Chapter 11 Simple Linear Regression

______________: comparing means across groups
______________: presenting relationships among numeric variables

11.1 Probabilistic Model

__________________**:** The model hypothesizes **an** __________ **relationship** between the variables.

For example, $y = 3.99x$, the cost of x boxes cereals, x: # of boxes,

$y = 35 + 0.5x$, the cost of renting a car for one day, x: # of miles driven.

(There is no allowance for error in this prediction.)

_________________**:** The model hypothesizes **a** ____________ **relationship** between the variables.

For example, x: the age, y: the height of people,
x: IQ level of students, y: the test score
x: the age of a car y: the price of a brand car

____

The simplest probabilistic model: _____________________or _________________________.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

y **:** ___________________or **response variable (**variable to be modeled**);**

x **:** ___________________or **predictor variable** (variable used as a predictor of y);

ε :_______________,

β_0 : ______________of the line, (the value of y when x=0);

β_1 : _____________of the line; (**the change of y for each unit increase of x**)

$\beta_1 > 0$, increasing function, __________________________ (upward)

$\beta_1 < 0$, decreasing function, __________________________ (downward)

In the probabilistic model, the deterministic component is referred to as the **line of means**, because the mean of y, $E(y)$, is equal to the straight-line component of the model. That is,

$$E(y) = \beta_0 + \beta_1 x$$

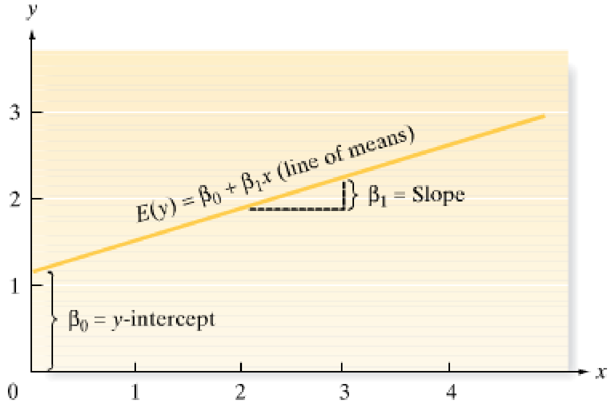
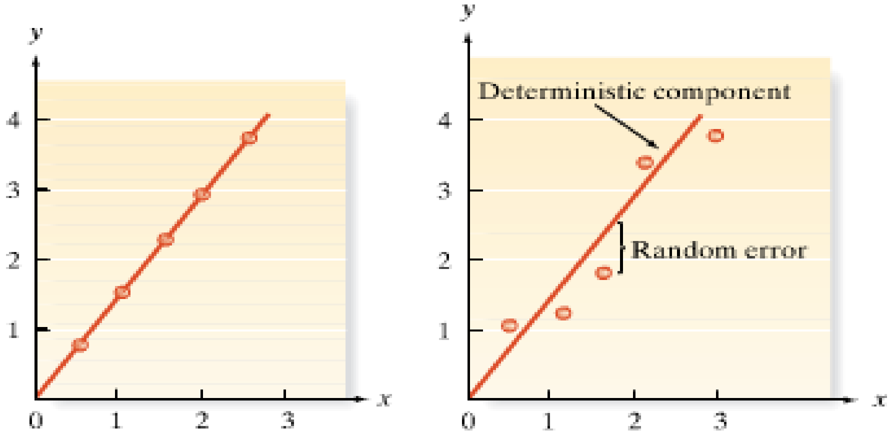


FIGURE 11.2
The straight-line model



a. Deterministic relationship: $y = 1.5x$

b. Probabilistic relationship: $y = 1.5x$ + Random error

Regression Analysis: five-step procedure:

Step1, Hypothesize the ________________________________of the model that relates the mean E(y) to the independent variable x;

Step 2, Use the sample data to estimate ___________________in the model;

Step 3, specify the probability distribution of the __________________ term and estimate the standard deviation of this distribution;

Step 4, statistically evaluate the _______________ of the model,

Step5, when satisfied that the model is useful, use it for___________, estimation and other purpose.

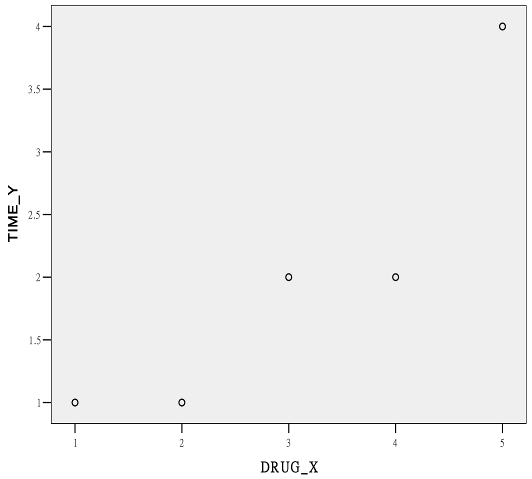
11.2 Fitting the model: the least squares approach

1. In order to determine whether a linear relationship between y and x is plausible, use a ______________________________.

Example 1: Reaction time (STIMULUS),

x: drug concentration (percentage) in bbodstream, y: reaction time (seconds)

X: drug concentration (%)	**Y: reaction time (seconds)**
1	**1**
2	**1**
3	**2**
4	**2**
5	**4**



Based on the scatter plot, It is ________________ to use the simple linear regression model to model the relationship between x and y.

We can have many fitted lines. The “best” one is the line that all the observed data are very close to this line. We use ____________________to evaluate how close the data from the line. There is one line and only one line **with the __________SSE**. We call this line __________________________,

the least squares line: ____________________________

How can we get it ($\hat{\beta}_0$, $\hat{\beta}_1$)?

Example1, Reaction time:

1. Find the least squares line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

x	y	xy	x^2
1	1		
2	1		
3	2		
4	2		
5	4		

So the least squares line is: ____________________________

2. Find the SSE. (SSE = $\sum(y-\hat{y})^2$)

x	y	$\hat{y}=-0.1+0.7x$	$y-\hat{y}$	$(y-\hat{y})^2$
1	1			
2	1			
3	2			
4	2			
5	4			

3. Give a practical interpretation of $\hat{\beta}_0$ and $\hat{\beta}_1$.

$\hat{\beta}_1 =$ _____ (estimated slope):

$\hat{\beta}_0 =$ _______(estimated intercept):

Note: the model parameters should be interpreted only within the _______________of the independent variable.

4. Predict the reaction time when x=2%. (a predicted mean value of y)

SPSS output for Example1: Reaction time:

Model Summary				
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**
1	.904(a)	.817	.756	.606
a Predictors: (Constant), DRUG_X				

ANOVA(b)						
Model		**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**

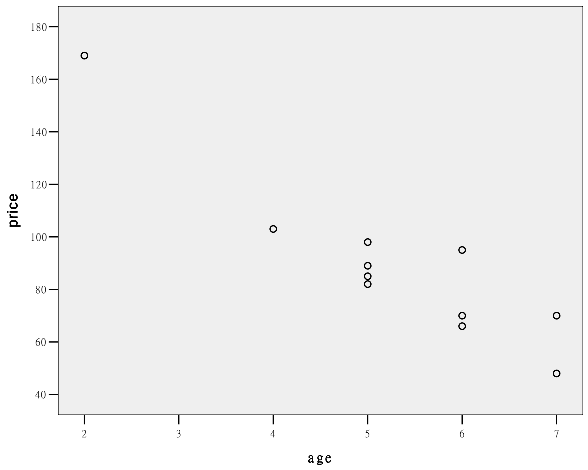
1	**Regression**	4.900	1	4.900	13.364	.035(a)
	Residual	1.100	3	.367		
	Total	6.000	4			
a Predictors: (Constant), DRUG_X						
b Dependent Variable: TIME_Y						

Coefficients(a)						
Model		**Unstandardized Coefficients**		**Standardized Coefficients**	**t**	**Sig.**
		B	**Std. Error**	**Beta**		
1	**(Constant)**	-.100	.635		-.157	.885
	DRUG_X	.700	.191	.904	3.656	.035
a Dependent Variable: TIME_Y						

Example2. Age and price of Orion car

car	Age(yr):x	Price($100):y	xy	X^2
1	5	85		
2	4	103		
3	6	70		
4	5	82		
5	5	89		
6	5	98		
7	6	66		
8	6	95		
9	2	169		
10	7	70		
11	7	48		

1. Determine the simple linear regression is reasonable to model the relationship between the price and age or not, using __________________.



2. Determine the regression equation (the least squares line) of the data.

3. Give practical interpretations to $\hat{\beta}_0$ and $\hat{\beta}_1$. (sampled range:________________)

$\hat{\beta}_0$ =_______,

$\hat{\beta}_1$ = __________,

4. Predict the price of a 3-year-old Orion car and a 4-year-old Orion car.

5. Find SSE of the analysis (from SPSS output). SSE = _______________

SPSS output for Example2: Age-price of Orion car:

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.924(a)	.853	.837	12.577

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8285.014	1	8285.014	52.380	.000(a)
	Residual	1423.532	9	158.170		
	Total	9708.545	10			

a Predictors: (Constant), age

b Dependent Variable: price

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	195.468	15.240		12.826	.000
	age	-20.261	2.800	-.924	-7.237	.000

a Dependent Variable: price

11.3 Model Assumptions

- Four basic assumptions about the probability distribution of random error ε **:**

1. ________________------ $E(y) = \beta_0 + \beta_1 x$ **;**

2. The variance of the probability distribution of ε is __________for all values of x;

3. The probability distribution of ε is_______________;

4. The values of ε associated with any two observed values of y are___________.

$\varepsilon \sim N(0, \sigma^2)$, **independent**

- An estimation of variance σ^2

When making inference (confidence interval and hypothesis test), we need an estimator of σ^2,

- **Estimation of variance σ^2 for simple linear regression model**

- **Estimation of standard deviation σ for simple linear regression model**

Example1: Reaction time, find an estimate of variance σ^2.

Example2: Age-price: compute an estimate of standard deviation σ **.**

- **Interpretation of estimated standard deviation s:**

Empirical rule: we expect ________________ of the observed y values to lie within 2S of their respective least squares predicted values $\hat{y}$.

11.4 Assessing the Utility of the model: Making inferences about the slope β_1

Under the four assumptions of ε, the sampling distribution of $\hat{\beta}_1$ ____________________,

$$\sigma_{\hat{\beta}} = \frac{\sigma}{\sqrt{SS_{xx}}},$$

Estimated standard error of the slope $\hat{\beta}_1$**:** ________________________

Example: Age-price: standard error $S_{\hat{\beta}_1} = \dfrac{S}{\sqrt{SS_{xx}}}$

- Hypothesis test of slope β_1

____________________, (variable y and x does not have significant linear relationship)
____________________, (**variable y and x have significant linear relationship**)

Test statistic:____________________ with **df =________________**

Rejection region:

Conclusion.

Example1: Reaction time:
Conduct a test of hypothesis to determine if there is a **linear** relationship between reaction time and the percentage of drug in bloodstream. Use $\alpha = 0.05$.

Example2: Age-Price:
Conduct a test of hypothesis to determine if there is a **negative linear** relationship between age and price of Orion cars. Use $\alpha = 0.05$.

- A 100(1- α)% Confidence interval for the slope β_1:

Example1: Reaction time: Form a 95% confidence interval for slope β_1 and interpret the interval.

Interpret: we are 95% confident that ____

Example2: Age-price of Orion car: Form a 95% confidence interval for slope β_1 and interpret the interval.

Interpret: we are 95% confident that ____

11.5 The coefficients of correlation and Determination

- The coefficients of correlation

r **:** measure the _______________________________between two variables x and y in the sample. It is a sample ____________ used as the _____________ of population correlation coefficient ρ .

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

Note: ____________________.

If r ______________________, it implies little or no linear relationship between y and x;

If r_______________________, the stronger linear relationship between y and x;

If___________, positive linear relationship between y and x, (y ___________ as x increase);

If___________, negative linear relationship between y and x, (y ____________as x increase).

Example1, **Reduction time:** Find the coefficient of correlation and interpret it.

.

x	**y**	**xy**	x^2	y^2
1	**1**			
2	**1**			
3	**2**			
4	**2**			
5	**4**			

So the correlation coefficient **r =** ___________, it implies _______________linear relationship between reaction time (y) and percentage of the drug in bloodstream (x).

Example2: **Age-price of Orion cars**: Find the coefficient of correlation and interpret it.

r = ____________, (SPSS output), it implies ______________linear relationship between price and age of the car.

Note: we can't infer a ____________relationship on the basis of high sample correlation. The only safe conclusions is that a linear trend may exist between x and y.

- The Coefficient of determination

The coefficient of determination r^2 represents the ____________of the total sample variability in y (dependent variable) can be attributed to the linear regression on x (independent variable).

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} =$$

Note: 1. ________________.

2. The value of coefficient of determination equals to ___________ of correlation coefficient.

Example 1: Reaction time: calculate the coefficient of determination and interpret it.

$r^2 = .817$, indicates that ________of the ______________________ in ____________can be attributed to the ____________regression on______________.

Example 2: Age-price: calculate the coefficient of determination and interpret it.

$r^2 = .853$, indicates

11.6 Using the Model for Estimation and Prediction (interval)

If the model is useful, we can use the model for estimation and prediction:

1. estimate ____.

2. predict ____.

Example1. Reaction time: We get: $\hat{y} = -0.1 + 0.7x$

1. What is 95% C.I. for the mean reaction time for all people with x=4?

2. What is 95% C.I. for the reaction time for an individual (Mr. John) if his x=4?

Example2. Age-price: We get $\hat{y} = 195.47 - 20.26x$

1. what is 95% C.I. for the mean price for all 3-year-old cars?

2. what is 95% C.I. for the price for an individual (Ms. Anna's) 3-year-old car?

- **A 100(1-α)% confidence interval for the mean value of y at $x = x_p$:**

_________________________ **df = n – 2**

- **A 100(1-α)% prediction interval for an individual new value of y at $x = x_p$**

_________________________ **df = n – 2**

Note: the prediction interval for an individual new value of y is always ___________than the corresponding confidence interval for the mean value of y.

Example 1, Reaction time:

1. Form a 95% confidence interval of the mean reaction time for the people whose drug concentration in bloodstream is 4%. Interpret the result.

2. Form a 95% prediction interval to predict the reaction time for Mr. John whose drug concentration in bloodstream is 4%. Interpret the result.

Example 2, Age-price:

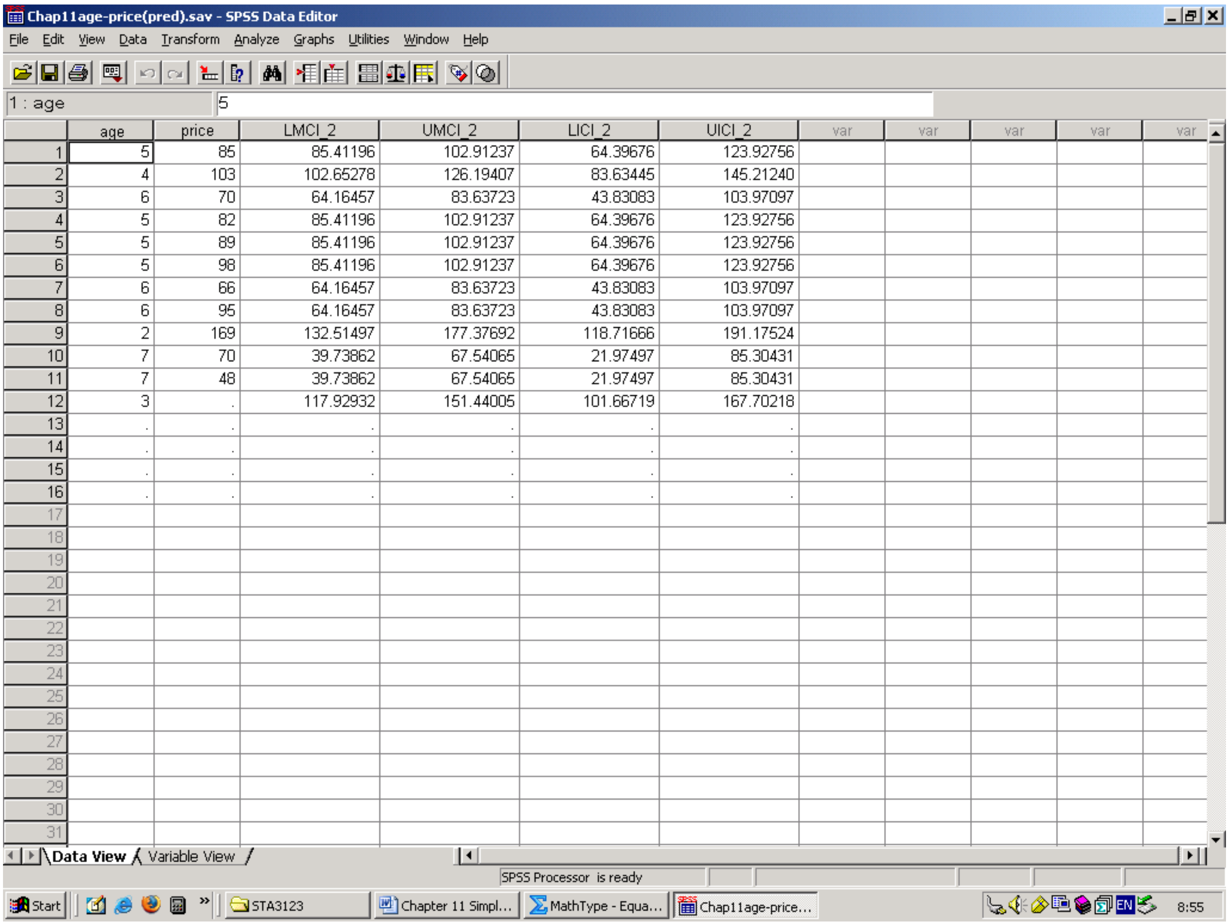
1. Form a 95% confidence interval of the mean price for 3-year-old cars. Interpret the result.

Interpretation: we are 95% confident that the ___________________ for all possible 3-year-old Orion cars is between _________________________.

2. Form a 95% prediction interval of the price for Anna's 3-year-old car. Interpret the result.

Interpretation: we are 95% confident that the ______________ for Anna's 3-year-old car is between __________________________.

Example: Age-price, SPSS output of prediction intervals.



	age	price	LMCI_2	UMCI_2	LICI_2	UICI_2
1	5	85	85.41196	102.91237	64.39676	123.92756
2	4	103	102.65278	126.19407	83.63445	145.21240
3	6	70	64.16457	83.63723	43.83083	103.97097
4	5	82	85.41196	102.91237	64.39676	123.92756
5	5	89	85.41196	102.91237	64.39676	123.92756
6	5	98	85.41196	102.91237	64.39676	123.92756
7	6	66	64.16457	83.63723	43.83083	103.97097
8	6	95	64.16457	83.63723	43.83083	103.97097
9	2	169	132.51497	177.37692	118.71666	191.17524
10	7	70	39.73862	67.54065	21.97497	85.30431
11	7	48	39.73862	67.54065	21.97497	85.30431
12	3	.	117.92932	151.44005	101.66719	167.70218
13	.	.	.	.	.	.
14	.	.	.	.	.	.
15	.	.	.	.	.	.
16	.	.	.	.	.	.

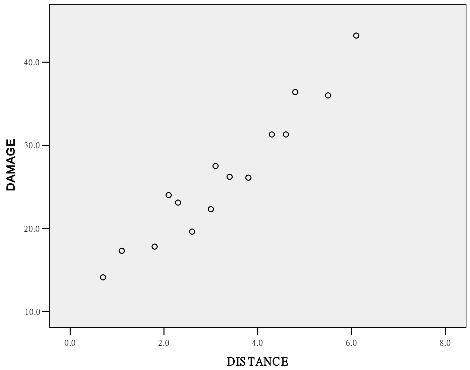
11.7 A complete example for simple linear regression model

Example: FIREDAM,

x (miles): the distance of a fire from the nearest fire station; y (thousand dollars): fire damage

x	3.4	1.8	4.6	2.3	3.1	5.5	0.7	3.0	2.6	4.3	2.1	1.1	6.1	4.8	3.8
y	26.2	17.8	31.3	23.1	27.5	36.0	14.1	22.3	19.6	31.3	24.0	17.3	43.2	36.4	26.1

Step1. Hypothesize a simple linear model to relate fire damage(y) to the distance a fire from the nearest fire station (x). (________________) $y = \beta_0 + \beta_1 x + \varepsilon$



Step2. Get the data to estimate the unknown parameters

From the SPSS output, $\hat{\beta}_0$ =___________, $\hat{\beta}_1$ =____________.

The least squares line is: ______________________________

Estimated slope, $\hat{\beta}_1 = 4.919$, it implies that the estimated mean damage will _______by ________ for each additional mile a fire from the fire station.

Estimated intercept, $\hat{\beta}_0 = 10.278$, Since x = 0 out of the sampled range (__________), no practical interpretation for $\hat{\beta}_0$.

Step3. Specify the probability distribution of the random error ε. Although the four assumptions are not completely satisfied (they rarely are for practical problems), we are willing to assume they are approx. satisfied for this example.

The estimate of the standard deviation σ of ε,

____.

Step4. Check the usefulness of the hypothesized (SLR) model.

- **making inference for** β_1

1. Hypothesis test for β_1

Do the data provide evidence that the distance and the damage have **positive linear** relationship? Use $\alpha = 0.05$.

2. 95% confidence interval for β_1**:**

$$\hat{\beta}_1 \pm t_{\alpha/2} S_{\hat{\beta}} = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

Interpret:

____.

- **the coefficient of determination** r^2 **=**________,

- **the coefficient of correlation** r **=**_______,

Overall, based on the inference on slope β_1, the values of r^2 and r, all signs show a ________________relationship between fire damage (y) and the distance (x), the SLR model is __________.

Step5, Use the model. Suppose the insurance company wants to predict the fire damage if a major residential fire was to occur 3.5 miles from the nearest fire station.

$$\hat{y} \pm t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

From SPSS output, a 95% prediction interval for x = 3.5: (____________________).

We are 95% confident that the ________________________ in a major residential fire 3.5 miles from the nearest station is between ____________ and ______________.

SPSS output for FIREDAM:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.961[a]	.923	.918	2.3163

a. Predictors: (Constant), DISTANCE

b. Dependent Variable: DAMAGE

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	841.766	1	841.766	156.886	.000[a]
	Residual	69.751	13	5.365		
	Total	911.517	14			

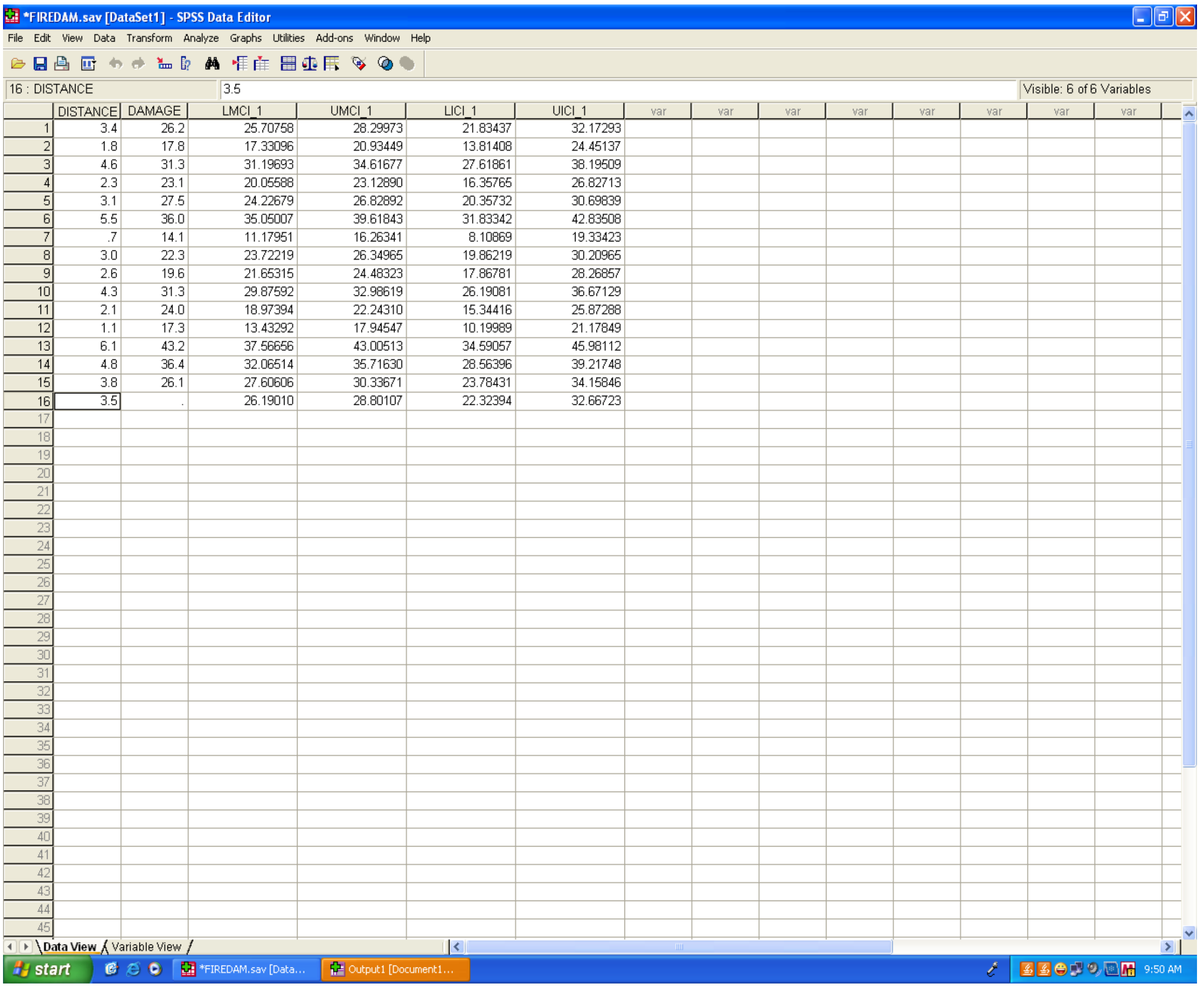
a. Predictors: (Constant), DISTANCE

b. Dependent Variable: DAMAGE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.278	1.420		7.237	.000
	DISTANCE	4.919	.393	.961	12.525	.000

a. Dependent Variable: DAMAGE


*FIREDAM.sav [DataSet1] - SPSS Data Editor

16 : DISTANCE 3.5

	DISTANCE	DAMAGE	LMCI_1	UMCI_1	LICI_1	UICI_1
1	3.4	26.2	25.70758	28.29973	21.83437	32.17293
2	1.8	17.8	17.33096	20.93449	13.81408	24.45137
3	4.6	31.3	31.19693	34.61677	27.61861	38.19509
4	2.3	23.1	20.05588	23.12890	16.35765	26.82713
5	3.1	27.5	24.22679	26.82892	20.35732	30.69839
6	5.5	36.0	35.05007	39.61843	31.83342	42.83508
7	.7	14.1	11.17951	16.26341	8.10869	19.33423
8	3.0	22.3	23.72219	26.34965	19.86219	30.20965
9	2.6	19.6	21.65315	24.48323	17.86781	28.26857
10	4.3	31.3	29.87592	32.98619	26.19081	36.67129
11	2.1	24.0	18.97394	22.24310	15.34416	25.87288
12	1.1	17.3	13.43292	17.94547	10.19989	21.17849
13	6.1	43.2	37.56656	43.00513	34.59057	45.98112
14	4.8	36.4	32.06514	35.71630	28.56396	39.21748
15	3.8	26.1	27.60606	30.33671	23.78431	34.15846
16	3.5	.	26.19010	28.80107	22.32394	32.66723