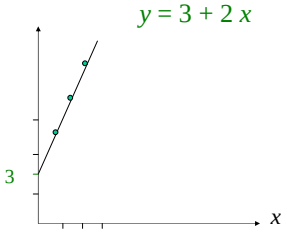
Examine Relation Between Two Quantitative Variables

1

Linear Relation

If the cost to get to the drug store is \$3.00, and each bottle of drug is \$2.00, then

$y = 3 + 2x$



2

Is there relation between "**number of handguns registered**" in the area and "**number of people killed by guns**"?

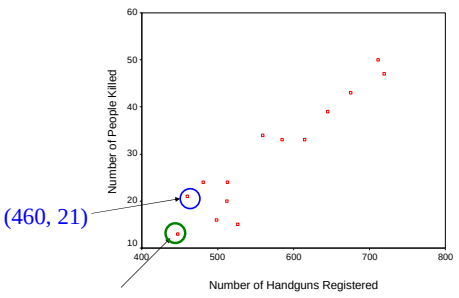
number of people killed by guns => Response variable

number of handguns registered => Explanatory variable

Year	NGR(x_i)	Nkill(y_i)
77	447	13
78	460	21
79	481	24
80	498	16
81	513	24
82	512	20
83	526	15
84	559	34
85	585	33
86	614	33
87	645	39
88	675	43
89	711	50
90	719	47

3

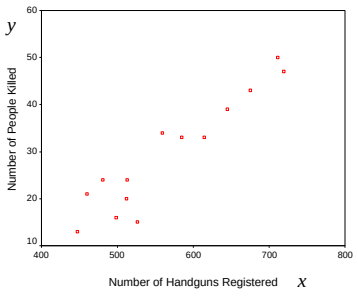
Scatter Plot



Year	NGR(x_i)	Nkill(y_i)
77	447	13
78	460	21
79	481	24
80	498	16
81	513	24
82	512	20
83	526	15
84	559	34
85	585	33
86	614	33
87	645	39
88	675	43
89	711	50
90	719	47

4

Scatter Plot



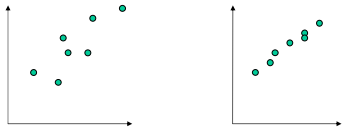
5

Information in Scatter Plot

- Form (Straight line or curve relation)
- Direction (Positive or negative relation)
- Strength (Strong or weak relation)
- Outliers

6

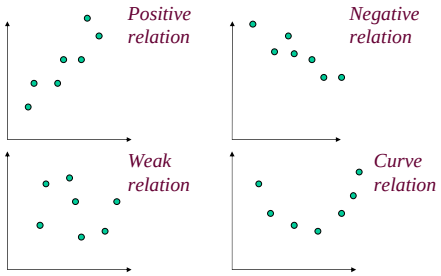
Correlation



Stronger linear trend

7

Correlation Coefficient



8

Pearson's Sample Correlation

$$r = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \cdot \left(\frac{y_i - \bar{y}}{s_y} \right)$$

$$= \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(z_{x_i} \cdot z_{y_i} \right)$$

s_x: Sample standard deviation of x's

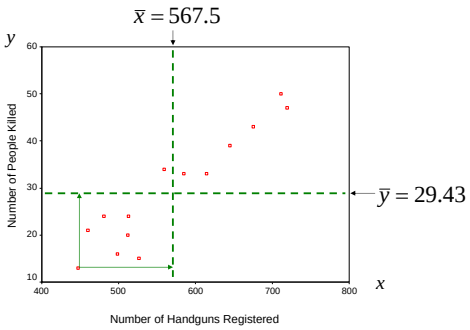
s_y: Sample standard deviation of y's

z_x: Sample standard score of x_i

z_y: Sample standard score of y_i

9

Scatter Plot



10

	x	y	$\frac{x-\bar{x}}{s_x}$	$\frac{y-\bar{y}}{s_y}$	$\left(\frac{x-\bar{x}}{s_x}\right) \cdot \left(\frac{y-\bar{y}}{s_y}\right)$
77	447	13	-1.31	-1.35	1.77
78	460	21	-1.17	-0.69	0.81
79	481	24	-0.94	-0.45	0.42
80	498	16	-0.76	-1.10	0.83
81	513	24	-0.59	-0.45	0.26
82	512	20	-0.60	-0.77	0.47
83	526	15	-0.45	-1.18	0.53
84	559	34	-0.09	0.38	-0.03
85	585	33	0.19	0.29	0.06
86	614	33	0.51	0.29	0.15
87	645	39	0.84	0.79	0.66
88	675	43	1.17	1.11	1.30
89	711	50	1.56	1.69	2.64
90	719	47	1.65	1.44	2.38
				Total =	12.24
s. d. =	s_x = 91.91	s_y = 12.19		r =	0.941477289
Mean =	$\bar{x}$ = 567.50	$\bar{y}$ = 29.43			

11

$$\frac{447 - 567.5}{91.91} = -1.31$$

$$\frac{13 - 29.43}{12.19} = -1.35$$

12

Shortcut Formula

$$r = \frac{1}{n-1}\cdot\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

$$= \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}}$$

$$s_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

$$s_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n},\ s_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

13

Interpretation of *r*

- $-1 < r < 1$
- It measures the **strength** and **direction** of the **linear relation** between two quantitative variables.
- $r = 1$ if all points lie exactly on a straight line.
- ρ is he notation for population correlation coefficient.

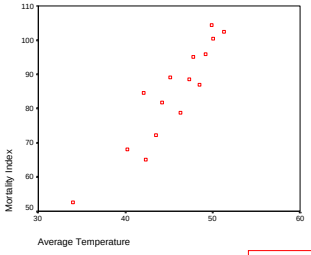
14

Correlation Does Not Imply Causation

The number of handguns registered may not be the direct cause for the number of people killed by guns.

15

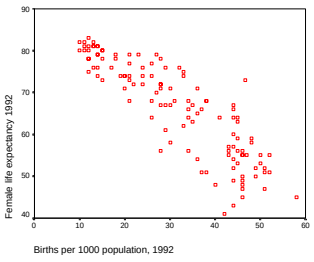
Example: The relation between the **average annual temperature** and **the mortality index** for a type of breast cancer in women in certain region of Europe.



$r = 0.924$

16

Example: In an investigation, 122 countries were included to study the relation between **female's life expectancy** and the **birthrate**.



17

Test of Linear Correlation

H_o: There is **NO** statistical correlation between Female Life Expectancy & Birth rates

H_a: There is statistical correlation between Female Life Expectancy & Birth rates

Correlations

		Female life expectancy 1992	Births per 1000 population, 1992
Female life expectancy 1992	Pearson Correlation	1	-.870**
	Sig. (2-tailed)	.	.000
	N	122	121
Births per 1000 population, 1992	Pearson Correlation	-.870**	1
	Sig. (2-tailed)	.000	.
	N	121	121

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

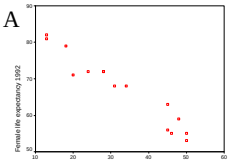
Correlation coefficient

p-value

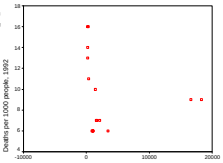
18

$r_1 = -0.968$ $r_2 = 0.982$ $r_3 = -0.165$ $r_4 = 0.8$

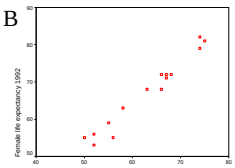
A



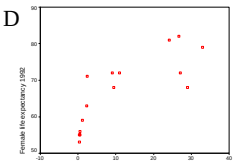
C



B



D



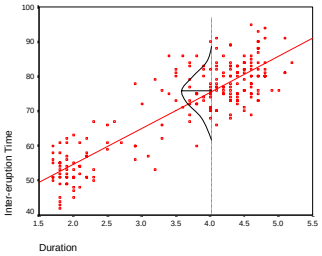
$n = 15$

19

How long should you wait till next eruption?

20

Duration and Inter-eruption Time



Inter-eruption time: Response variable (Outcome, Dependent)

Duration of eruption: Explanatory variable (Predictor, Independent)

21

Is there relation between "**number of handguns registered**" in the area and "**number of people killed by guns**"?

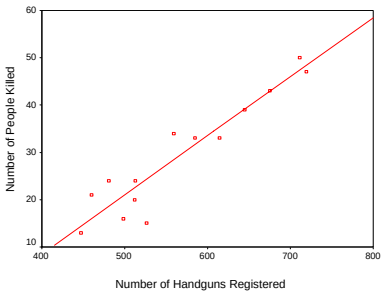
number of people killed by guns
=> Response variable
(Outcome, Dependent)

number of handguns registered
=> Explanatory variable
(Predictor, Independent)

Year	NGR(x_i)	Nkill(y_i)
77	447	13
78	460	21
79	481	24
80	498	16
81	513	24
82	512	20
83	526	15
84	559	34
85	585	33
86	614	33
87	645	39
88	675	43
89	711	50
90	719	47

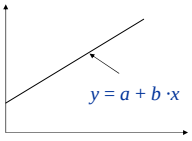
22

Graph with a Fitted Line



23

Equation of a Straight Line



$y = a + b \cdot x$

a is the slope

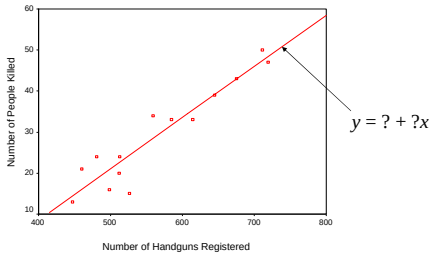
b is the y-intercept

y response or dependent variable

x explanatory, dependent, or predictor variable

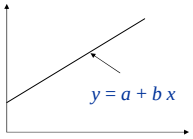
24

Graph with a Fitted Line



25

The Least Square Regression Line



The formula for the y-intercept, *a*, and slope, *b*, of the lease-squares regression line is:

$$b = r \cdot \frac{s_y}{s_x}, \qquad a = \bar{y} - b \cdot \bar{x}$$

26

	x	y	$\frac{x-\bar{x}}{s_x}$	$\frac{y-\bar{y}}{s_y}$	$\left(\frac{x-\bar{x}}{s_x}\right)\cdot\left(\frac{y-\bar{y}}{s_y}\right)$
77	447	13	-1.31	-1.35	1.77
78	460	21	-1.17	-0.69	0.81
79	481	24	-0.94	-0.45	0.42
80	498	16	-0.76	-1.10	0.83
81	513	24	-0.59	-0.45	0.26
82	512	20	-0.60	-0.77	0.47
83	526	15	-0.45	-1.18	0.53
84	559	34	-0.09	0.38	-0.03
85	585	33	0.19	0.29	0.06
86	614	33	0.51	0.29	0.15
87	645	39	0.84	0.79	0.66
88	675	43	1.17	1.11	1.30
89	711	50	1.56	1.69	2.64
90	719	47	1.65	1.44	2.38
				Total =	12.24
s. d. =	s_x = 91.91	s_y = 12.19		r =	0.941477289
Mean =	$\bar{x}$ = 567.50	$\bar{y}$ = 29.43			

27

Handgun Example

$$b = .941 \cdot \frac{12.19}{91.91} = .124862, \qquad s_y = 12.19, \quad s_x = 91.91$$

$$a = 29.43 - .124862 \times 567.50 = -41.430439$$

The regression (prediction) equation:

$$\hat{y} = a + b \cdot x$$

$$\hat{y} = -41.430369 + .124862 \cdot x$$

28

An Estimation

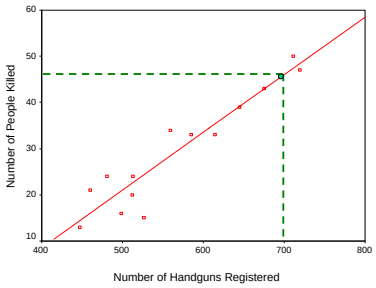
If at a certain year the number of handguns registered is 700,000, estimate how many people on average would be killed by guns.

$$\begin{aligned} \hat{y} &= -41.430439 + .124862 \cdot x \\ &= -41.430439 + .124862 \cdot 700 \\ &= 45.973 \end{aligned}$$

The average response at x = 700 is 45.973.

29

Graph with a Fitted Line



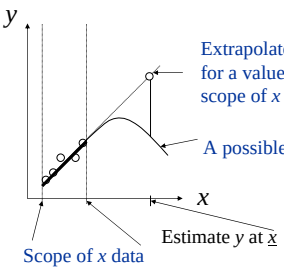
30

Cautions

- Avoid unsure extrapolation.
- Causality?

31

Problem of extrapolation



32

Regression and Causality

Regression itself provides no information about causal patterns and must supplemented by additional analysis (with designed and controlled experiments) to obtain insight about causal relationship.

33

Evaluation of the Model

Coefficient of Determination, r^2 : It is the proportion of variation in observed ***y*** that can be explained by the variable ***x*** with the linear regression model.

34

88.6% of the variability in *y* can be explained by *x*.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.941[a]	.886	.877	4.28

a. Predictors: (Constant), NPOWERBT

Coefficient of determination

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1711.979	1	1711.979	93.615	.000[a]
	Residual	219.450	12	18.287		
	Total	1931.429	13			

a. Predictors: (Constant), NPOWERBT

b. Dependent Variable: MANKILL

Mean Square Error (MSE) = $s^2_{y|x}$

35

SPSS Output for Linear Regression

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-41.430	7.412		-5.589	.000
	Number of Handguns Registered	.125	.013	.941	9.675	.000

a. Dependent Variable: Number of People Killed

Regression coefficients: $a = -41.430$; $b = .125$

Statistical tests for $a = 0$ and $b = 0$ Sig. Values < 0.05 implies that a and b are both not zero.

Equation of the regression line:

$\hat{y} = a + b \cdot x;$ $\hat{y} = -41.430 + .125 \cdot x$

36