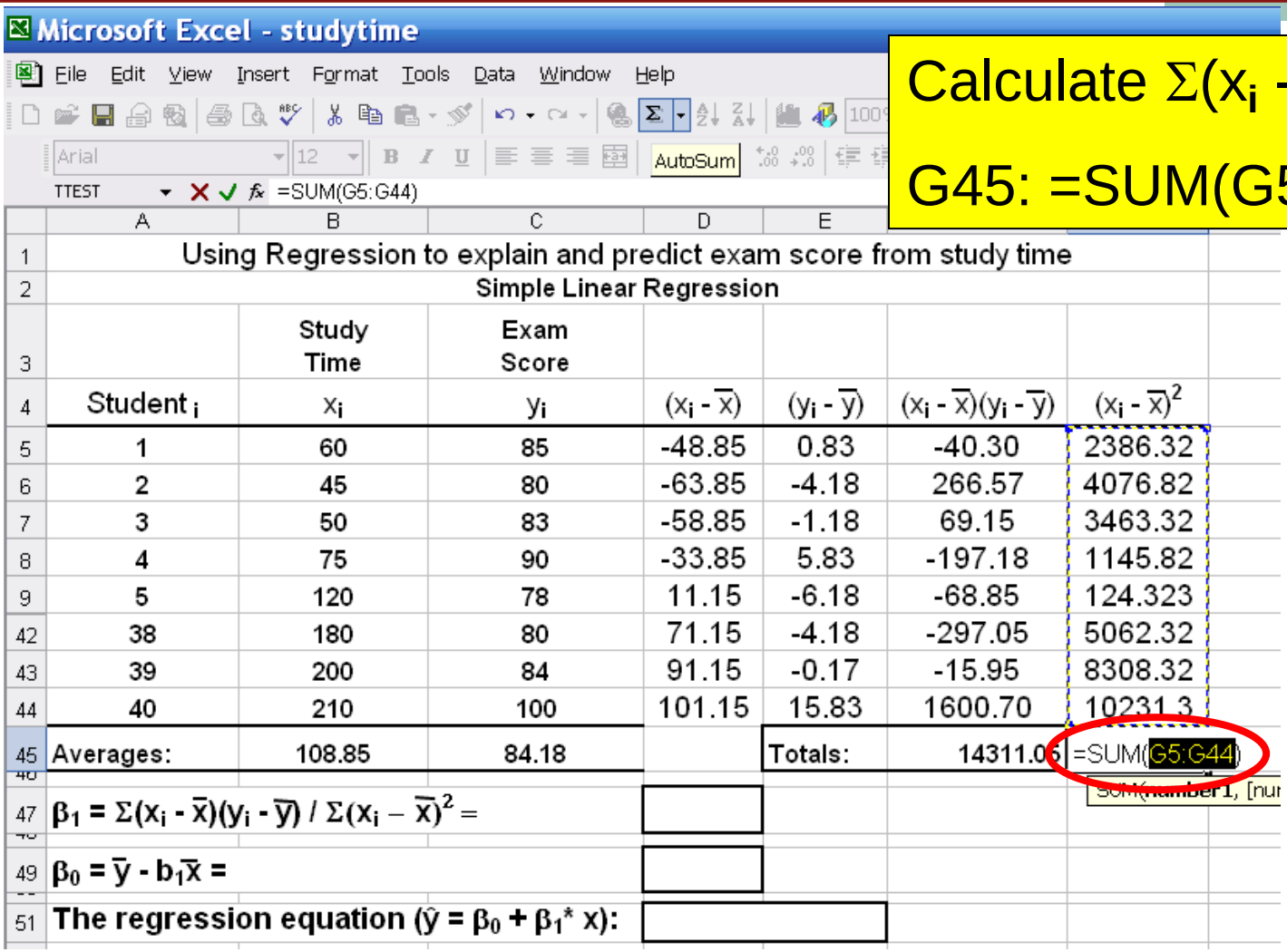
SUM(number1, [number2], ...)

Calculate $\Sigma(x_i - \bar{x})(y_i - \bar{y})$:

F45: =SUM(F5:F44)

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Calculating the coefficients of the regression line



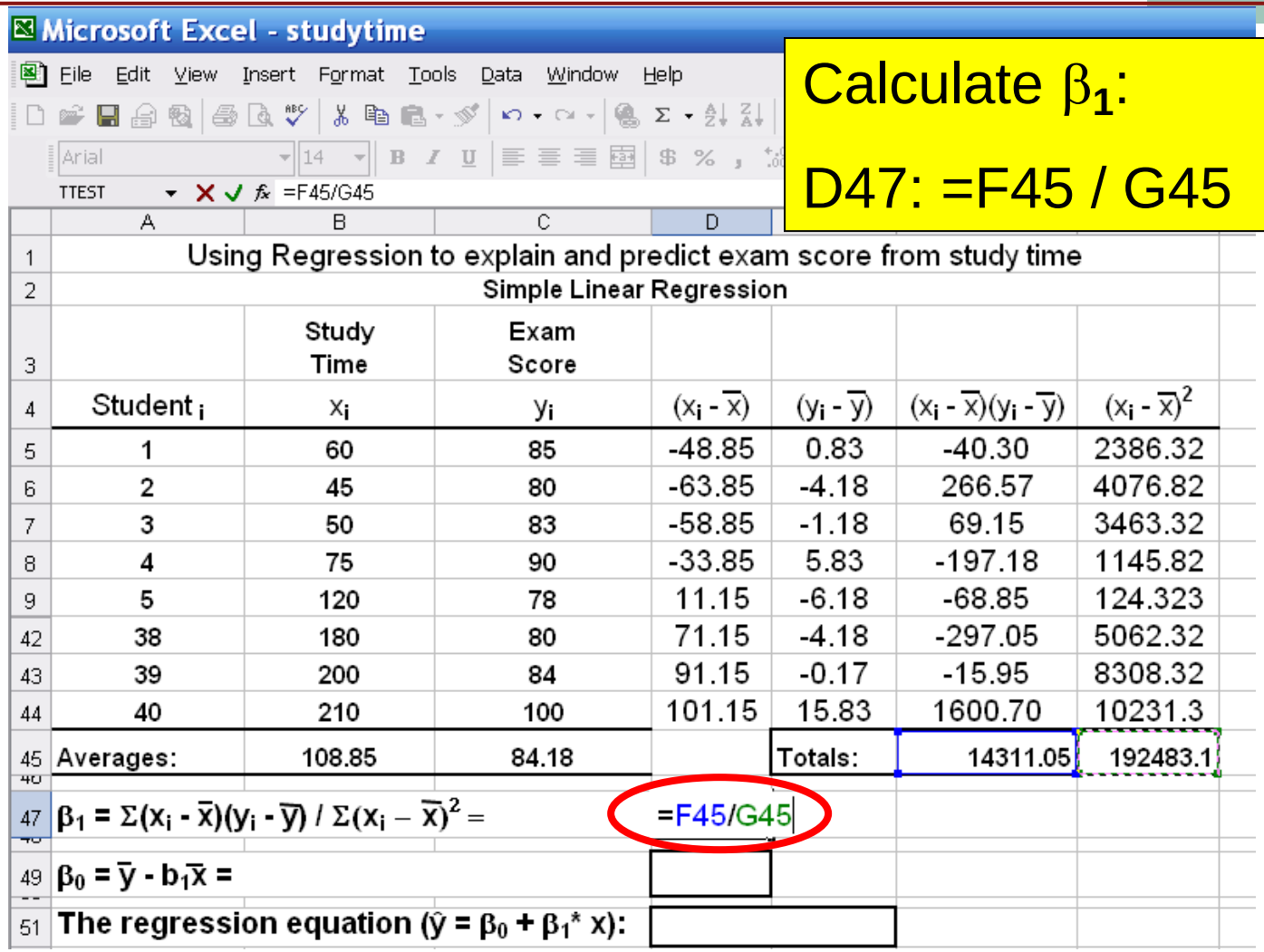
	A	B	C	D	E	F	G
1	Using Regression to explain and predict exam score from study time						
2	Simple Linear Regression						
3		Study Time	Exam Score				
4	Student $_i$	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
5	1	60	85	-48.85	0.83	-40.30	2386.32
6	2	45	80	-63.85	-4.18	266.57	4076.82
7	3	50	83	-58.85	-1.18	69.15	3463.32
8	4	75	90	-33.85	5.83	-197.18	1145.82
9	5	120	78	11.15	-6.18	-68.85	124.323
42	38	180	80	71.15	-4.18	-297.05	5062.32
43	39	200	84	91.15	-0.17	-15.95	8308.32
44	40	210	100	101.15	15.83	1600.70	10231.3
45	Averages:	108.85	84.18		Totals:	14311.05	=SUM(G5:G44)
47	$\beta_1 = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \Sigma(x_i - \bar{x})^2 =$						
49	$\beta_0 = \bar{y} - b_1\bar{x} =$						
51	The regression equation ($\hat{y} = \beta_0 + \beta_1^* x$):						

Calculate $\Sigma(x_i - \bar{x})^2$:

G45: =SUM(G5:G44)

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Calculating the coefficients of the regression line



	A	B	C	D	E	F	G
1	Using Regression to explain and predict exam score from study time						
2	Simple Linear Regression						
3		Study Time	Exam Score				
4	Student $_i$	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
5	1	60	85	-48.85	0.83	-40.30	2386.32
6	2	45	80	-63.85	-4.18	266.57	4076.82
7	3	50	83	-58.85	-1.18	69.15	3463.32
8	4	75	90	-33.85	5.83	-197.18	1145.82
9	5	120	78	11.15	-6.18	-68.85	124.323
42	38	180	80	71.15	-4.18	-297.05	5062.32
43	39	200	84	91.15	-0.17	-15.95	8308.32
44	40	210	100	101.15	15.83	1600.70	10231.3
45	Averages:	108.85	84.18		Totals:	14311.05	192483.1
47	$\beta_1 = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \Sigma(x_i - \bar{x})^2 =$			=F45/G45			
49	$\beta_0 = \bar{y} - b_1\bar{x} =$						
51	The regression equation ($\hat{y} = \beta_0 + \beta_1 * x$):						

Calculate β_1:

D47: =F45 / G45

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Calculating the coefficients of the regression line

Microsoft Excel - studytime

File Edit View Insert Format Tools Data Window Help

Arial 14 B I U

TTEST =C45-(D47*B45)

	A	B	C	D	E	F	G
1	Using Regression to explain and predict exam score from study time						
2	Simple Linear Regression						
3		Study Time	Exam Score				
4	Student $_i$	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
5	1	60	85	-48.85	0.83	-40.30	2386.32
6	2	45	80	-63.85	-4.18	266.57	4076.82
7	3	50	83	-58.85	-1.18	69.15	3463.32
8	4	75	90	-33.85	5.83	-197.18	1145.82
9	5	120	78	11.15	-6.18	-68.85	124.323
42	38	180	80	71.15	-4.18	-297.05	5062.32
43	39	200	84	91.15	-0.17	-15.95	8308.32
44	40	210	100	101.15	15.83	1600.70	10231.3
45	Averages:	108.85	84.18		Totals:	14311.05	192483.1
47	$\beta_1 = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \Sigma(x_i - \bar{x})^2 =$			0.0743			
49	$\beta_0 = \bar{y} - b_1\bar{x} =$			=C45-(D47*B45)			
51	The regression equation ($\hat{y} = \beta_0 + \beta_1 * x$):						

Calculate β_0:

D49: =C45 – (D47 * B45)

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Calculating the coefficients of the regression line

Microsoft Excel - studytime

File Edit View Insert Format Tools Data Window Help

Arial 14 B I U

D51 fx ŷ=76.082+.0743*x

	A	B	C	D	E	F	G
1	Using Regression to explain and predict exam score from study time						
2	Simple Linear Regression						
3		Study Time	Exam Score				
4	Student $_i$	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
5	1	60	85	-48.85	0.83	-40.30	2386.32
6	2	45	80	-63.85	-4.18	266.57	4076.82
7	3	50	83	-58.85	-1.18	69.15	3463.32
8	4	75	90	-33.85	5.83	-197.18	1145.82
9	5	120	78	11.15	-6.18	-68.85	124.323
42	38	180	80	71.15	-4.18	-297.05	5062.32
43	39	200	84	91.15	-0.17	-15.95	8308.32
44	40	210	100	101.15	15.83	1600.70	10231.3
45	Averages:	108.85	84.18		Totals:	14311.05	192483.1
47	$\beta_1 = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \Sigma(x_i - \bar{x})^2 =$			0.0743			
49	$\beta_0 = \bar{y} - b_1\bar{x} =$			76.082			
51	The regression equation ($\hat{y} = \beta_0 + \beta_1 * x$)			ŷ=76.082+.0743*x			

Write the regression equation:

D51: $\hat{y} = 76.08 + .0743x$

All material, both content and format contained herein including the slide notes but except for the textbook exhibits are the property of and copyright by Dr. Dan Brandon. Memphis TN USA

Measuring the Fit of the Regression Model

- Regression models can be developed for any variables X and Y
- How do we know that the model is actually helpful in predicting Y based on X?
- Three measures of variability are
  - SST – Total variability about the mean
  - SSE – Variability about the regression line
  - SSR – Total variability that is explained by the model

Fit of the Regression Model (con't)

- Sum of the squares total (actual minus average, Y bar is overall average)

$$SST = \sum (Y - \bar{Y})^2$$

- Sum of the squared error (actual minus model)

$$SSE = \sum e^2 = \sum (Y - \hat{Y})^2$$

- Sum of squares due to regression (model minus overall average)

$$SSR = \sum (\hat{Y} - \bar{Y})^2$$

- An important relationship

$$SST = SSR + SSE$$

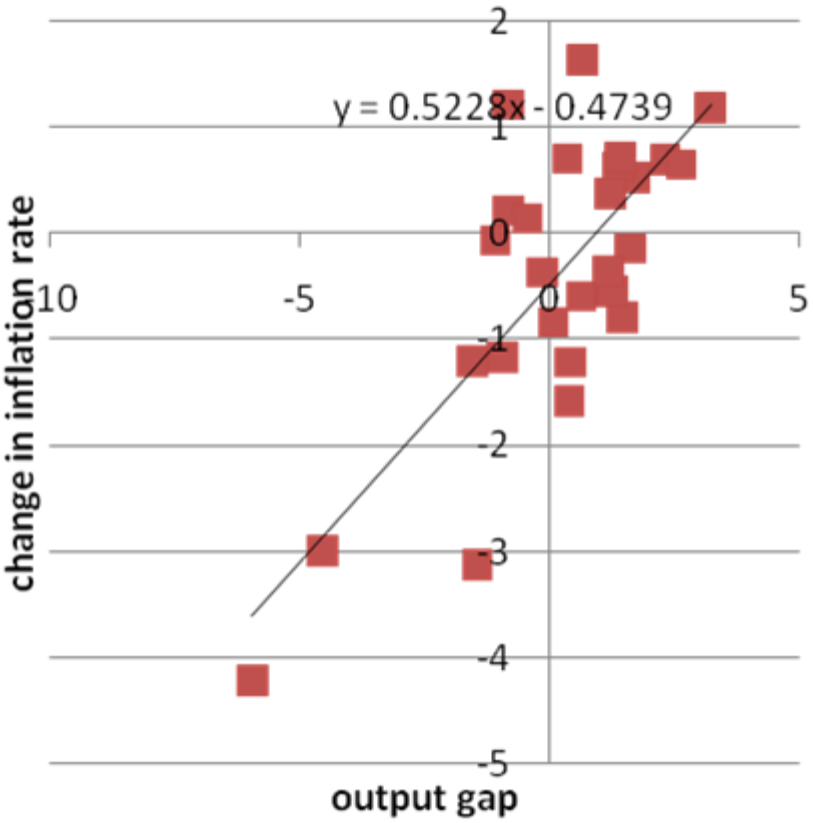
Coefficient of Determination

- The proportion of the variability in Y explained by regression equation is called the *coefficient of determination*
- The coefficient of determination is r^2, which is between zero and 1
- A smaller SSE/SST ratio yields a larger r^2

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

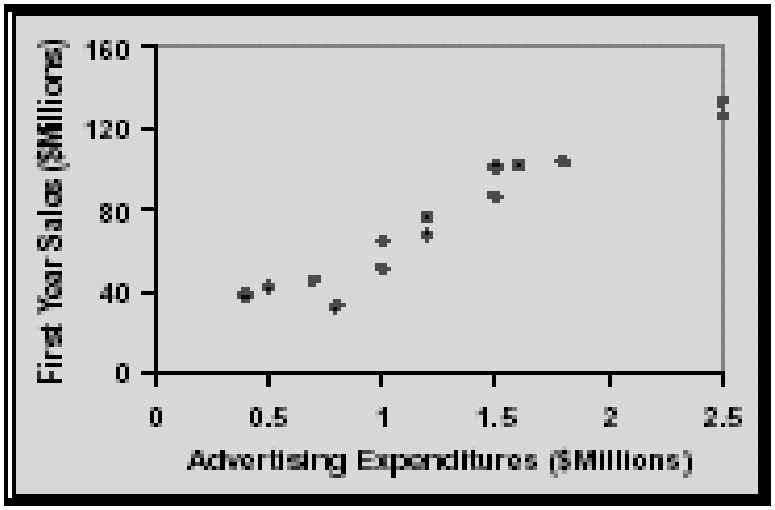
Coefficient of Determination (con't)

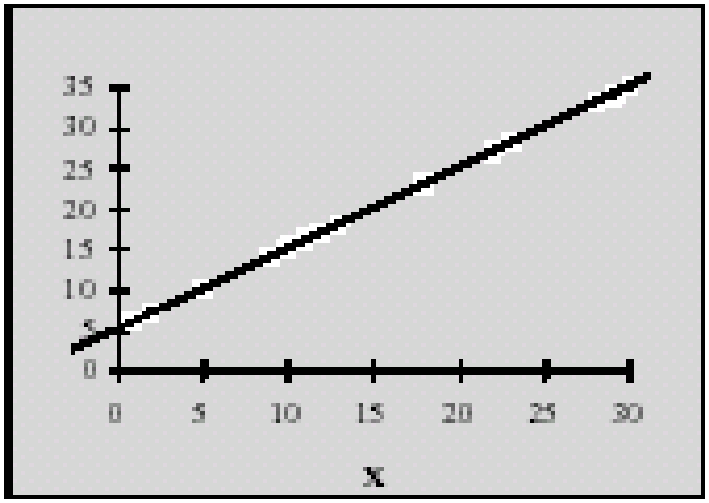
- It is a measure of the overall quality of the regression
- r^2 (or $R^{2)}$ takes on values between 0 and 1; it is a ratio of that portion of the dependent variable's variation that is explained by the fitted model
- R^2 is pearson's r squared



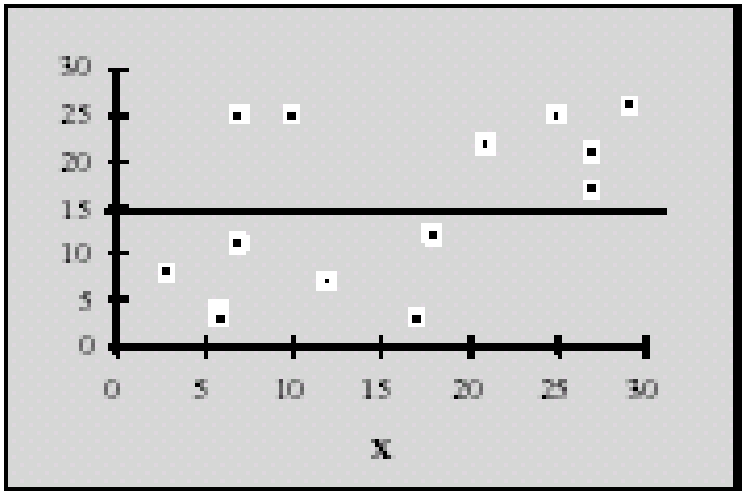
R^2 takes values between 0 and 1 (it is a percentage).

R^2 = 0.833 in our Appleglo Example





R^2 = 1; x values account for all variation in the Y values



R^2 = 0; x values account for none variation in the Y values

Assumptions of the Regression Model

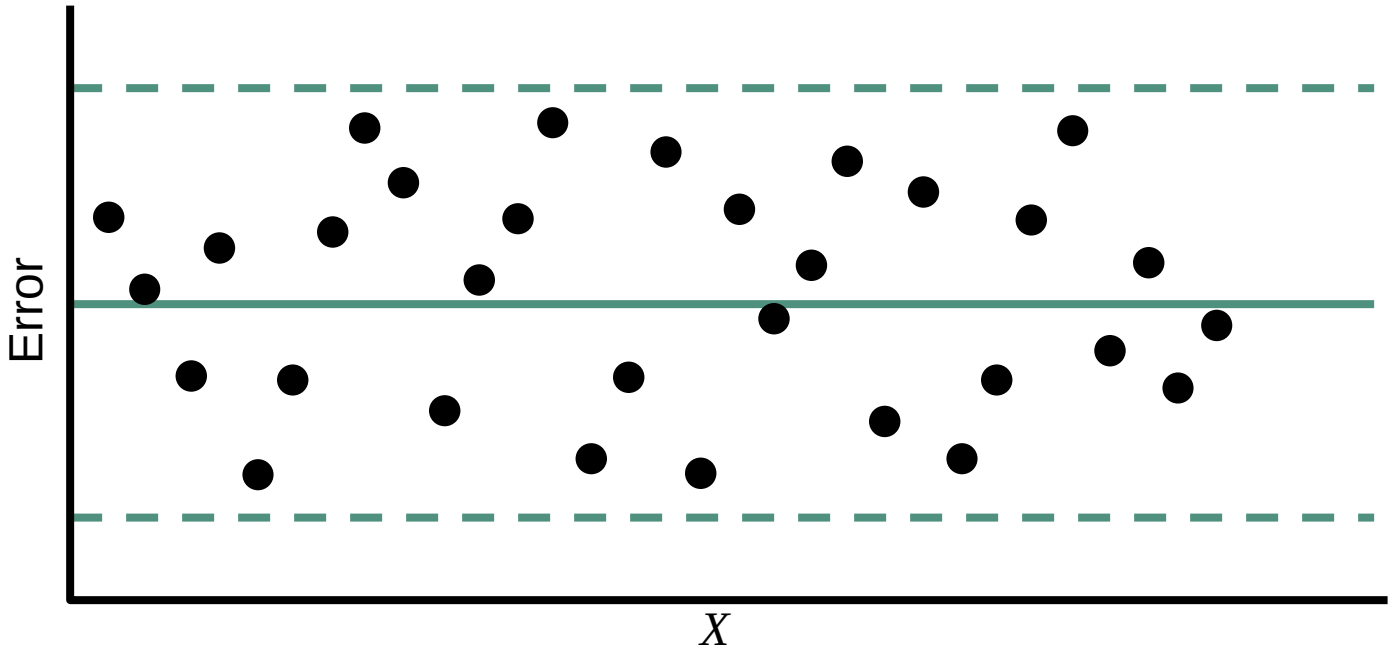
- If we make certain assumptions about the errors in a regression model, we can perform statistical tests to determine if the model is useful

  1. Errors are independent
  2. Errors are normally distributed
  3. Errors have a mean of zero
  4. Errors have a constant variance

- A plot of the residuals (errors) will often highlight any glaring violations of the assumptions

Residual Plots

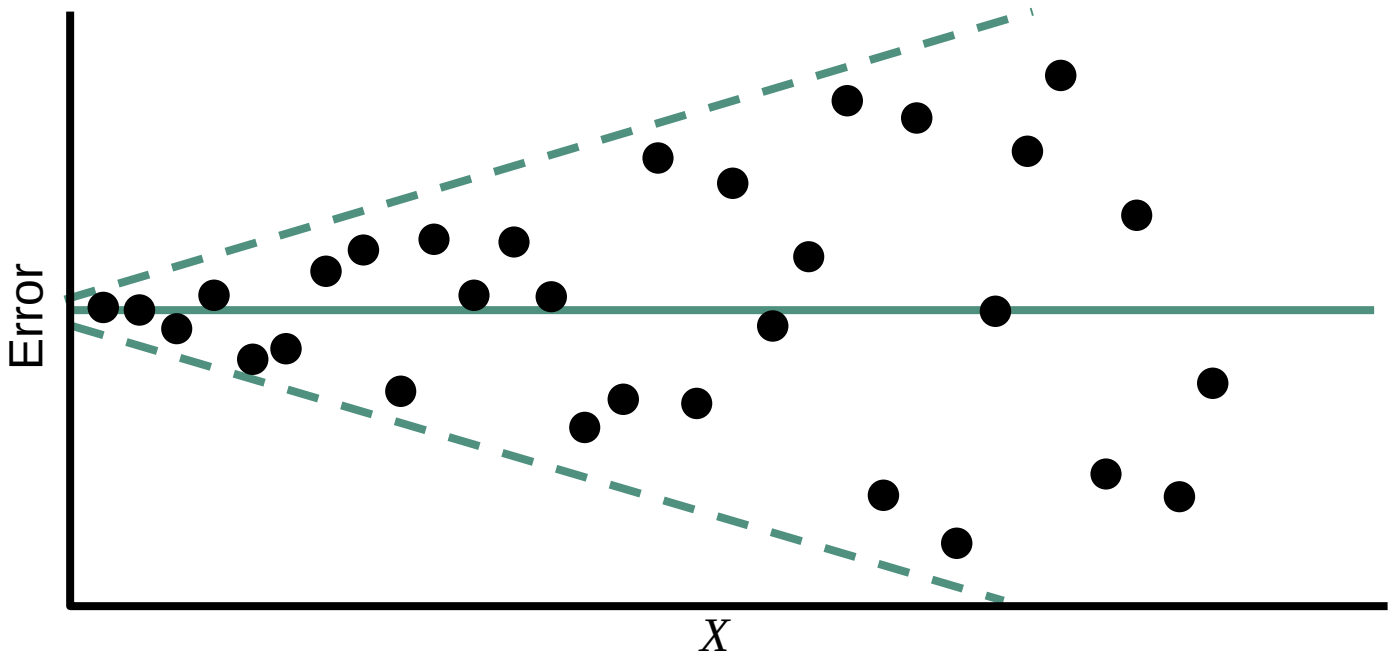
- A random plot of residuals



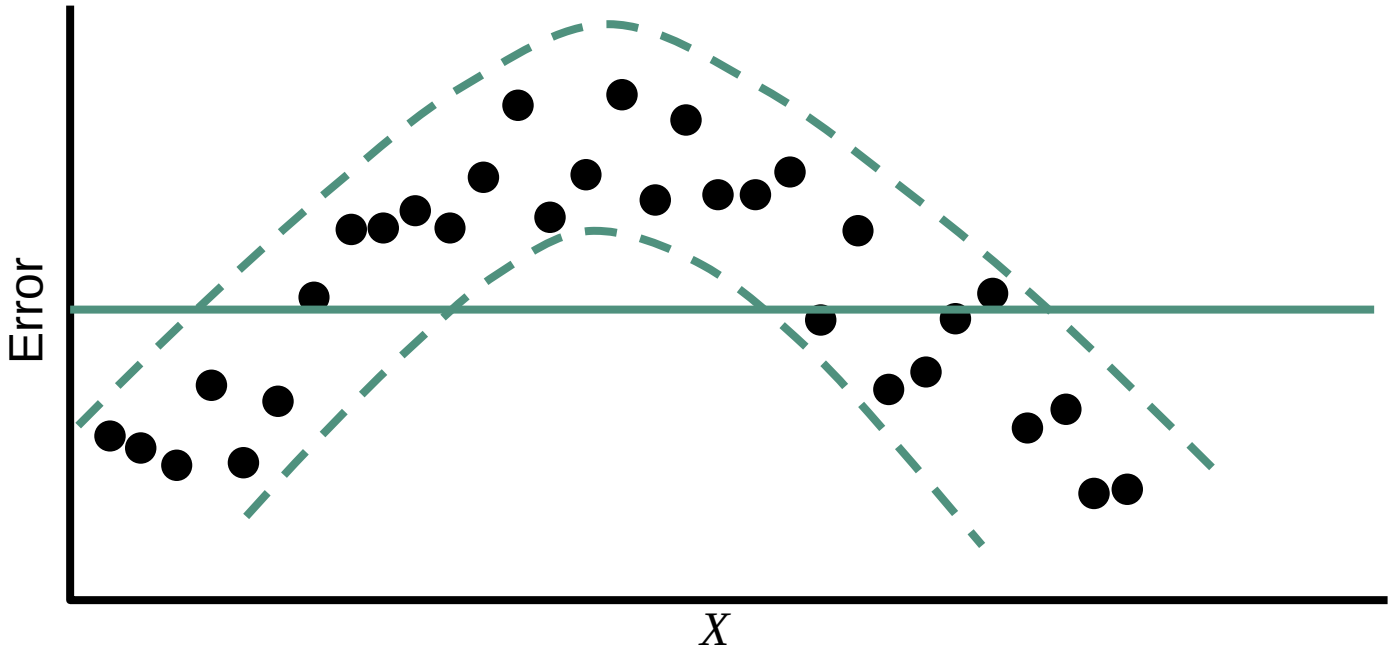
This is the type of random pattern we want to see for a good linear model.

Residual Plots (con't)

- Nonconstant error variance



Non Linear Residual Plot



Does the western conference NBA team with the largest salary cap win more games?

Team	Win %	03-04 Salary Cap
Timberwolves	70.7	72
Lakers	68.3	66
Spurs	69.5	47
Kings	67.1	70
Mavericks	63.4	79
Grizzlies	61.0	58
Rockets	54.9	52
Nuggets	52.4	36
Jazz	51.2	28
Trailblazers	50.0	84
Warriors	45.1	52
Supersonics	45.1	51
Suns	35.4	65
Clippers	34.1	38

How is the NBA salary cap determined ?

NBA Salary Cap

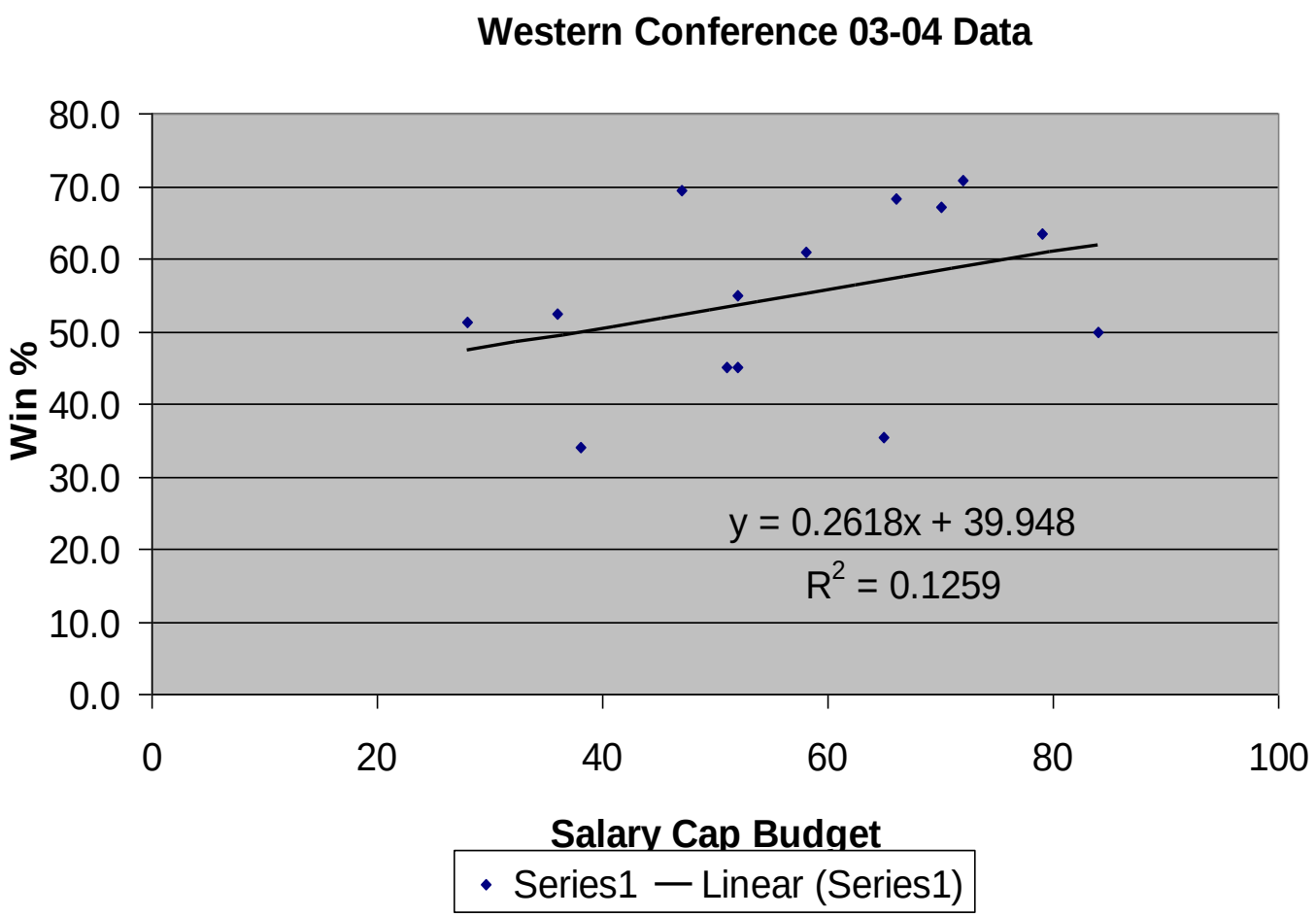
- This is the limit to the total amount of money that National Basketball Association teams are allowed to pay their players
- It is defined by the league's collective bargaining agreement (CBA), and is subject to a complex system of rules and exceptions and as such is considered a "soft" cap
- The actual amount of the salary cap varies on a year-to-year basis, and is calculated as a percentage of the league's revenue from the previous season
- Like many professional sports leagues, the NBA has a salary cap to control cost and foster team equality
- The maximum amount of money a player can sign for is based on the number of years that player has played and the total of the salary cap; for example, the maximum salary of a player with 6 or fewer years of experience is either $9,000,000 or 25% of the total salary cap (2013–14: $14,670,000), whichever is greater; however there are many types of “exceptions”

Copyright – Dan Brandon

VIEWING: 2014 Team Salary Caps GO

RANK	TEAM	ACTIVE PLAYERS	GUARD CAP	FORWARD CAP	CENTER CAP	TOTAL CAP	MAX TYPE	SPACE
1	Atlanta Hawks	15	$19,566,923	$27,451,111	$12,000,000	60,035,044	Hard	$20,790,956
2	Philadelphia 76ers	15	$10,141,055	$15,023,903	$5,730,363	42,296,914	Max	$20,765,086
3	Phoenix Suns	15	$35,642,326	$12,555,460	$5,649,920	61,841,966	Hard	$19,157,034
4	Minnesota Timberwolves	14	$24,093,277	$21,946,749	$13,561,505	64,359,797	Hard	$16,439,203
5	Los Angeles Lakers	14	$25,407,655	$10,302,023	$9,915,243	46,707,154	Max	$16,357,846
6	Portland Trail Blazers	15	$16,561,503	$39,215,372	$13,242,649	69,352,524	Hard	$11,476,176
7	Denver Nuggets	14	$26,554,042	$24,234,923	$15,394,430	69,665,395	Hard	$11,160,605
8	Chicago Bulls	14	$23,456,302	$29,515,343	$14,122,579	67,432,557	LT	$9,396,443
9	Sacramento Kings	15	$13,463,266	$41,431,300	$15,661,243	71,514,559	Hard	$9,314,411
10	New Orleans Pelicans	14	$37,032,120	$20,941,935	$10,172,212	65,760,395	LT	$5,046,605
11	Golden State Warriors	15	$33,360,565	$23,460,958	$14,902,335	72,965,125	Hard	$7,540,572
12	Milwaukee Bucks	16	$19,767,155	$19,060,161	$16,200,000	55,597,669	Max	$7,167,331
13	San Antonio Spurs	16	$32,264,116	$19,253,559	$17,250,000	69,951,359	LT	$6,547,641
14	Indiana Pacers	15	$16,145,406	$39,247,262	$19,406,274	74,795,942	Hard	$6,030,058
15	Memphis Grizzlies	15	$27,632,596	$27,919,671	$15,629,666	75,019,956	Hard	$5,509,042
16	Houston Rockets	15	$26,100,274	$25,159,615	$21,436,271	75,771,592	Hard	$5,057,105
17	Washington Wizards	14	$24,657,364	$24,221,164	$26,350,025	75,528,553	Hard	$5,000,447
18	Miami Heat	14	$22,276,646	$26,936,762	$22,230,763	71,967,393	LT	$4,561,607
19	Toronto Raptors	15	$36,743,960	$32,334,671	$6,356,253	76,096,523	Hard	$4,732,477
20	Orlando Magic	15	$17,556,450	$21,002,006	$2,751,260	55,530,516	Max	$4,234,462
21	Utah Jazz	17	$26,350,200	$24,751,992	$6,622,074	59,562,424	Max	$3,462,576
22	Dallas Mavericks	14	$30,673,404	$25,993,355	$15,795,050	74,366,567	LT	$2,460,433
23	Charlotte Hornets	14	$23,962,206	$15,456,403	$21,404,456	60,823,067	Max	$2,241,933
24	Los Angeles Clippers	13	$35,560,663	$23,670,430	$16,745,123	79,244,755	Hard	$1,554,245
25	Boston Celtics	14	$21,570,136	$31,559,466	$1,703,760	62,271,847	Max	$793,153
26	Detroit Pistons	14	$21,562,450	$14,595,656	$11,545,393	63,034,444	Max	$30,556
27	Oklahoma City Thunder	15	$25,245,637	$35,094,229	$11,839,302	76,422,492	LT	$-1,593,492
28	New York Knicks	15	$11,552,634	$52,654,950	$915,242	80,257,876	LT	$-3,525,876
29	Cleveland Cavaliers	15	$15,556,390	$45,949,354	$16,565,233	81,593,356	LT	$-5,064,356
30	Brooklyn Nets	14	$53,157,649	$7,153,436	$29,592,625	90,549,514	LT	$-13,720,514

Salary Cap vs Win %



Analysis of Trendline

- The prediction or “trendline” formula is:
  - Wins = 0.2618 * Salary + 39.948
  - So for about every 10 million of salary cap, the team will win 2.6% more games

The teams would win about 40% of their games if the players played for free ???

Determination Coefficient Analysis

- $R^2 = 0.1259$
  - This is the measure of the overall quality of the linear regression
  - This value is between 0 and 1
  - Here it is closer to zero, so as to suggest that there is not a very good correlation between salary and win%

Multiple Linear Regression

- SLR only involves one independent variable
- When there is more than one independent variable, then the analysis becomes more complicated
- The concepts are the same, except we are working in multi-dimensional space instead of just two dimensions
  - $Y = b_0 + b_1*X_1 + b_2*X_2 + \ldots + b_n*X_n$
- For our advertising example here, there may be other factors influencing sales values in addition to advertising expenditures

Multiple Independent Variables

region	sales	advertising	promotions	competitor's sales
Selkirk	101.8	1.3	0.2	20.40
Susquehanna	44.4	0.7	0.2	30.50
Kittery	108.3	1.4	0.3	24.60
Acton	85.1	0.5	0.4	19.60
Finger Lakes	77.1	0.5	0.6	25.50
Berkshire	158.7	1.9	0.4	21.70
Central	180.4	1.2	1.0	6.80
Providence	64.2	0.4	0.4	12.60
Nashua	74.6	0.6	0.5	31.30
Dunster	143.4	1.3	0.6	18.60
Endicott	120.6	1.6	0.8	19.90
Five-Towns	69.7	1.0	0.3	25.60
Waldeboro	67.8	0.8	0.2	27.40
Jackson	106.7	0.6	0.5	24.30
Stowe	119.6	1.1	0.3	13.70

What are the independent and dependent variables ?

Minimize the Sum of the Residual Squares

- Choosing the coefficients (b_0 …b_n) to minimize the sum of the residual squares gives us the prediction formula
- The prediction formula here is:
  - Sales = 65.705 + (48.979 * Advertising) + (59.654 * Promotions) - (1.838 * CompetitorSales)

Typical MLR Tool Output

Regression Statistics	
Multiple R	0.913
R Square	0.833
Adjusted R Square	0.787
Standard Error	17.600
Observations	15

Analysis of Variance

	df	Sum of Squares	Mean Square	F	Significance F
Regression	3	16997.537	5665.85	18.290	0.000
Residual	11	3407.473	309.77		
Total	14	20405.009			

	Coefficients	Standard Error	t Statistic	P-value	Lower 95%	Upper 95%
Intercept	65.71	27.73	2.37	0.033	4.67	126.74
Advertising	48.98	10.66	4.60	0.000	25.52	72.44
Promotions	59.65	23.63	2.53	0.024	7.66	111.65
Competitor's Sales	-1.84	0.81	-2.26	0.040	-3.63	-0.047

Typical MLR Tool Output (con't)

b_0 = 65.705 (its interpretation is context dependent .

b_1 = 48.979 (an additional $1 million in advertising is expected to result in an additional $49 million in sales)

b_2 = 59.654 (an additional $1 million in promotions is expected to result in an additional $60 million in sales)

b_3 = -1.838 (an increase of $1 million in competitor sales is expected to decrease sales by $1.8 million)

Adjusted R Squared

- The formula for R squared is:
  - R^2 = 1 – (variation not accounted for/total variation)
  - $R^2 = 1 – (SS_{error}/SS_{total})$
- The Adjusted R Squared adjusts for the degrees of freedom in the model, and penalizes an unnecessarily complex model (too many independent variables)
- The formula is
  - $aR^2 = 1 – (SS_{error}/df_{error})/(SS_{total}/df_{total})$
  - Where df_{error} is the degree of freedom [n-1-k]
  - And df_{total} is total degrees of freedom [n-1]

Testing the Model for Significance

- When the sample size is too small, you can get good values for MSE and r^2 even if there is no relationship between the variables
- Testing the model for significance helps determine if the values are meaningful
- We do this by performing a statistical hypothesis test

Testing the Model for Significance (con't)

- The overall significance is checked by performing an F-test, with the null hypothesis being that all the slopes are zero
  - When F is large then the significance is small and the overall model is good
- We can also do a hypothesis test for each slope (variable) using a t-test
  - When the p-value is small, then there is a strong relationship for that variable
- For simple linear regression, the F test and t-test give the same results

Copyright – Dan Brandon

Testing the Model for Significance (con't)

- The F statistic is based on the MSE and MSR

$$\boldsymbol{MSR = \frac{SSR}{k}}$$

where

k = number of independent variables in the model

- The F statistic is

$$\boldsymbol{F = \frac{MSR}{MSE}}$$

- This describes an F distribution with

  degrees of freedom for the numerator = $\mathrm{df}_1 = k$

  degrees of freedom for the denominator = $\mathrm{df}_2 = n - k - 1$

Testing the Model for Significance (con't)

Regression Statistics	
Multiple R	0.913
R Square	0.833
Adjusted R Square	0.787
Standard Error	17.600
Observations	15

Analysis of Variance

	df	Sum of Squares	Mean Square	F	Significance F
Regression	3	16997.537	5665.85	18.290	0.000
Residual	11	3407.473	309.77		
Total	14	20405.009			

	Coefficients	Standard Error	t Statistic	P-value	Lower 95%	Upper 95%
Intercept	65.71	27.73	2.37	0.033	4.67	126.74
Advertising	48.98	10.66	4.60	0.000	25.52	72.44
Promotions	59.65	23.63	2.53	0.024	7.66	111.65
Competitor's Sales	-1.84	0.81	-2.26	0.040	-3.63	-0.047

Copyright – Dan Brandon

Testing the Model for Significance (con't)

- **Looking at the coefficients (slopes), can we tell which independent variables are the most influential ?**

Regression Statistics	
Multiple R	0.913
R Square	0.833
Adjusted R Square	0.787
Standard Error	17.600
Observations	15

Analysis of Variance

	df	Sum of Squares	Mean Square	F	Significance F
Regression	3	16997.537	5665.85	18.290	0.000
Residual	11	3407.473	309.77		
Total	14	20405.009			

	Coefficients	Standard Error	t Statistic	P-value	Lower 95%	Upper 95%
Intercept	65.71	27.73	2.37	0.033	4.67	126.74
Advertising	48.98	10.66	4.60	0.000	25.52	72.44
Promotions	59.65	23.63	2.53	0.024	7.66	111.65
Competitor's Sales	-1.84	0.81	-2.26	0.040	-3.63	-0.047

Copyright – Dan Brandon

Multicollinearity

- **Multicollinearity** is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated
- In this situation the coefficient estimates may change erratically in response to small changes in the model or the data
- Multicollinearity does not reduce the predictive power or reliability of the model as a whole – the overall F test is still valid
- It only affects calculations regarding individual predictors (independent variables)

Copyright – Dan Brandon

Singularity

- Singularity is the extreme form of multicollinearity - when a near perfect linear relationship exists between variables
- Such absolute multicollinearity could arise when independent variables are linearly related in their definition
- A simple example: two variables "height in centimeters" and "height in inches" are included in the regression model

Copyright – Dan Brandon

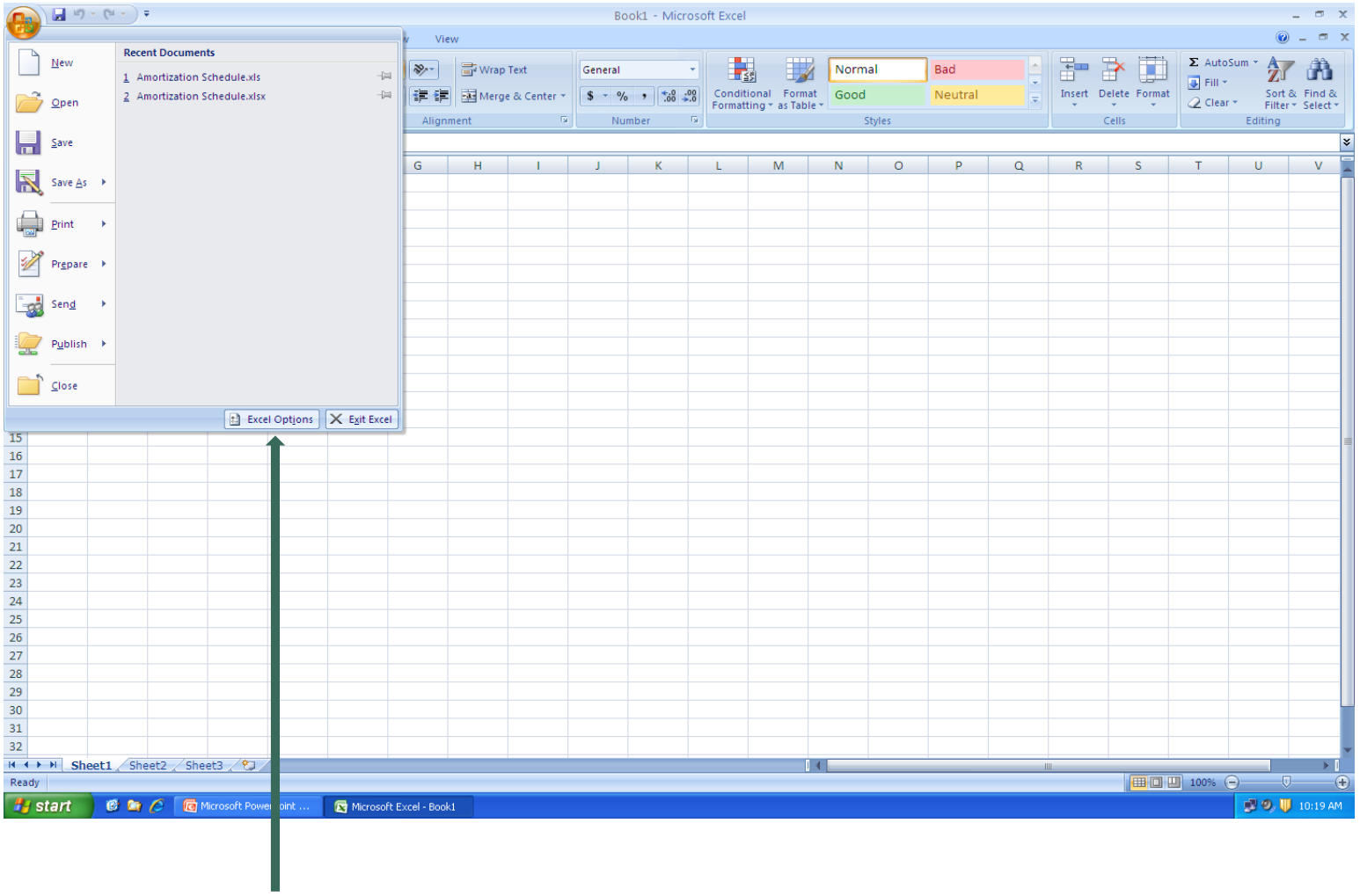
Matrix Solution of MLR

- To perform MLR, a matrix formulation and solution is normally used
- $b = (X'X)^{-1} X'y$
  - where X is a matrix of the data values for the independent variables, y is a vector of the values of the dependent variable, b is the solution vector (weighting parameters), -1 denotes the matrix inverse, and ' denotes the transpose of the matrix.
- Numerically it is not efficient for large models to directly calculate matrix inverses, so other numerical techniques are used such as Gauss, Householder, etc.
- “Sparse Matrix” techniques are also used for large regression problems

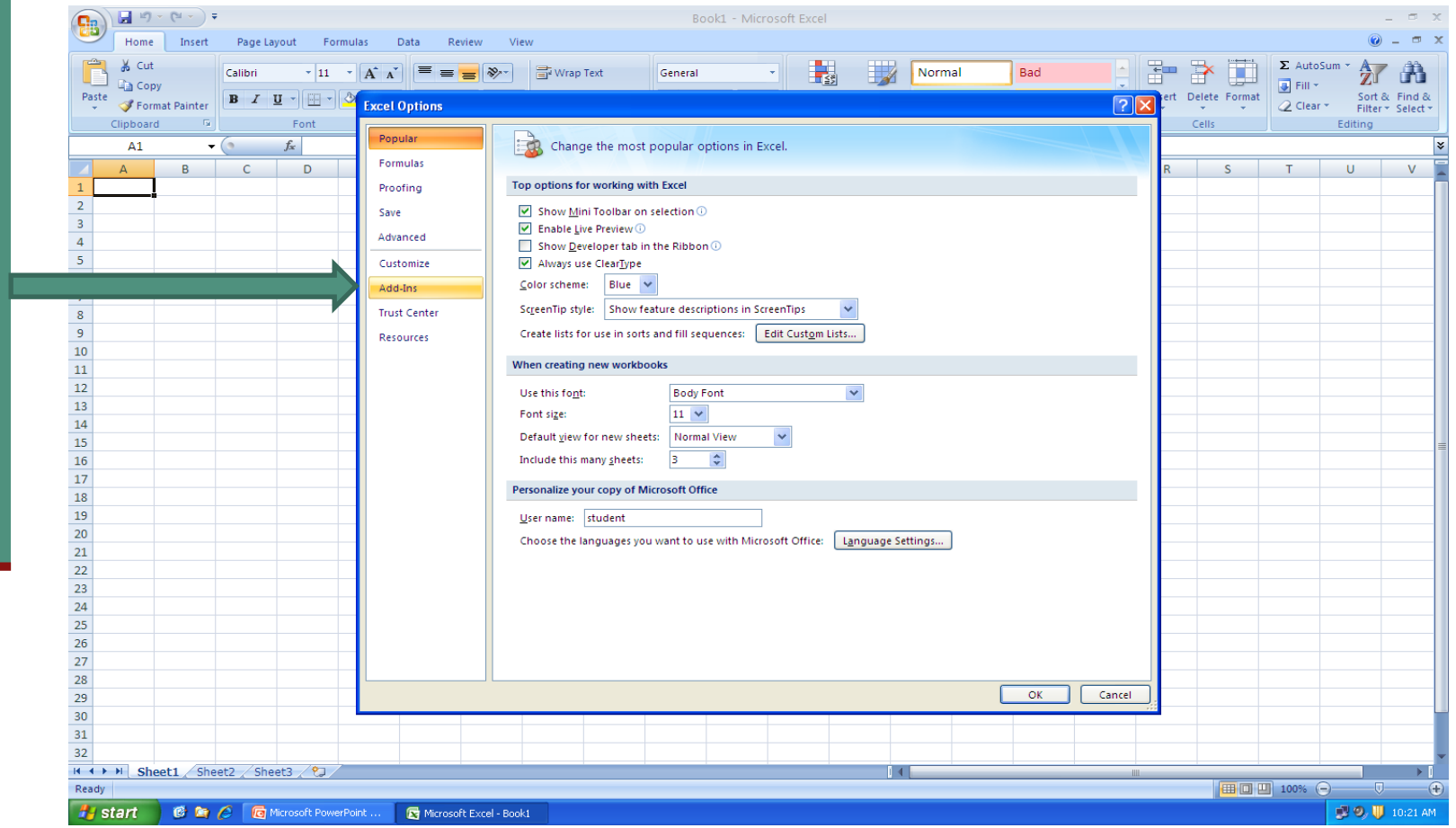
MLR in Excel

- Install the “Analysis ToolPak” if not already installed
  - Excel 2003: Tools…Add-ins…Analysis ToolPak
  - Excel 2007/2010+: Office Button (File in 2010)…Excel Options…Add-ins…Analysis ToolPak
- Enter dependent and independent variable data (see next slide)
- Tools…Data Analysis…Regression

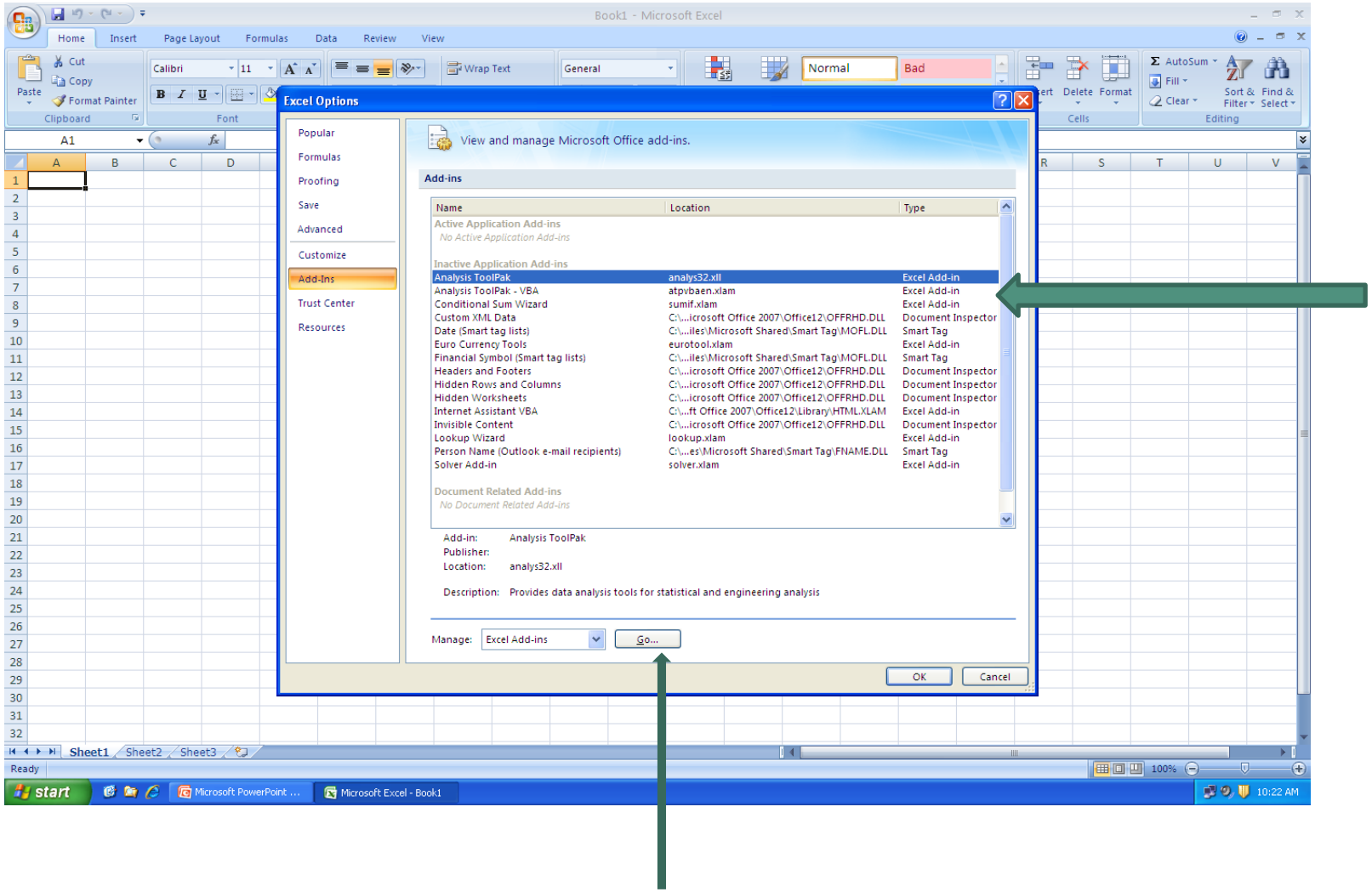
Office Button → Excel Options



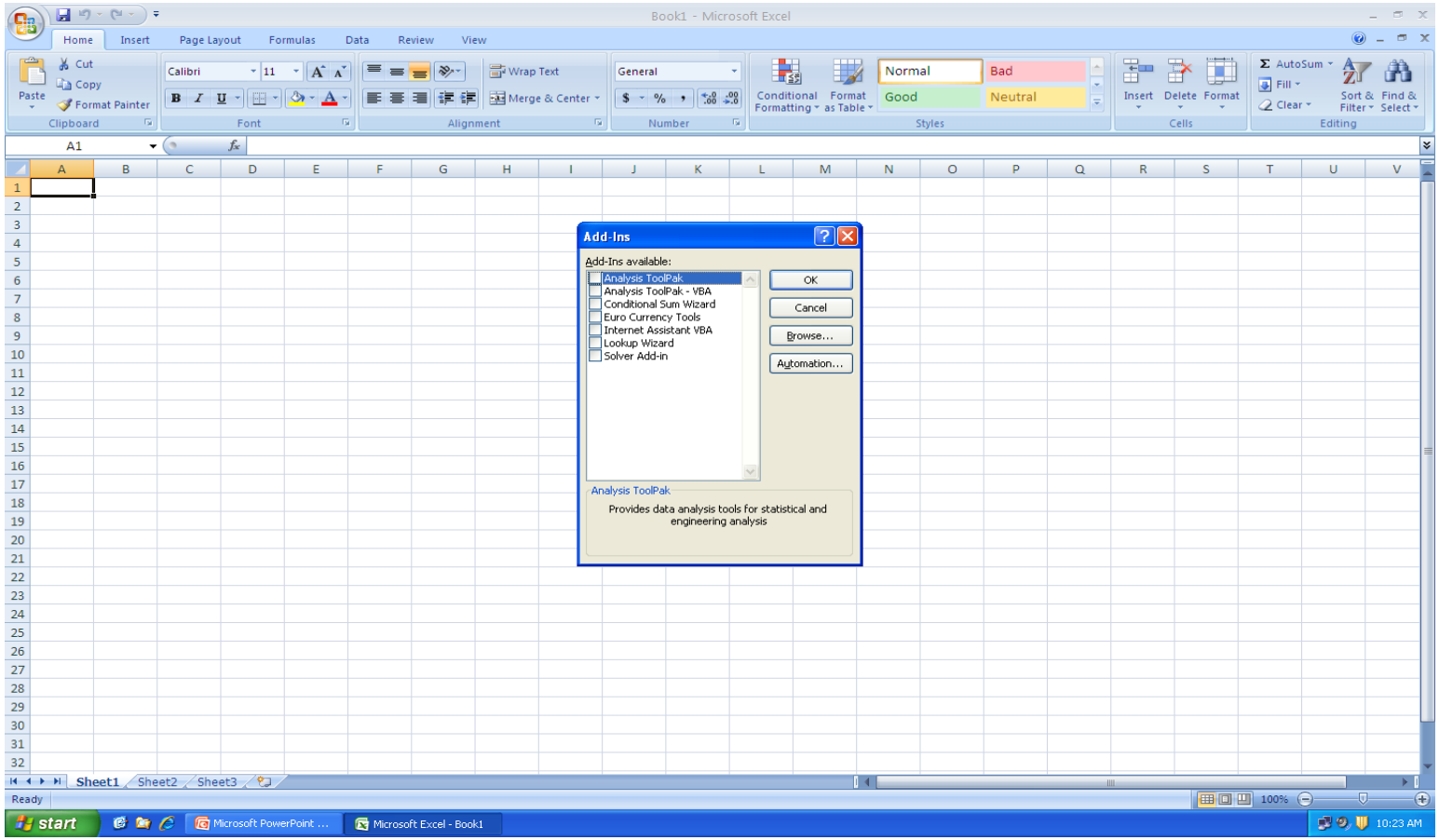
Add-Ins


Book1 - Microsoft Excel
Excel Options
Popular
Formulas
Proofing
Save
Advanced
Customize
Add-Ins
Trust Center
Resources
Change the most popular options in Excel.
Top options for working with Excel
Show Mini Toolbar on selection
Enable Live Preview
Show Developer tab in the Ribbon
Always use ClearType
Color scheme: Blue
ScreenTip style: Show feature descriptions in ScreenTips
Create lists for use in sorts and fill sequences: Edit Custom Lists...
When creating new workbooks
Use this font: Body Font
Font size: 11
Default view for new sheets: Normal View
Include this many sheets: 3
Personalize your copy of Microsoft Office
User name: student
Choose the languages you want to use with Microsoft Office: Language Settings...
OK
Cancel
Sheet1
Sheet2
Sheet3
Ready
start
Microsoft PowerPoint ...
Microsoft Excel - Book1
10:21 AM

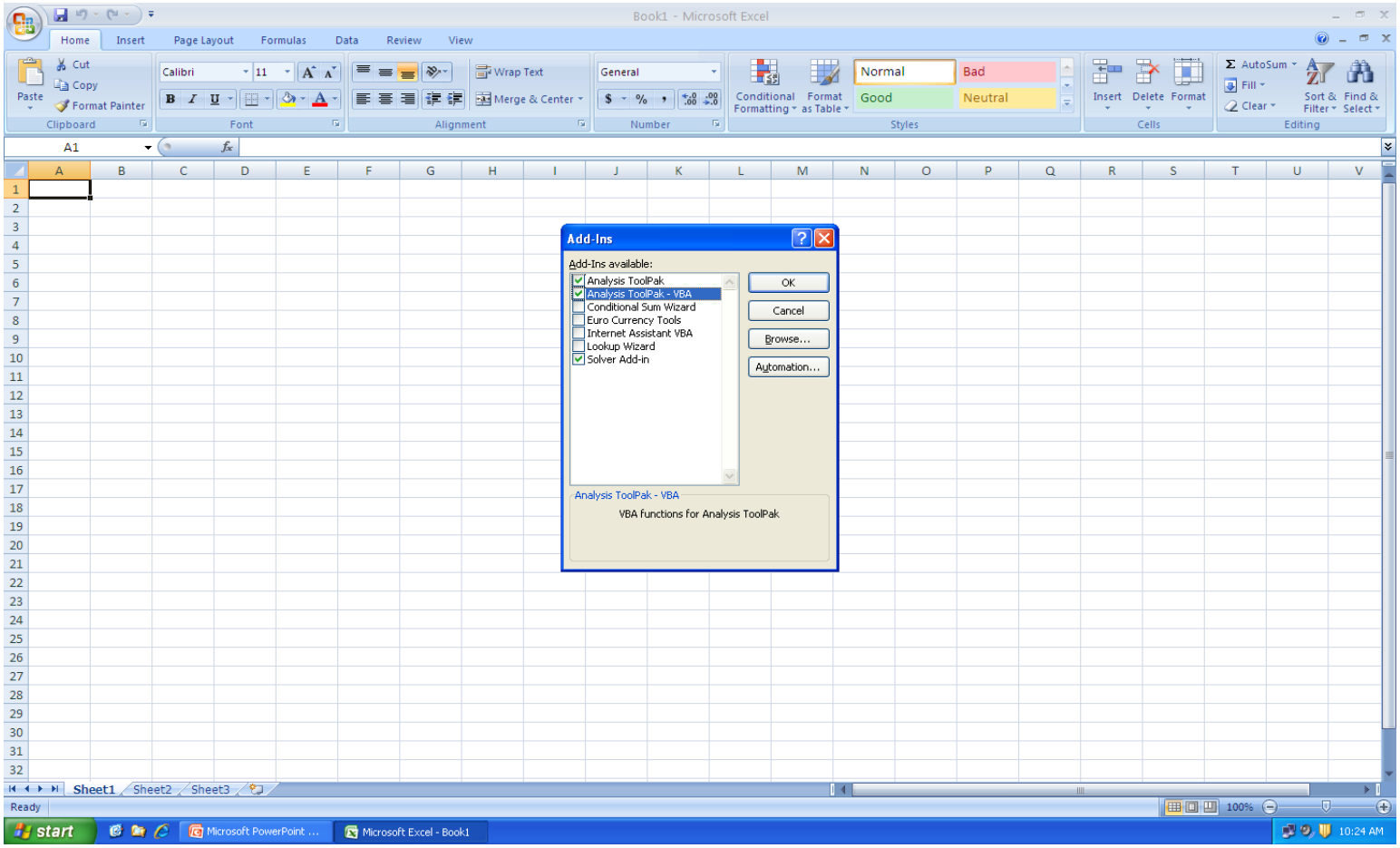
Analysis ToolPak → Go



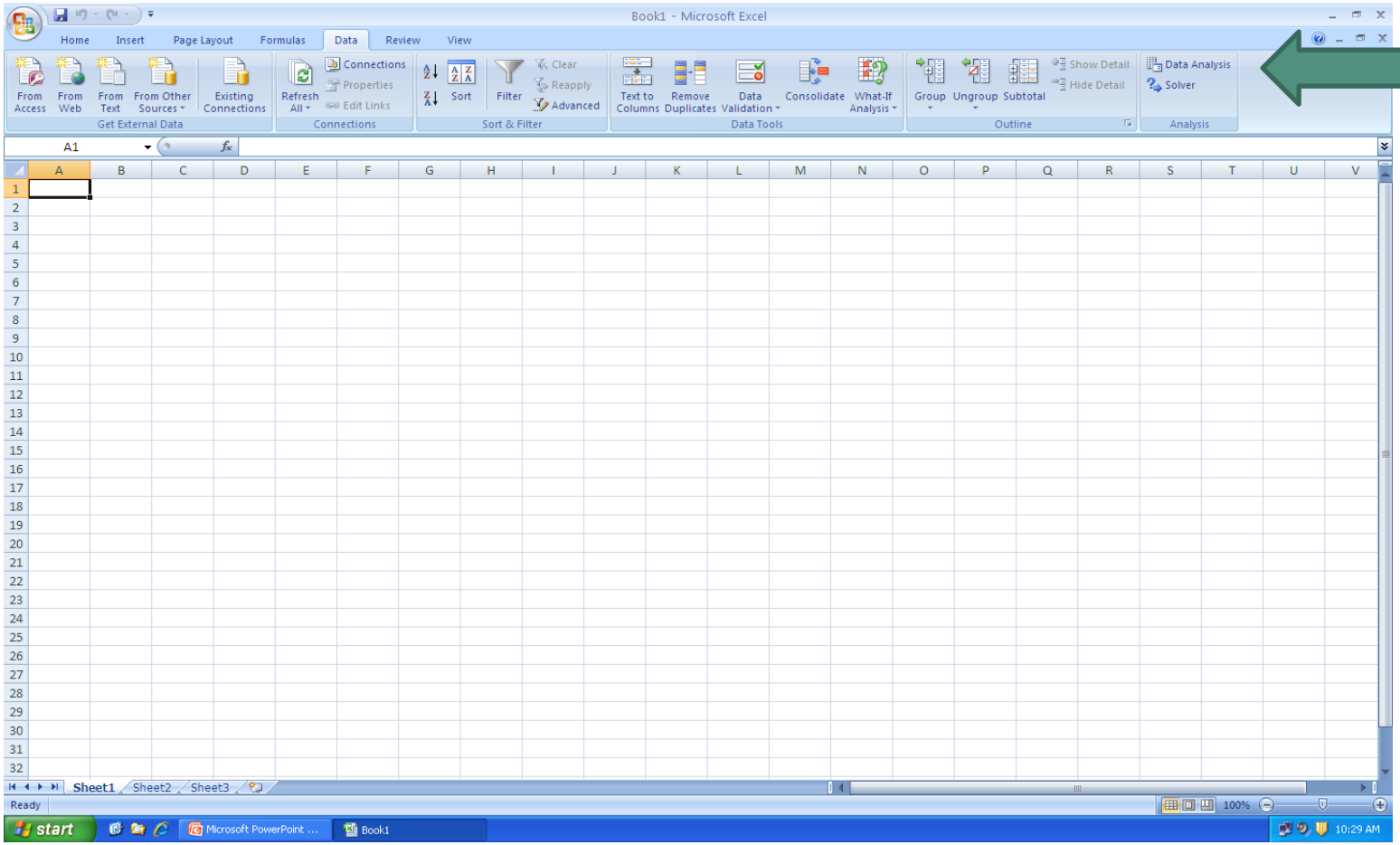
Add In Selections



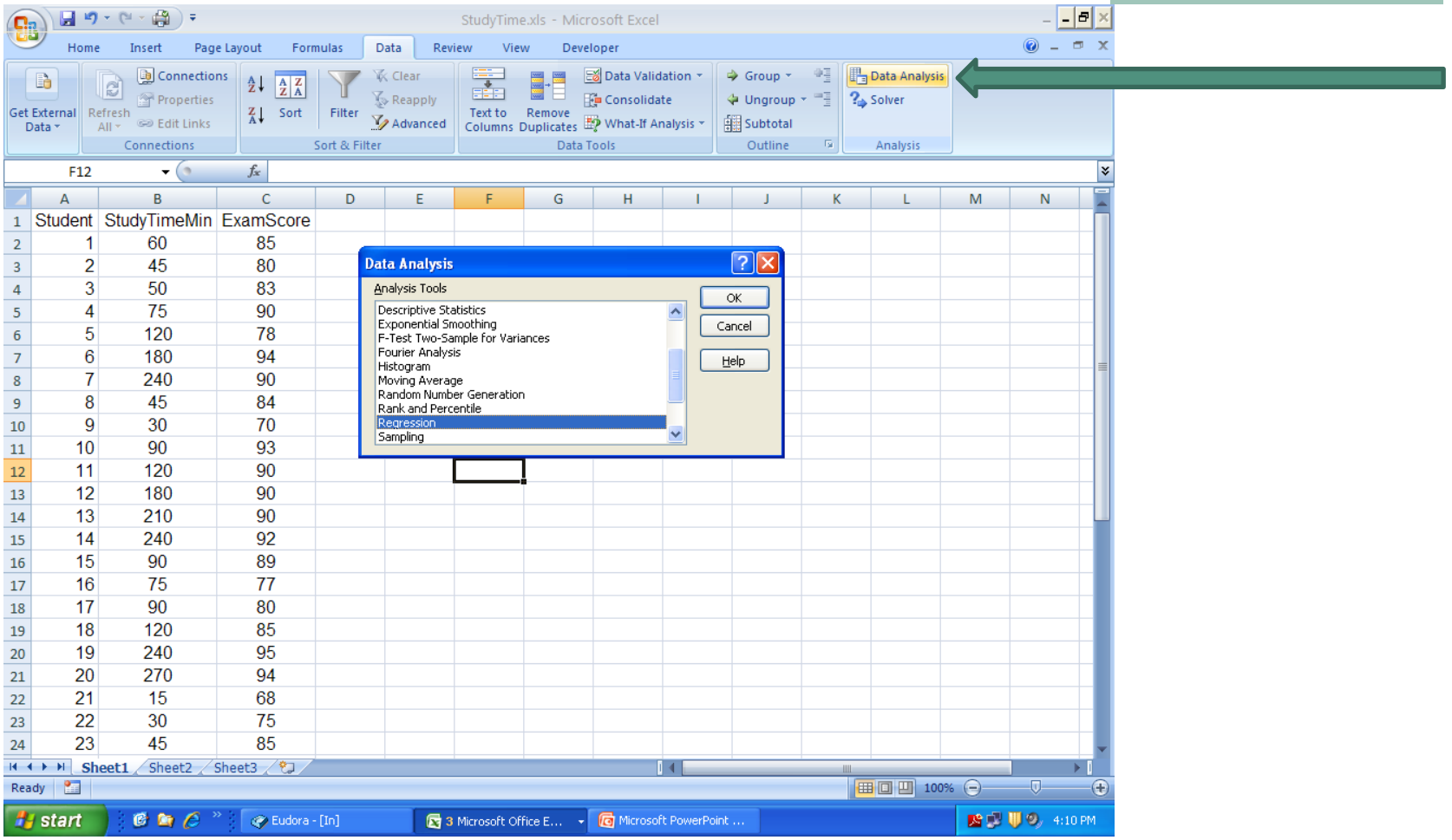
Check-off Options → Ok



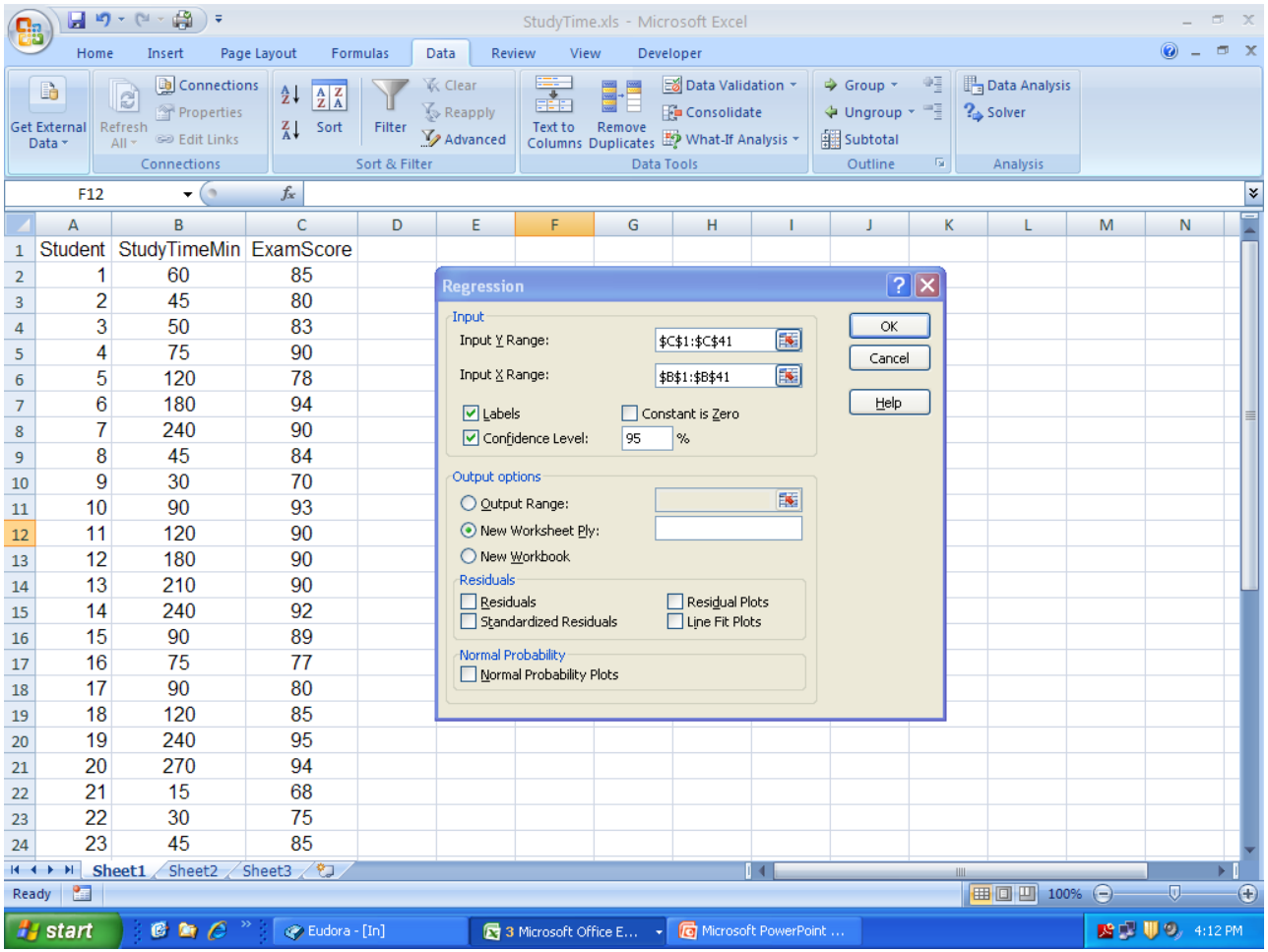
Data Tab → Analysis



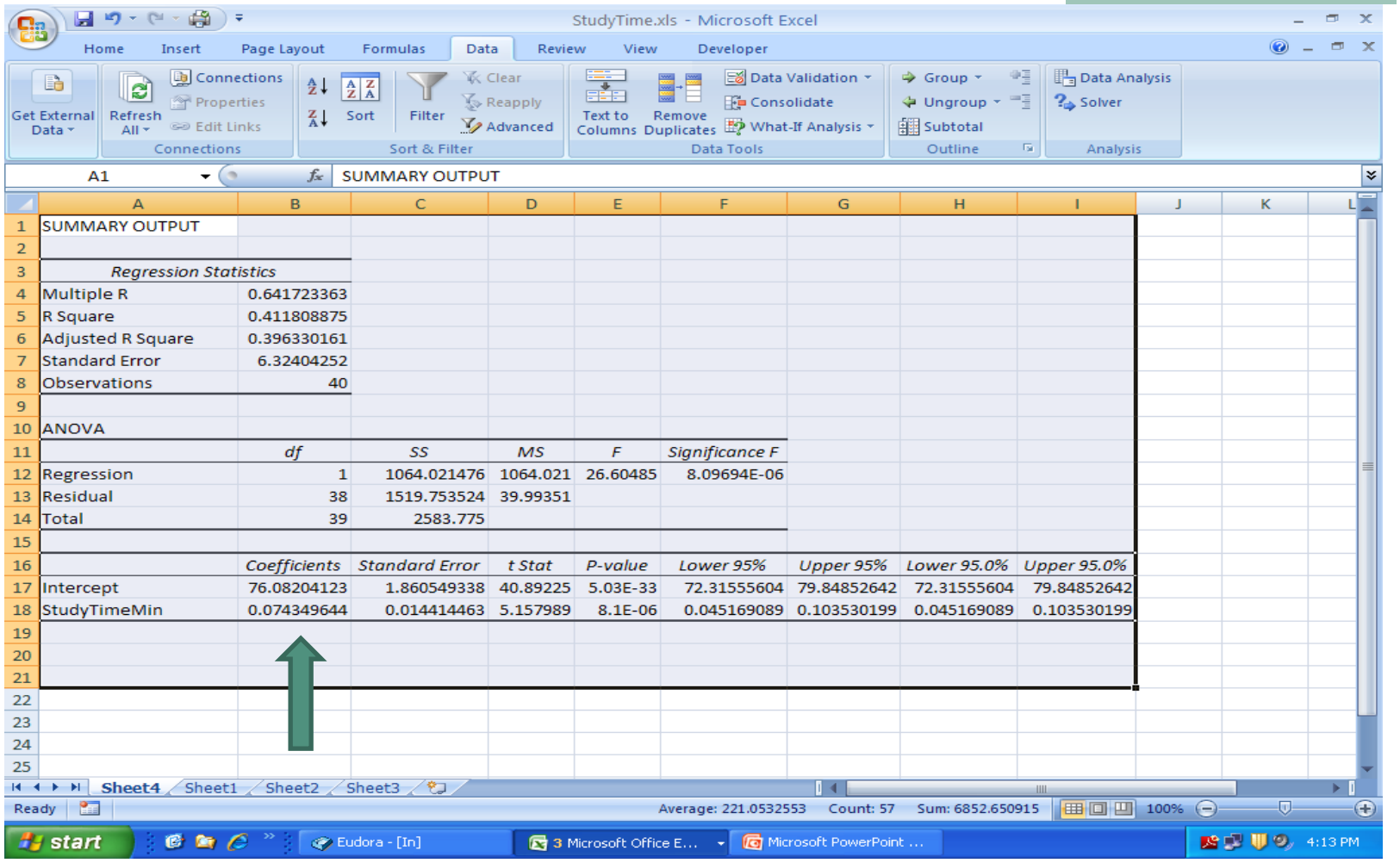
Study Time Model Using Excel



Regression in Excel (con't)

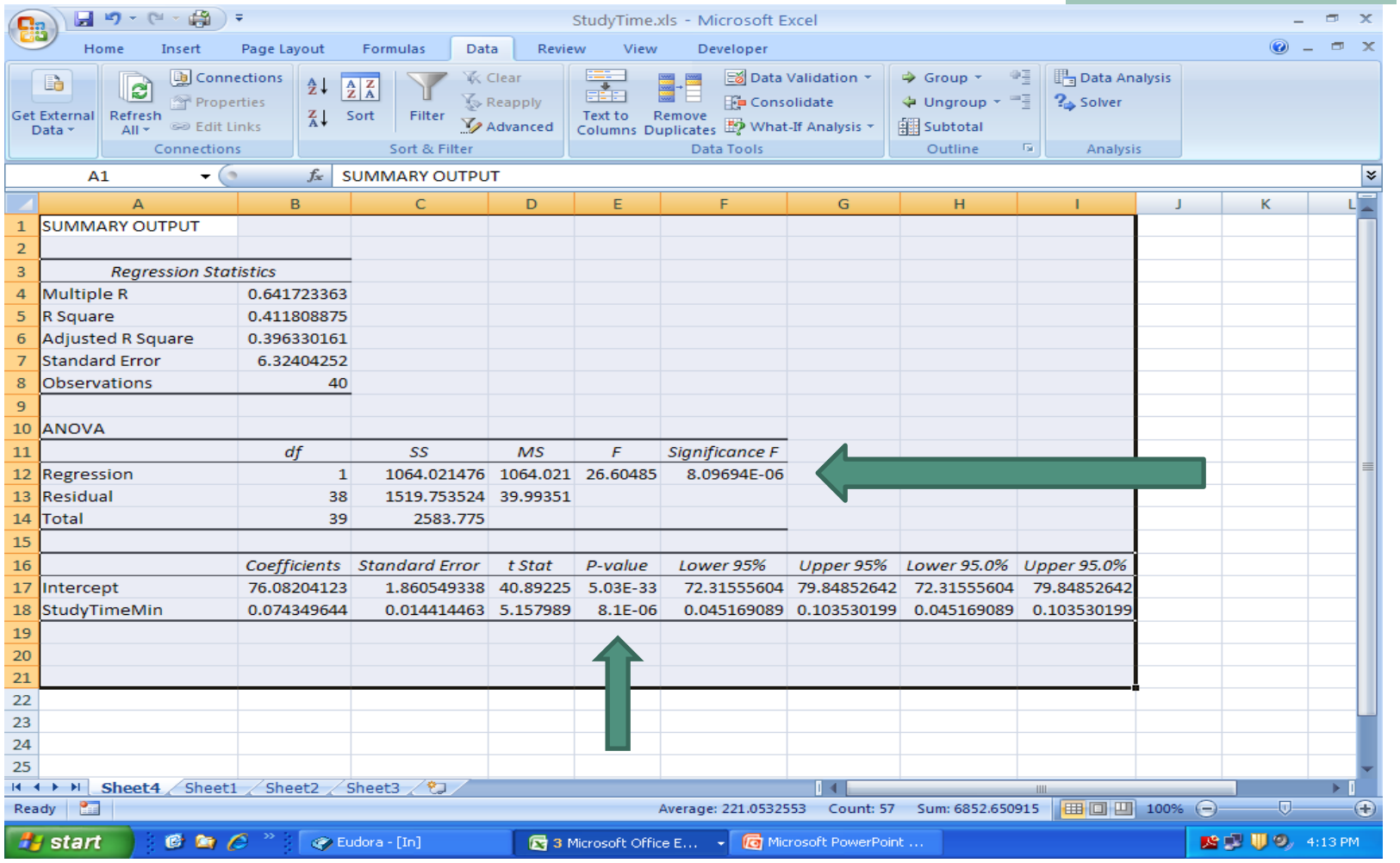


Coefficients: Intercept & Slope



Regression in Excel (con't)

[F and p-value are the same for SLR]



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.641723363							
5	R Square	0.411808875							
6	Adjusted R Square	0.396330161							
7	Standard Error	6.32404252							
8	Observations	40							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	1064.021476	1064.021	26.60485	8.09694E-06			
13	Residual	38	1519.753524	39.99351					
14	Total	39	2583.775						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	76.08204123	1.860549338	40.89225	5.03E-33	72.31555604	79.84852642	72.31555604	79.84852642
18	StudyTimeMin	0.074349644	0.014414463	5.157989	8.1E-06	0.045169089	0.103530199	0.045169089	0.103530199
19									
20									
21									
22									
23									
24									
25									

Dimensions of Regression Coefficients

- The units of the intercept are the same units as the dependent variable
- The units of the coefficients for the independent variables are the dependent variable units divided by the independent variable units
- For example in the relationship between house price in dollars (dependent variable) and house square footage (independent variable):
  - The intercept is in dollars
  - The slope is in dollars per square foot

- For our example here of study time (in minutes) versus exam score (in %):
  - What are the units of the intercept ?
  - What are the units of the slope ?

Wait….

?

Don’t look ahead, until you have your answer !

- For our example here of study time (in minutes) versus exam score (points):
  - Units of the intercept - points
  - Units of the slope - points/min

Using The Regression Equation For Prediction

- If you wanted to predict what your exam score would be if you studied for 90 minutes:

  $\hat{y} = 76.072 + .0743(90)$

  $\hat{y} = 82.7$

- You can expect to get an 82.7% on the exam if you study for 90 minutes
- The regression equation is supposed to improve our ability to predict an outcome value (exam score) because we know the value of the related variable (study time)

MLR in Excel - Sales Data Set

region	sales	advertising	promotions	competitor's sales
Selkirk	101.8	1.3	0.2	20.40
Susquehanna	44.4	0.7	0.2	30.50
Kittery	108.3	1.4	0.3	24.60
Acton	85.1	0.5	0.4	19.60
Finger Lakes	77.1	0.5	0.6	25.50
Berkshire	158.7	1.9	0.4	21.70
Central	180.4	1.2	1.0	6.80
Providence	64.2	0.4	0.4	12.60
Nashua	74.6	0.6	0.5	31.30
Dunster	143.4	1.3	0.6	18.60
Endicott	120.6	1.6	0.8	19.90
Five-Towns	69.7	1.0	0.3	25.60
Waldeboro	67.8	0.8	0.2	27.40
Jackson	106.7	0.6	0.5	24.30
Stowe	119.6	1.1	0.3	13.70

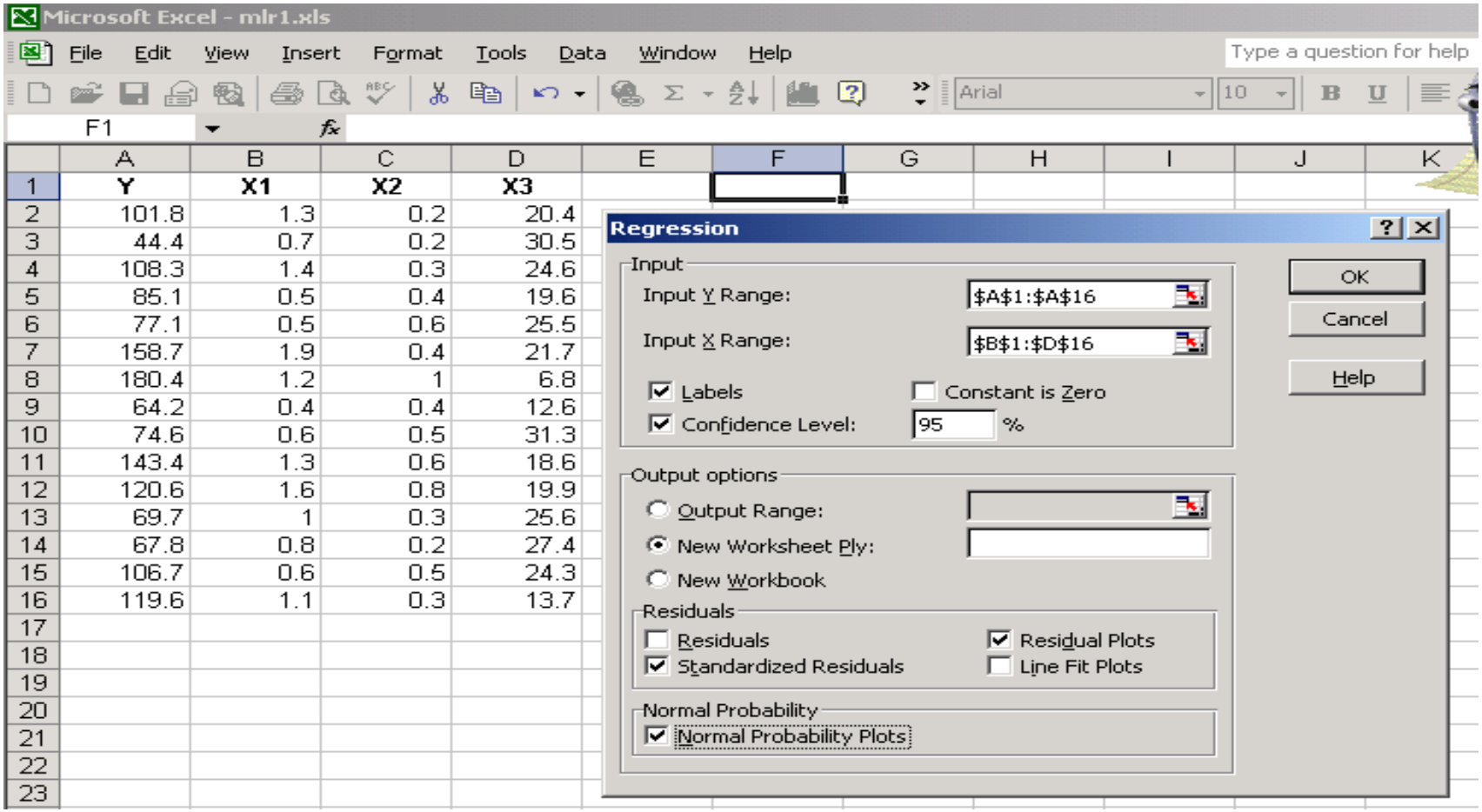
Microsoft Excel - mlr1.xls

File Edit View Insert Format Tools Data

F1

	A	B	C	D
1	**Y**	**X1**	**X2**	**X3**
2	101.8	1.3	0.2	20.4
3	44.4	0.7	0.2	30.5
4	108.3	1.4	0.3	24.6
5	85.1	0.5	0.4	19.6
6	77.1	0.5	0.6	25.5
7	158.7	1.9	0.4	21.7
8	180.4	1.2	1	6.8
9	64.2	0.4	0.4	12.6
10	74.6	0.6	0.5	31.3
11	143.4	1.3	0.6	18.6
12	120.6	1.6	0.8	19.9
13	69.7	1	0.3	25.6
14	67.8	0.8	0.2	27.4
15	106.7	0.6	0.5	24.3
16	119.6	1.1	0.3	13.7
17				

Regression



	A	B	C	D
1	Y	X1	X2	X3
2	101.8	1.3	0.2	20.4
3	44.4	0.7	0.2	30.5
4	108.3	1.4	0.3	24.6
5	85.1	0.5	0.4	19.6
6	77.1	0.5	0.6	25.5
7	158.7	1.9	0.4	21.7
8	180.4	1.2	1	6.8
9	64.2	0.4	0.4	12.6
10	74.6	0.6	0.5	31.3
11	143.4	1.3	0.6	18.6
12	120.6	1.6	0.8	19.9
13	69.7	1	0.3	25.6
14	67.8	0.8	0.2	27.4
15	106.7	0.6	0.5	24.3
16	119.6	1.1	0.3	13.7

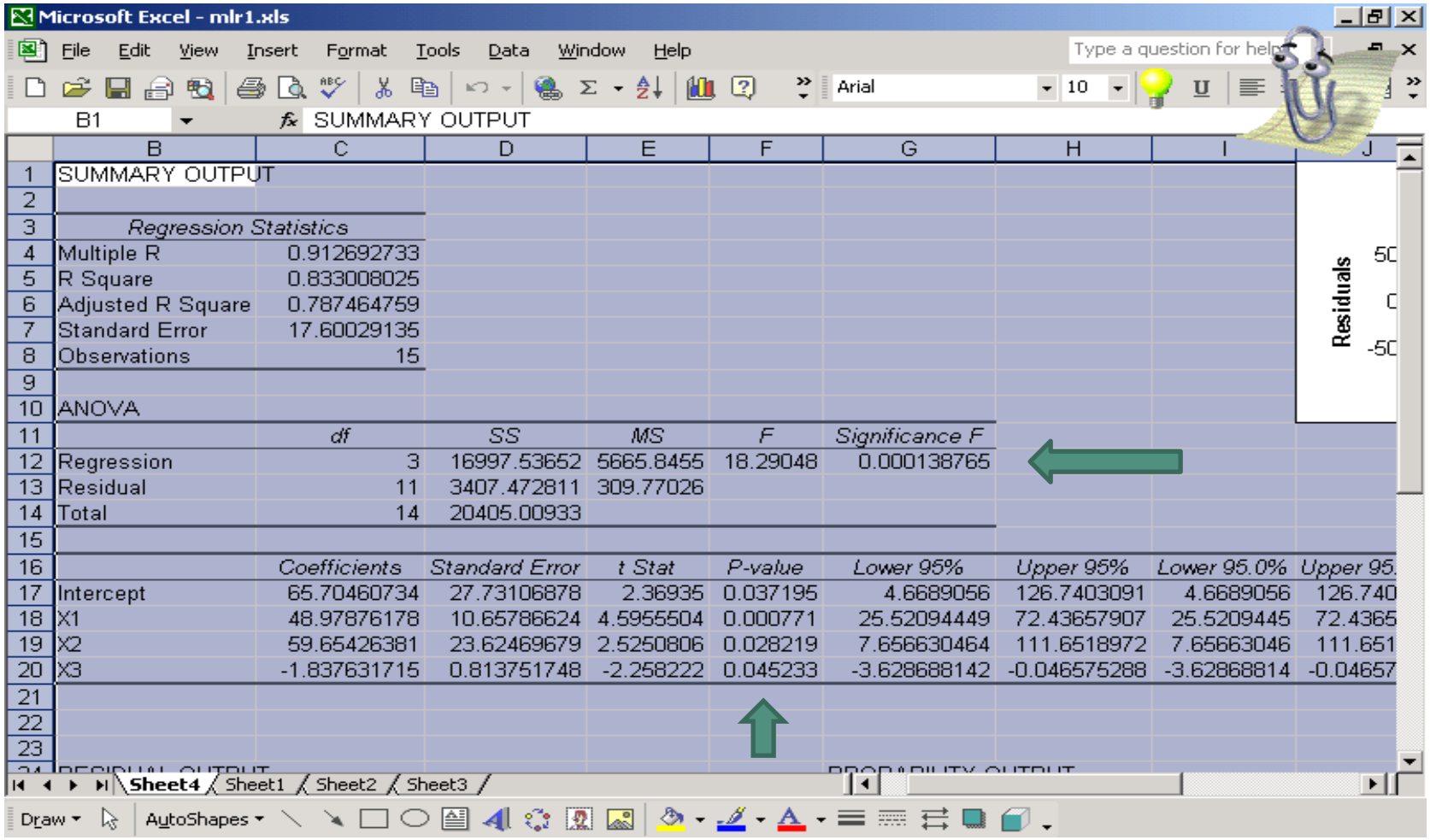
Note that if you include the labels (Y, X1, X2, X3) in the range, you must check the labels box.

Options

- Residuals is the difference between the observed value and the fitted value
- Standardized residuals are scaled to be invariant to the scale of the independent variable; for a good fit (no “outliers”) they should all be between 3 and -3
- Residual plots allow one to visually check the fit and see if they are random or have some pattern; if not random then the MLR model is not appropriate
- The Normal Probability Plot is used to check the normality assumption of the error term; there should be an even spread of data on either side of the 50% mark

Excel Output

[F and p-values are different for MLR]



Microsoft Excel - mlr1.xls

File Edit View Insert Format Tools Data Window Help

Type a question for help

Arial 10

B1 SUMMARY OUTPUT

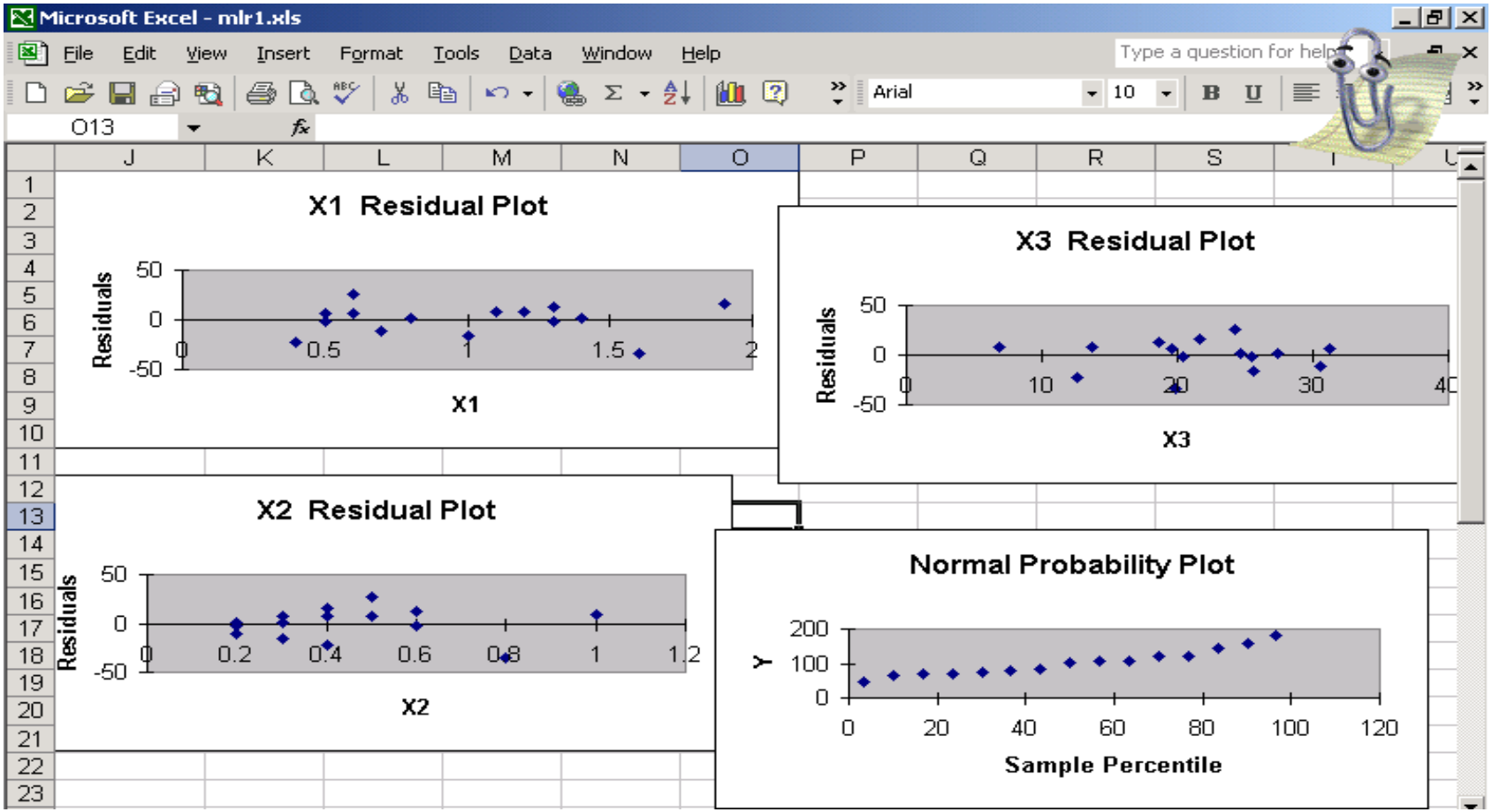
	B	C	D	E	F	G	H	I	J
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.912692733							
5	R Square	0.833008025							
6	Adjusted R Square	0.787464759							
7	Standard Error	17.60029135							
8	Observations	15							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	16997.53652	5665.8455	18.29048	0.000138765			
13	Residual	11	3407.472811	309.77026					
14	Total	14	20405.00933						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.*
17	Intercept	65.70460734	27.73106878	2.36935	0.037195	4.6689056	126.7403091	4.6689056	126.740
18	X1	48.97876178	10.65786624	4.5955504	0.000771	25.52094449	72.43657907	25.5209445	72.4365
19	X2	59.65426381	23.62469679	2.5250806	0.028219	7.656630464	111.6518972	7.65663046	111.651
20	X3	-1.837631715	0.813751748	-2.258222	0.045233	-3.628688142	-0.046575288	-3.62868814	-0.04657
21									
22									
23									

Residuals 50 0 -50

Sheet4 Sheet1 Sheet2 Sheet3

Draw AutoShapes

Residual Plots



Pizza Sales – What is the regression formula ?

	A	B	C
1	PizzaStore	CampusSize	Sales
2	1	2	58
3	2	6	105
4	3	8	88
5	4	8	118
6	5	12	117
7	6	16	137
8	7	20	157
9	8	20	169
10	9	22	149
11	10	26	202

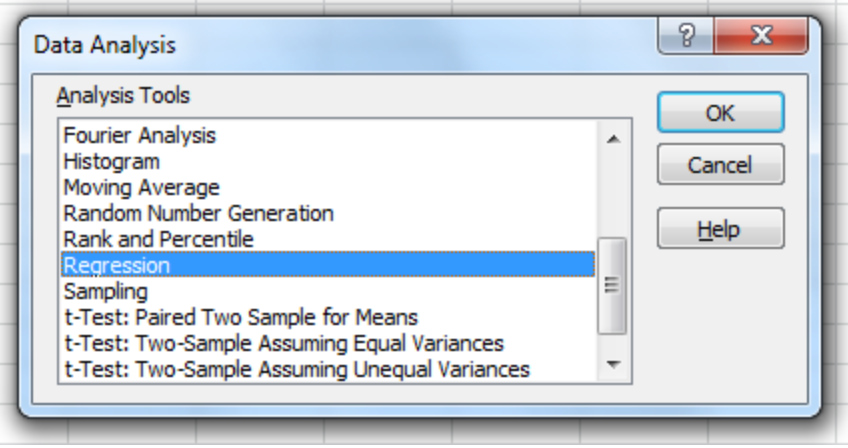
Copyright – Dan Brandon

Wait….

?

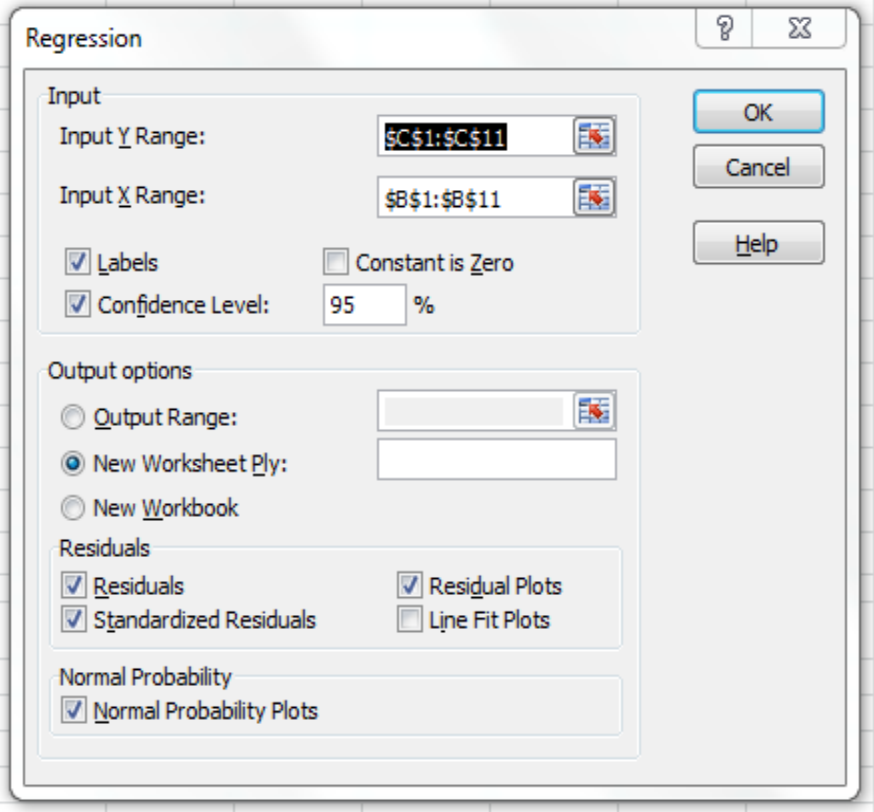
Don’t look ahead, until you have your answer !

	A	B	C
1	PizzaStore	CampusSize	Sales
2	1	2	58
3	2	6	105
4	3	8	88
5	4	8	118
6	5	12	117
7	6	16	137
8	7	20	157
9	8	20	169
10	9	22	149
11	10	26	202
12			
13			
14			
15			



Copyright – Dan Brandon

	A	B	C
1	PizzaStore	CampusSize	Sales
2	1	2	58
3	2	6	105
4	3	8	88
5	4	8	118
6	5	12	117
7	6	16	137
8	7	20	157
9	8	20	169
10	9	22	149
11	10	26	202



Copyright – Dan Brandon

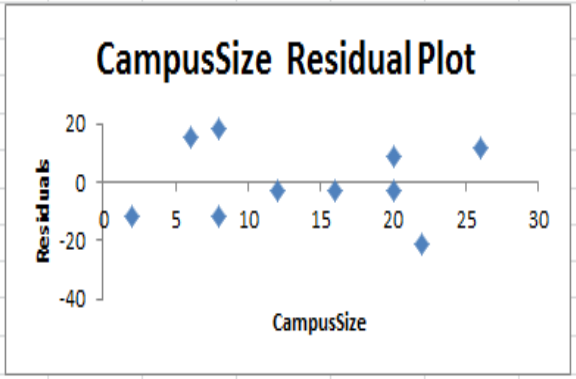
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.950122955
R Square	0.90273363
Adjusted R Square	0.890575334
Standard Error	13.82931669
Observations	10

ANOVA

	df	SS	MS	F	Significance F
Regression	1	14200	14200	74.24836601	2.54887E-05
Residual	8	1530	191.25		
Total	9	15730			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	60	9.22603481	6.503336	0.000187444	38.72472558	81.27527	38.72473	81.27527
CampusSize	5	0.580265238	8.616749	2.54887E-05	3.661905962	6.338094	3.661906	6.338094



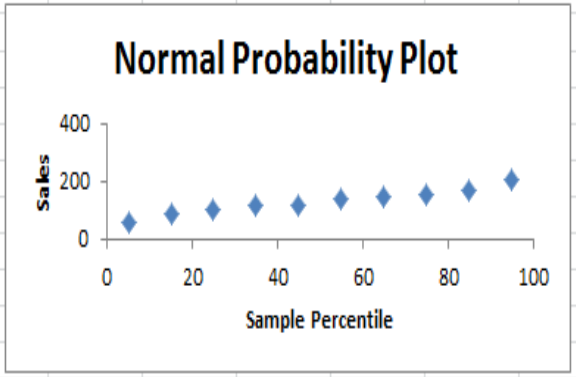
Y(sales) = 60 + 5*X(size)

RESIDUAL OUTPUT

Observation	Predicted Sales	Residuals	dard Residuals
1	70	-12	-0.92036
2	90	15	1.150447
3	100	-12	-0.92036
4	100	18	1.380537
5	120	-3	-0.23009
6	140	-3	-0.23009
7	160	-3	-0.23009
8	160	9	0.690268
9	170	-21	-1.61063
10	190	12	0.920358

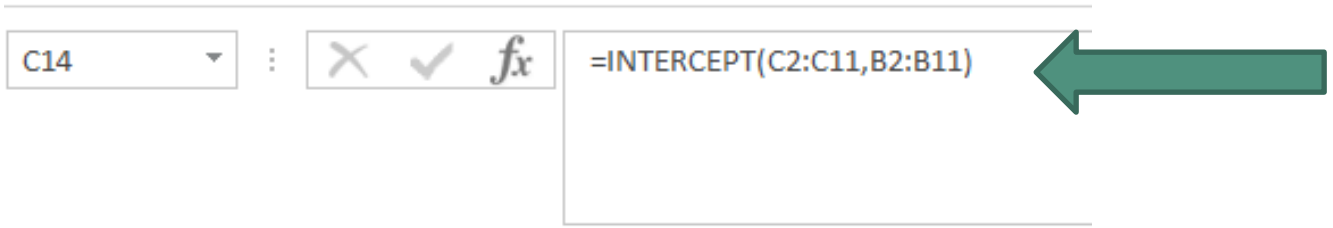
PROBABILITY OUTPUT

Percentile	Sales
5	58
15	88
25	105
35	117
45	118
55	137
65	149
75	157
85	169
95	202



Copyright – Dan Brandon

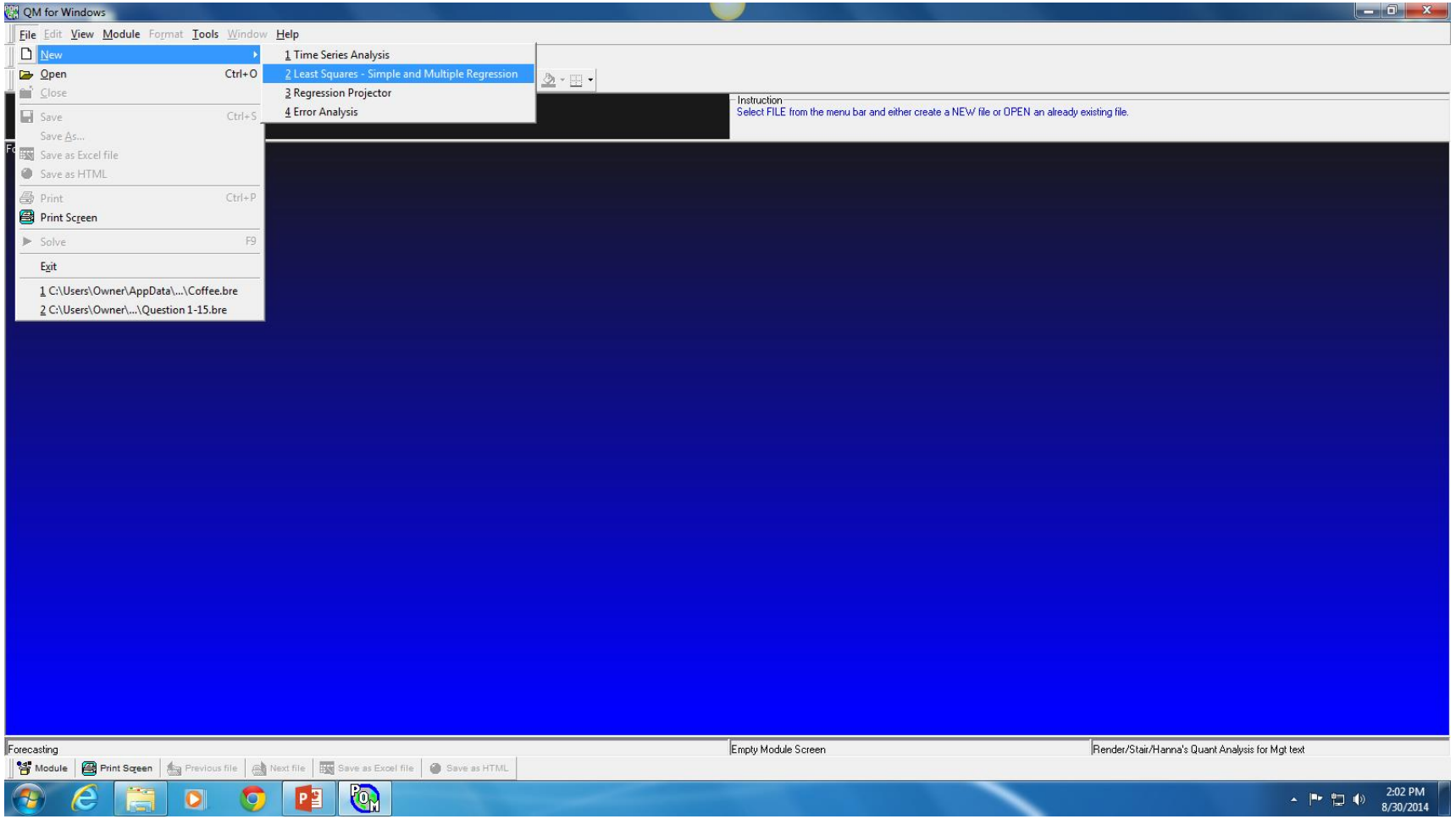
Excel Intercept and Slope Functions



	A	B	C	D	E	F
1	PizzaStore	CampusSize	Sales			
2	1	2	58			
3	2	6	105			
4	3	8	88			
5	4	8	118			
6	5	12	117			
7	6	16	137			
8	7	20	157			
9	8	20	169			
10	9	22	149			
11	10	26	202			
12						
13						
14		Intercept	60			
15		Slope	5			
16						

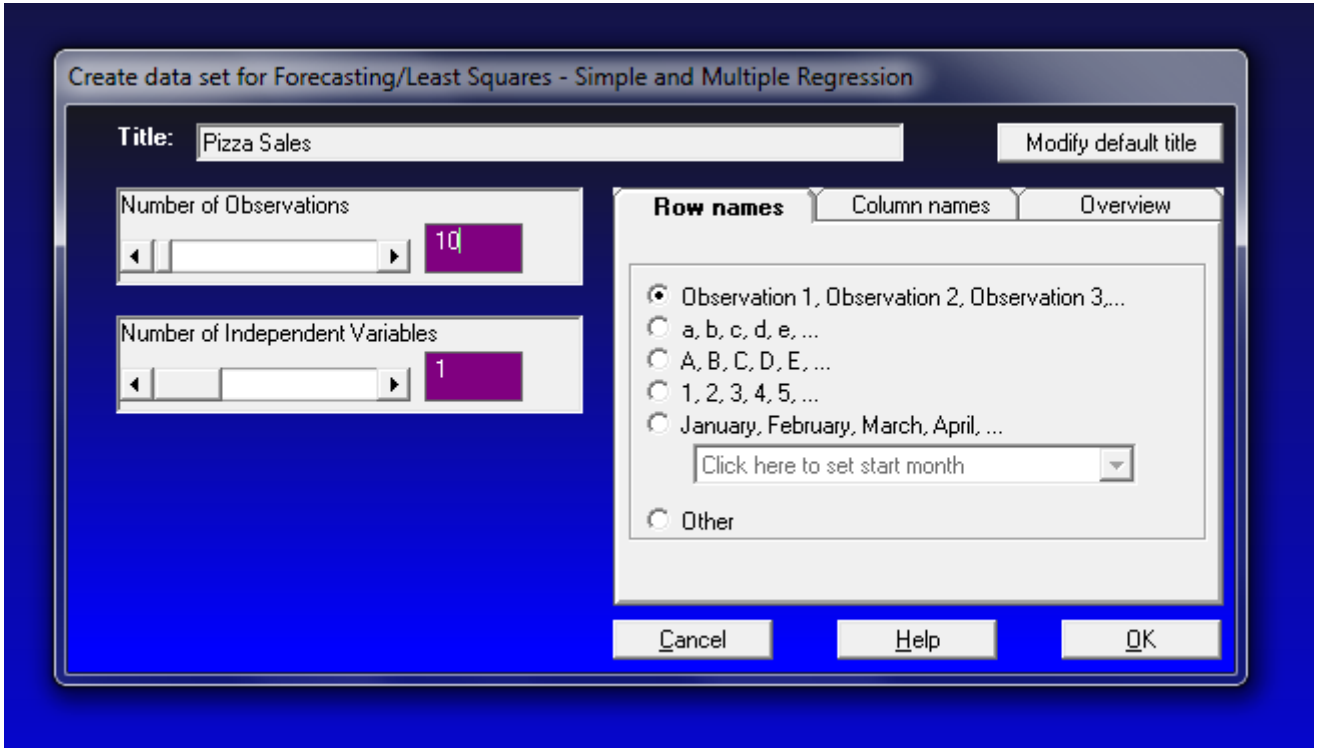
Copyright – Dan Brandon

Using QM
[need to go into “forecasting”]



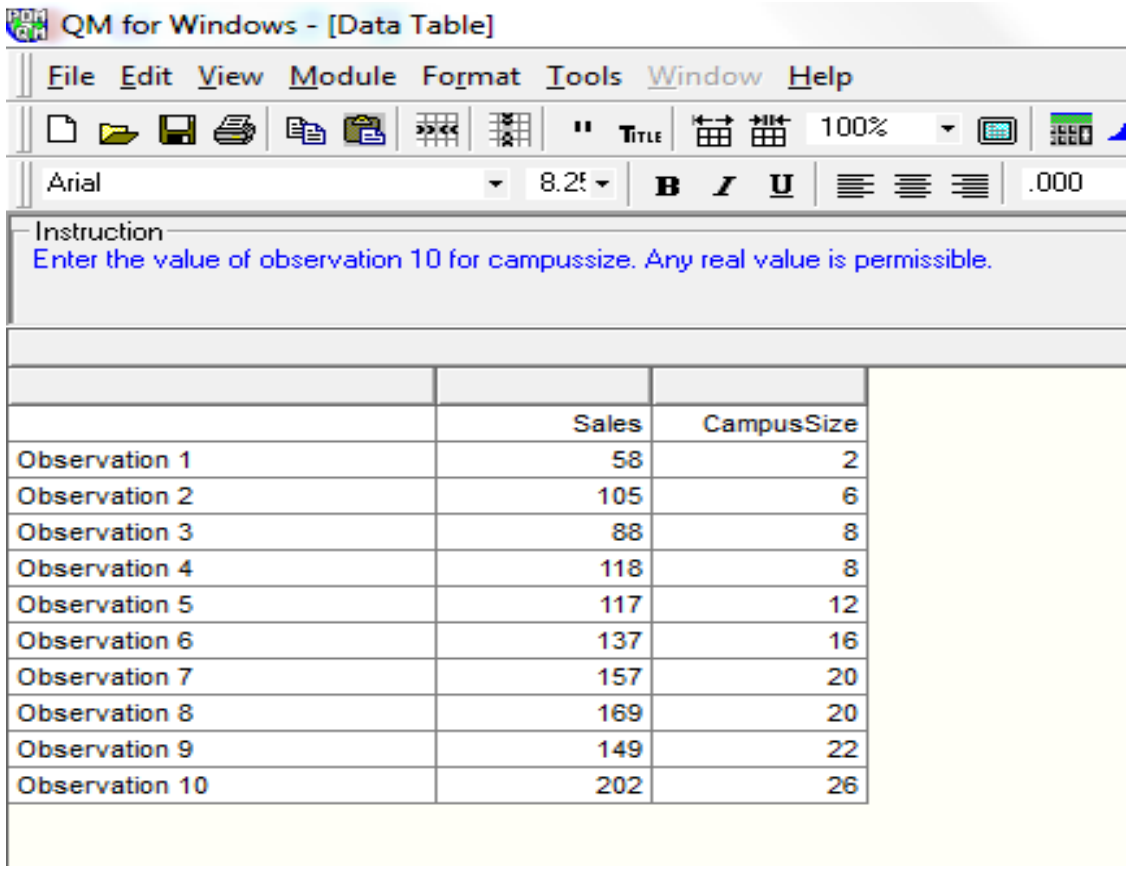
Copyright – Dan Brandon

Using QM (con't)



Copyright – Dan Brandon

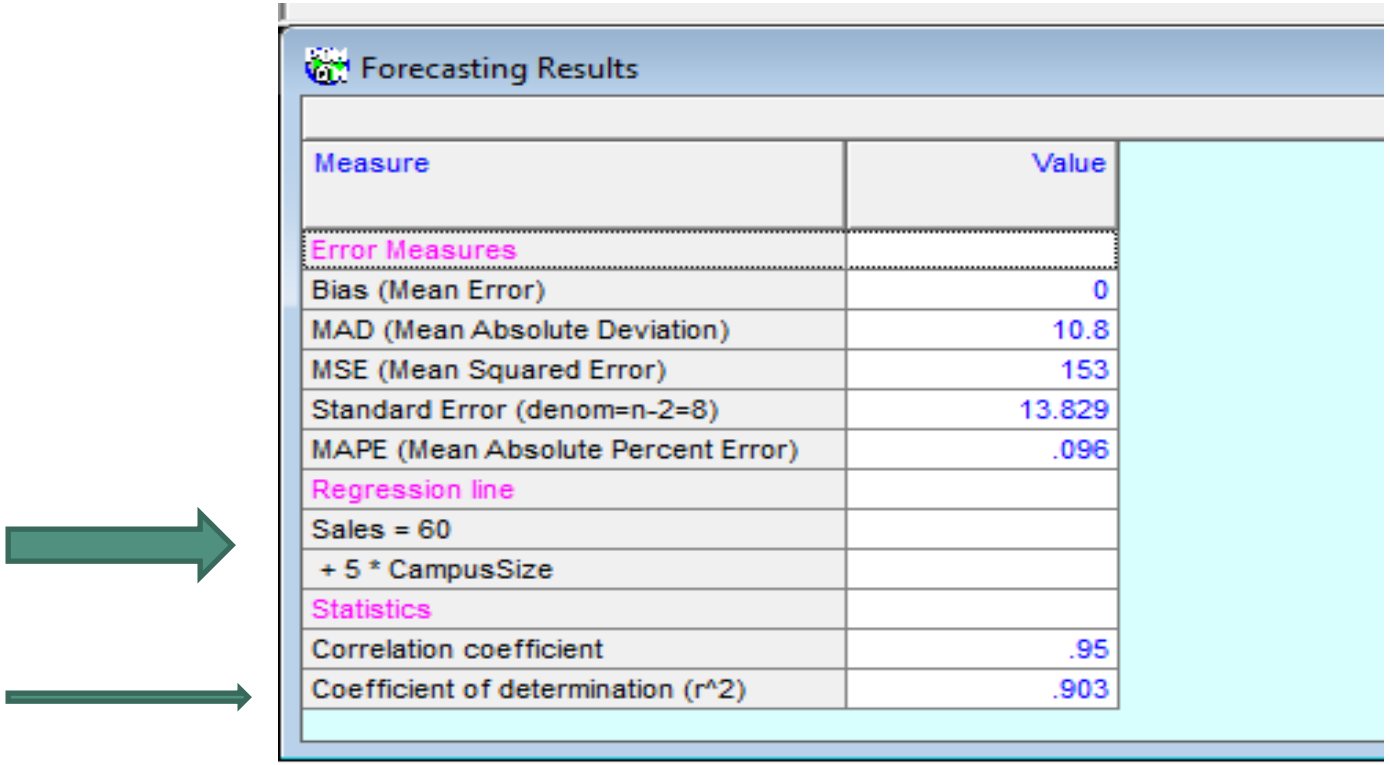
Using QM (con't)



	Sales	CampusSize
Observation 1	58	2
Observation 2	105	6
Observation 3	88	8
Observation 4	118	8
Observation 5	117	12
Observation 6	137	16
Observation 7	157	20
Observation 8	169	20
Observation 9	149	22
Observation 10	202	26

Copyright – Dan Brandon

Using QM (con't)

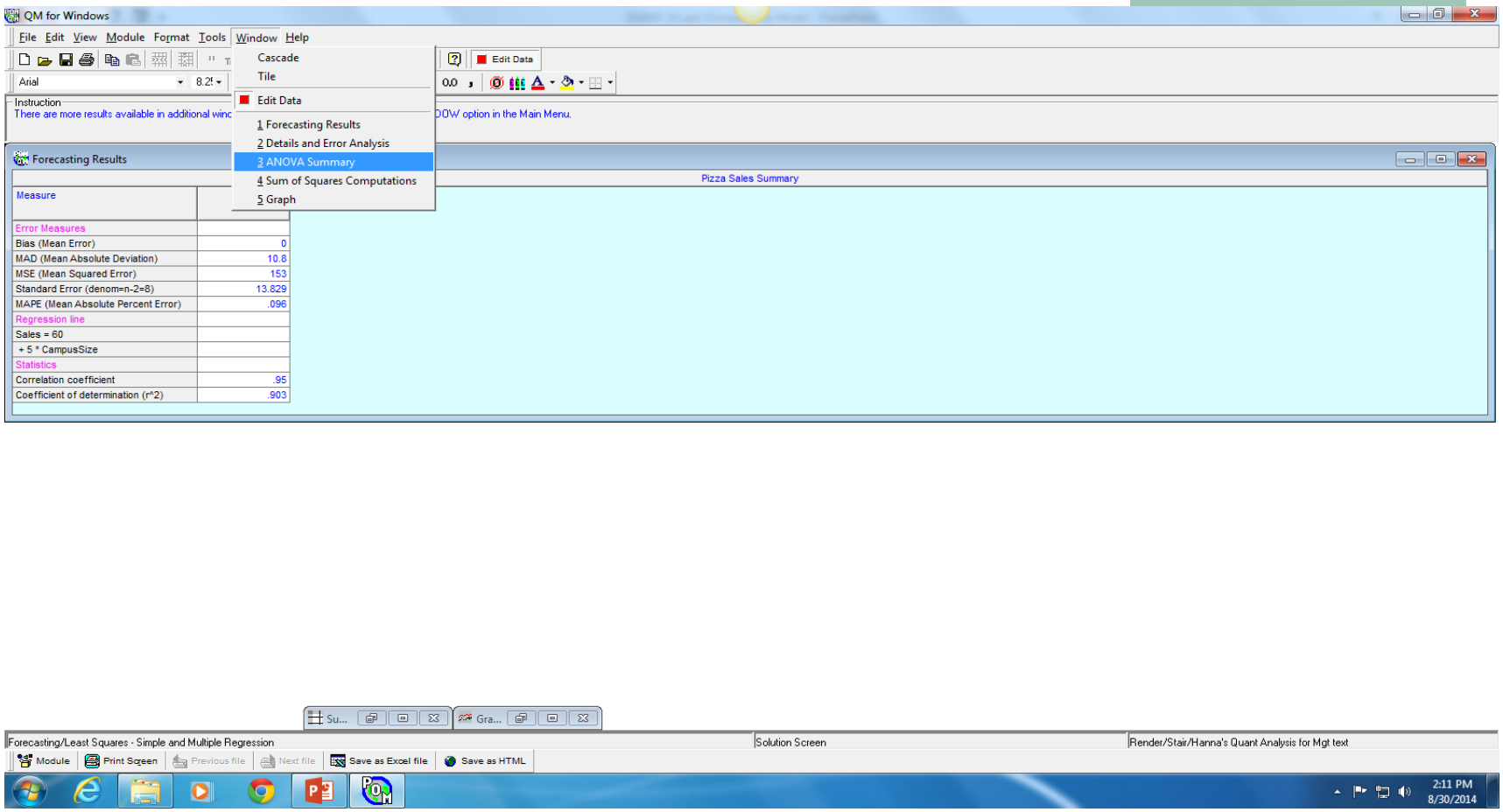


Forecasting Results

Measure	Value
Error Measures	
Bias (Mean Error)	0
MAD (Mean Absolute Deviation)	10.8
MSE (Mean Squared Error)	153
Standard Error (denom=n-2=8)	13.829
MAPE (Mean Absolute Percent Error)	.096
Regression line	
Sales = 60	
+ 5 * CampusSize	
Statistics	
Correlation coefficient	.95
Coefficient of determination (r^2)	.903

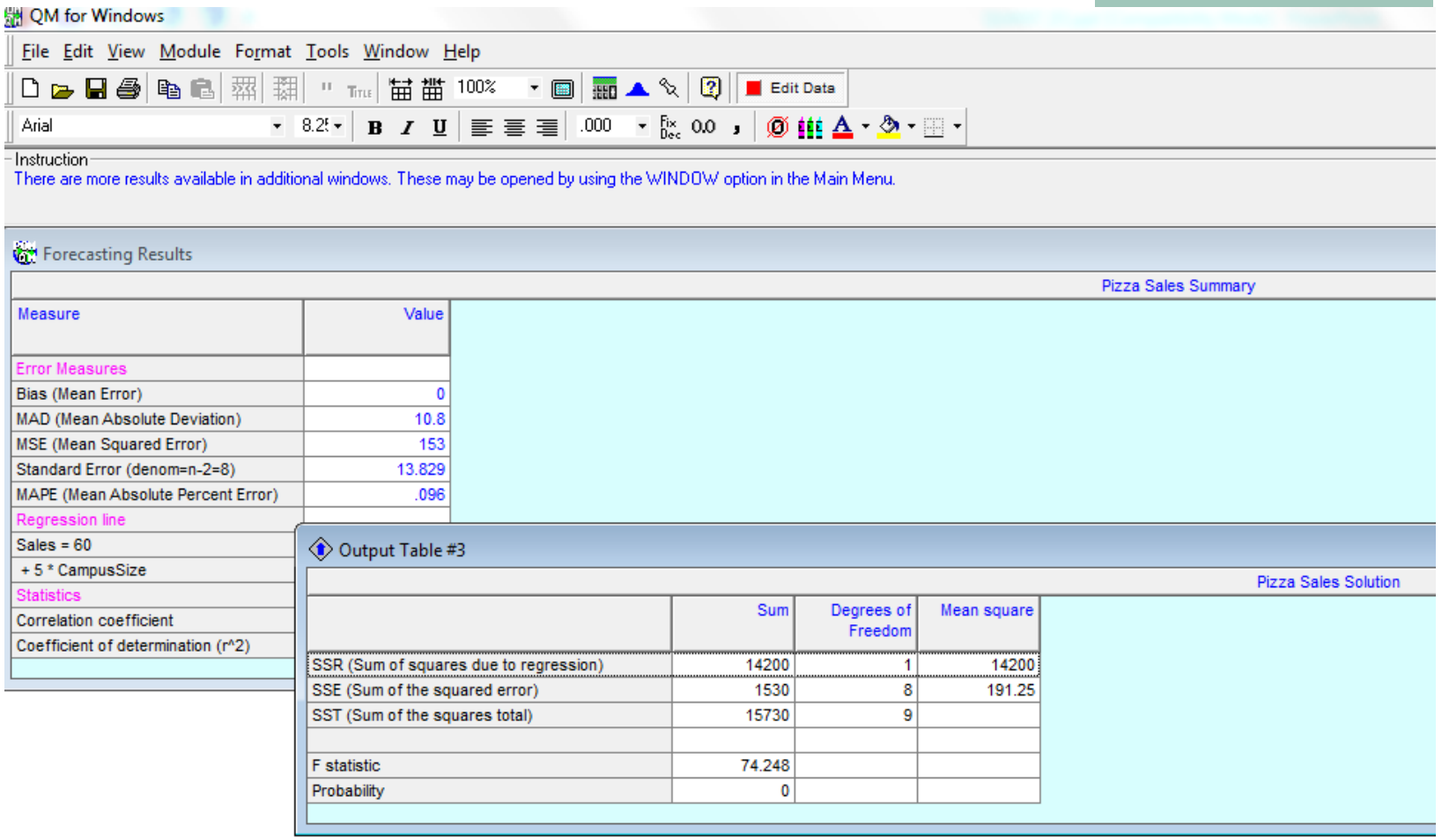
Copyright – Dan Brandon

Using QM (con't)



Copyright – Dan Brandon

Using QM (con't)



Copyright – Dan Brandon

Binary or Dummy Variables

- *Binary* (or *dummy* or *indicator*) variables are special variables created for qualitative data (nominal or ordinal)
- A dummy variable is assigned a value of 1 if a particular condition is met and a value of 0 otherwise
- The number of dummy variables must equal one less than the number of categories of the qualitative variable

Realty Example

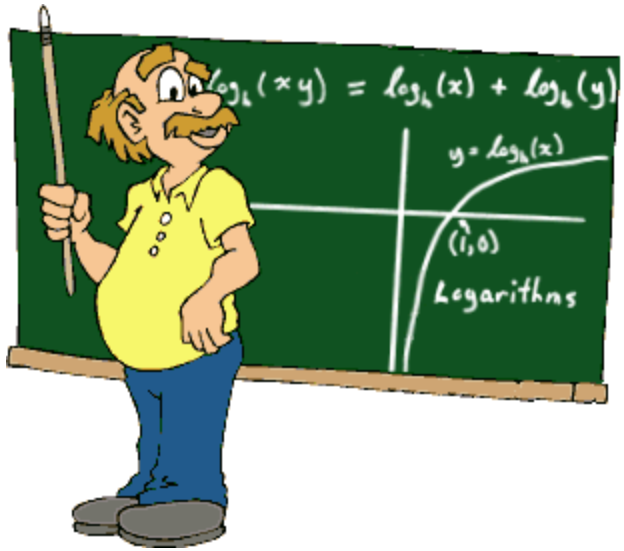
- In a model for house values, a regression model may have independent variables for square footage (X_1) and number of bedrooms (X_2)
- A better model could possibly be developed if information about the condition of the property was included:

X_3 = 1 if house is in excellent condition
= 0 otherwise
X_4 = 1 if house is in mint condition
= 0 otherwise

- Two dummy variables can be used to describe the three categories of condition (good, excellent, mint)
- No variable is needed for "good" condition since if both X_3 and X_4 = 0, the house must be in good condition

Transformation of Data

- Some non-linear data may be transformed into related linear data by applying a function to one or more independent variables:
  - Log function
  - Square root
  - Power function
  - Exponential

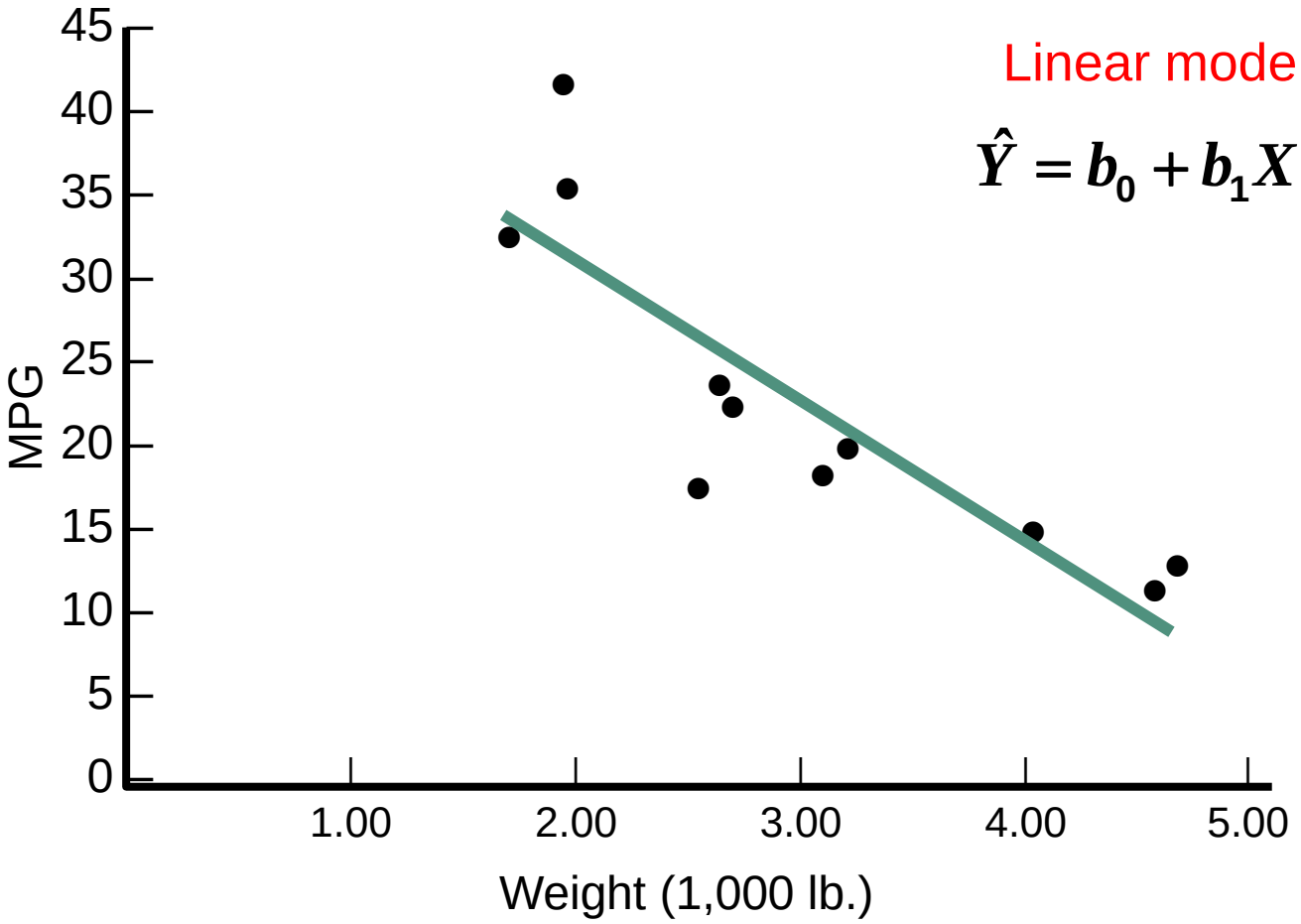


Transformation Example

- Engineers want to use regression analysis to improve fuel efficiency
- They have been asked to study the impact of weight on miles per gallon (MPG)

MPG	WEIGHT (1,000 LBS.)	MPG	WEIGHT (1,000 LBS.)
12	4.58	20	3.18
13	4.66	23	2.68
15	4.02	24	2.65
18	2.53	33	1.70
19	3.09	36	1.95
19	3.11	42	1.92

Transformation Example (con't)



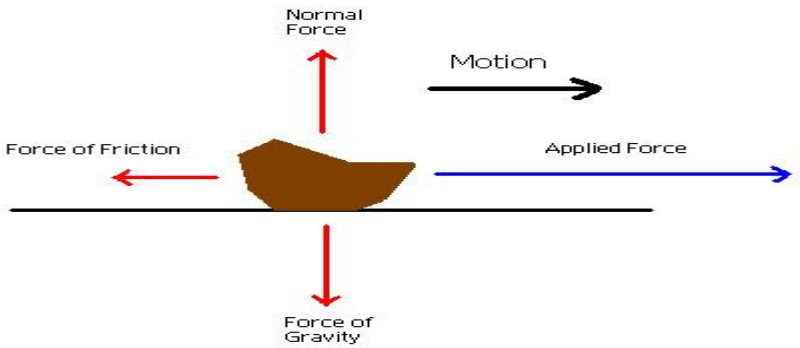
Transformation Example (con't)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Automobile Weight vs. MPG**			SUMMARY OUTPUT									
2													
3	**MPG (Y)**	**Weight (X1)**		*Regression Statistics*									
4	12	4.58		Multiple R	0.8629								
5	13	4.66		R Square	0.7446								
6	15	4.02		Adjusted R Sc	0.7190								
7	18	2.53		Standard Erro	5.0076								
8	19	3.09		Observations	12								
9	19	3.11											
10	20	3.18		ANOVA									
11	23	2.68			*df*	*SS*	*MS*	*F*	*Significance F*				
12	24	2.65		Regression	1	730.9090	730.9090	29.1480	0.0003				
13	33	1.70		Residual	10	250.7577	25.0758						
14	36	1.95		Total	11	981.6667							
15	42	1.92											
16					*Coefficients*	*Standard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0*	*Upper 95.0%*	
17				Intercept	47.6193	4.8132	9.8936	0.0000	36.8950	58.3437	36.8950	58.3437	
18				Weight	-8.2460	1.5273	-5.3989	0.0003	-11.6491	-4.8428	-11.6491	-4.8428	

- A useful model with a small F-test for significance and a good r^2 value

Transformation Example (con't)

- What is a more likely relationship between MPG and weight ?
- Between the power needed to move an object of some weight ?
- Need force to accelerate an object (F=MA) and need force to overcome friction (F=Mg)

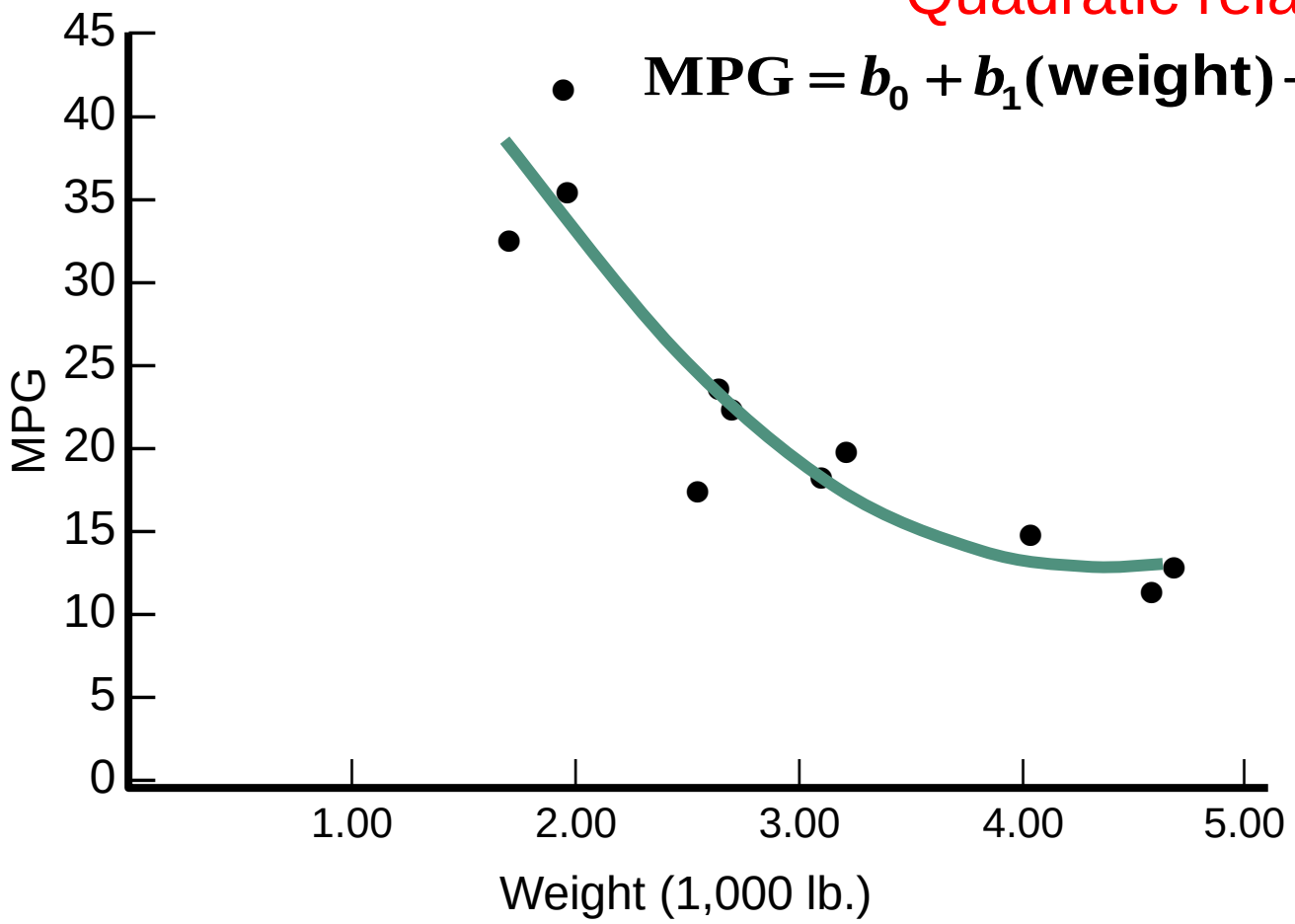


Copyright – Dan Brandon

Transformation Example (con't)

Quadratic relationship

$$\mathbf{MPG} = b_0 + b_1(\mathbf{weight}) + b_2(\mathbf{weight})^2$$



Transformation Example (con't)

- The nonlinear model is a quadratic model
- The easiest way to work with this model is to develop a new variable

$$X_2 = (\textbf{weight})^2$$

- This gives us a model that can be solved with linear regression software

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

Transformation Example (con't)

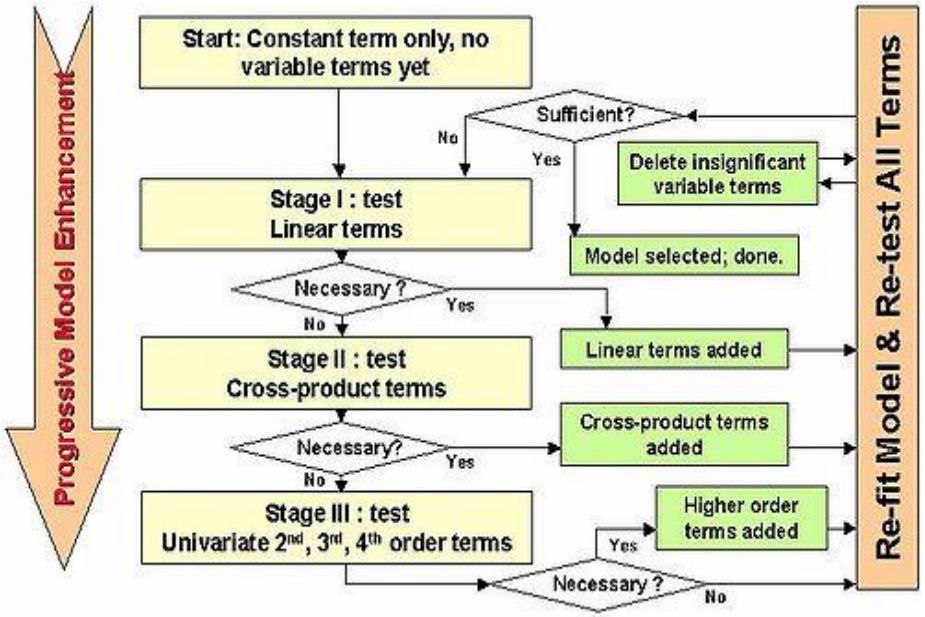
	A	B	C	D	E	F	G	H	I	J	K	L
1	**Automobile Weight vs. MPG**			SUMMARY OUTPUT								
2												
3	**MPG (Y)**	**Weight (X1)**	**WeightSq. (X2)**	*Regression Statistics*								
4	12	4.58	20.98	Multiple R	0.9208							
5	13	4.66	21.72	R Square	0.8478							
6	15	4.02	16.16	Adjusted R Sc	0.8140							
7	18	2.53	6.40	Standard Erro	4.0745							
8	19	3.09	9.55	Observations	12							
9	19	3.11	9.67									
10	20	3.18	10.11	ANOVA								
11	23	2.68	7.18		*df*	*SS*	*MS*	*F*	*Significance F*			
12	24	2.65	7.02	Regression	2	832.2557	416.1278	25.0661	0.0002			
13	33	1.70	2.89	Residual	9	149.4110	16.6012					
14	36	1.95	3.80	Total	11	981.6667						
15	42	1.92	3.69									
16					*Coefficient*	*Standard Er*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0*	*Upper 95.0%*
17				Intercept	79.7888	13.5962	5.8685	0.0002	49.0321	110.5454	49.0321	110.5454
18				Weight	-30.2224	8.9809	-3.3652	0.0083	-50.5386	-9.9061	-50.5386	-9.9061
19				Weight2	3.4124	1.3811	2.4708	0.0355	0.2881	6.5367	0.2881	6.5367

$$\hat{Y} = 79.8 - 30.2X_1 + 3.4X_2$$

- A better model with a smaller F-test for significance and a larger adjusted r^2 value

Stepwise MLR

- Brings variables into the regression one at a time to determine which variables have the most impact and which variables are really necessary
- Uses adjusted R squared to judge "improvement"
- Can also investigate variable relationships



Regression Pitfalls

- If the assumptions are not met, the statistical test may not be valid
- Correlation does not necessarily mean causation
- Multicollinearity makes interpreting coefficients problematic, but the model may still be good
- Using a regression model beyond the range of X is questionable, the relationship may not hold outside the sample data
- A linear relationship may not be the best relationship, even if the F-test returns an acceptable value
- A nonlinear relationship can exist even if a linear relationship does not
- Just because a relationship is statistically significant doesn't mean it has any practical value

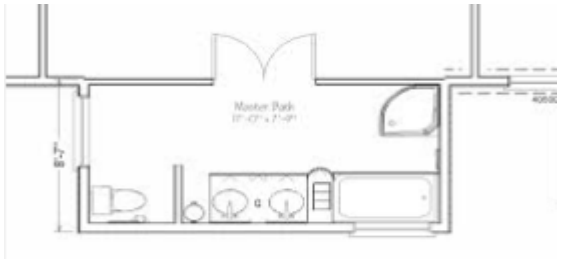
References

- Data Analysis Using Regression and Multilevel/Hierarchical Models by Andrew Gelman and Jennifer Hill
- Applied Regression Analysis and Generalized Linear Models by John Fox
- Regression Analysis by Example (Wiley Series in Probability and Statistics) by Samprit Chatterjee and Ali S. Hadi

Homework

- Textbook Chapter 4
- Quiz on these slides and Chapter 4
- Questions to be answered: 1, 2, 3, 5 from Chapter 4
- Project Two →

Copyright – Dan Brandon



Project 2

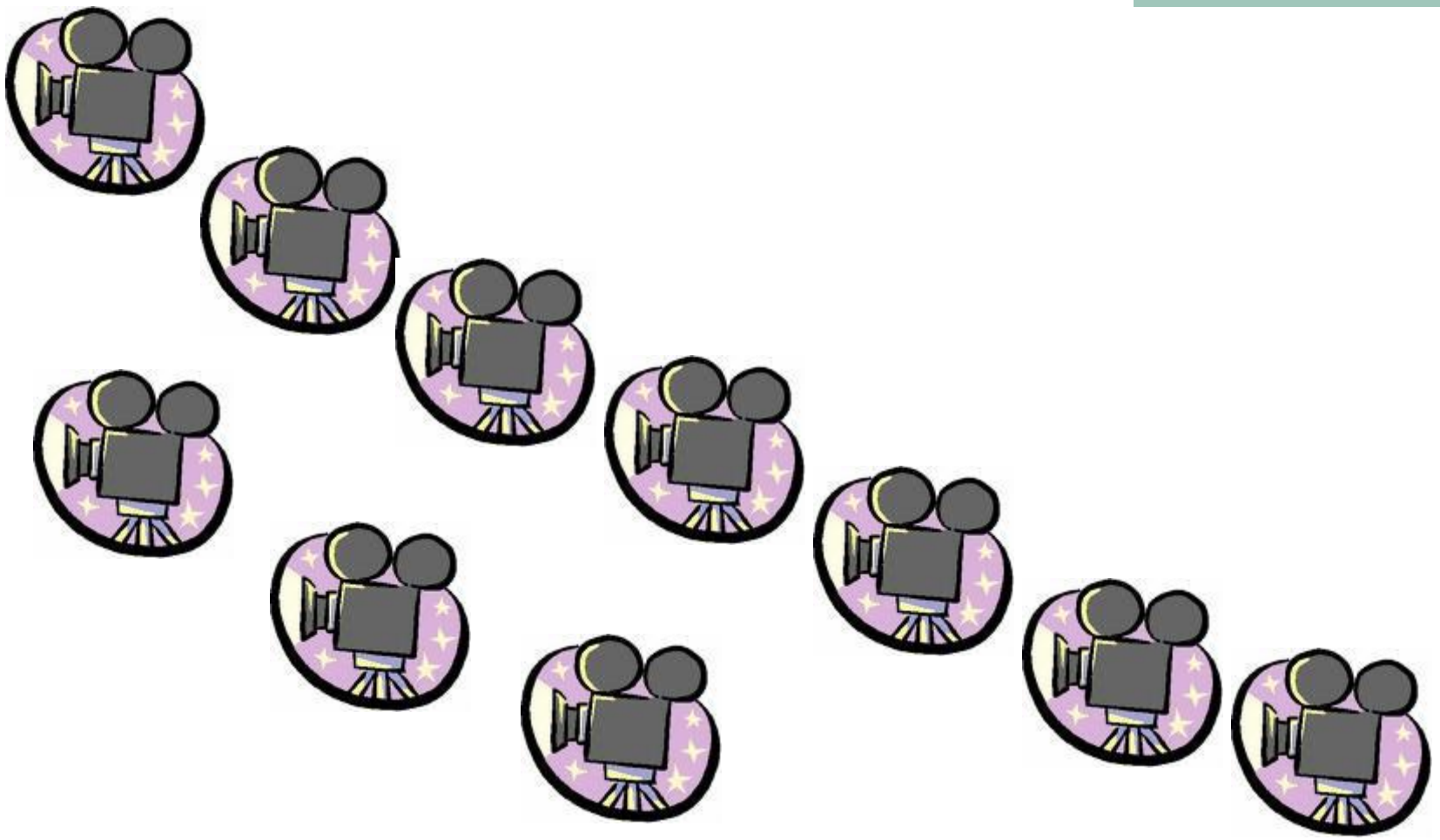
- Betty Byte lives in a 4br/2.5ba 2900 sq ft house on .6 acres
- She has obtained recent sales data on comparable houses in her subdivision (shown on the next slide)
- She is considering adding another bath to her house which would be 300 sq ft
- What is the most she should spend on adding that extra bath ?
  - Use MLR to find the relevant value formula and how much adding one bath at 300 sq ft to her house would increase the value of her house

Project (con't)

Value	Acres	Sq Ft	BRs	Baths
253000	0.5	3000	3	2
310000	0.6	3400	4	3.5
260000	0.4	3100	3	2.5
340000	0.7	3600	5	2.5
320000	0.8	3200	4	3
305000	0.55	3300	5	2.5
285000	0.65	2900	4	2.5
278000	0.45	3200	4	2.5
325000	0.75	3300	4	3
315000	0.5	3600	5	3

Regression Derivation

[Khan Academy Videos for Detailed Derivation]



J