PRESENTATION

ON

REGRESSION ANALYSIS

INTRODUCTION TO REGRESSION ANALYSIS

- **Regression analysis** is the most often applied technique of statistical analysis and modeling.
- If two variables are involved, the variable that is the basis of the estimation, is conventionally called the **independent variable** and the variable whose value is to be estimated+ is called the **dependent variable.**
- In general, it is used to model a response variable (Y) as a function of one or more driver variables (X_1, X_2, …, X_p).
- The functional form used is:

  $Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon$
- The dependent variable is variously known as **explained variables**, **predictand**, **response** and **endogenous variables**.
- While the independent variable is known as **explanatory**, **regressor** and **exogenous variable**.

DEFINITION

The Regression Analysis is a technique of studying the dependence of one variable (called dependant variable), on one or more variables (called explanatory variable), with a view to estimate or predict the average value of the dependent ariables in terms of the known or fixed values of the independent variables.

THE REGRESSION TECHNIQUE IS PRIMARILY USED TO :

- Estimate the relationship that exists, on the average, between the dependent variable and the explanatory variable
- Determine the effect of each of the explanatory variables on the dependent variable, controlling the effects of all other explanatory variables
- Predict the value of dependent variable for a given value of the explanatory variable

HISTORY

The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as .egression towards the mean[1] For Galton, regression had only this biological meaning, but his work was later extended by Uday Ule and Karl Pearson to a more general statistical context.I n the work of Yule and Pearson, the joint distribution of the response and explanatory variables is assumed to be Gaussian. This assumption was weakened by R.A Fisher in his works of 1922 and 1925 Fisher assumed that the conditional distribution of the response variable is Gaussian, but the joint distribution need not be. In this respect, Fisher's assumption is closer to Gauss's formulation of 1821.

ASSUMPTIONS OF THE LINEAR REGRESSION MODEL

1. Linear Functional form
2. Fixed independent variables
3. Independent observations
4. Representative sample and proper specification of the model (no omitted variables)
5. Normality of the residuals or errors
6. Equality of variance of the errors (homogeneity of residual variance)
7. No multicollinearity
8. No autocorrelation of the errors
9. No outlier distortion

Derivation of the Intercept

$$y = a + bx + e$$

$$e = y - a - bx$$

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a_i - b\sum_{i=1}^{n} x_i$$

Because by definition $\sum_{i=1}^{n} e_i = 0$

$$0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a_i - b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i$$

$$na = \sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i$$

$$a = \bar{y} - b\bar{x}$$

The term ε in the model is referred to as a "random error term" and may reflect a number of things including the general idea that knowledge of the driver variables will not ordinarily lead to perfect reconstruction of the response.

- If there is only one driver variable, X, then we usually speak of "simple" linear regression analysis.
- When the model involves
  - (a) multiple driver variables,
  - (b) a driver variable in multiple forms, or
  - (c) a mixture of these, the we speak of "multiple linear regression analysis".
- The "linear" portion of the terminology refers to the response variable being expressed as a "linear combination" of the driver variables.

EXAMPLE

An agronomist may be interested in studying the dependence of paddy on temperature, rainfall, amoont of fertilizer and soil fertility. such a dependency analysis maybe enable the forecasting of the average yield, given information about the explanatory variables

In regression analysis, the data used to describe the relationship between the variables are primarily measured on interval scale. the chief advantage of using the interval level of measurement is that, with such data it is possible to describe the relationship between variables more exactly employing mathematical equation. This in turn allows more accurate prediction of one variable from the knowledge of the other variables, which is one of the most important objectives of regression analysis.

It is important to note that if the **relationship** between X and Y is **curvilinear** , the **regression line** will be a **curved line** rather than **straight line**. The **greater** the **strength of relationships** between X and Y the **better** is the **prediction**.

The problem is presented to the mathematician as follows: "The values of a and b in the linear model $Y'_i = a + b X_i$ are to be found which minimize the algebraic expression ."
The mathematician begins as follows:

$\sum(Y_i - Y'_i)^2$ is the expression to be minimized

$\sum(Y_i - (a + bX_i))^2$ substituting $a + bX$ for Y'

$\sum(Y_i - a - bX_i)^2$ deleting the innermost parentheses

$\sum(Y^2 + a^2 + b^2X^2 - 2aY - 2bXY + 2abX)$ squaring the expression

$\sum Y^2 + \sum a^2 + \sum b^2X^2 - 2\sum aY - 2\sum bXY + 2\sum abX$ taking the summation sign inside

The result becomes:

$$b = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2}$$

Using a similar procedure to find the value of a yields:

$$a = \overline{Y} - b\overline{X}$$

	Y_i	X_i^2	X_iY_i	
	13	23	169	299
	20	18	400	360
	10	35	100	350
	33	10	1089	330
	15	27	225	405
SUM	91	113	1983	1744

$$N = 5$$
$$\sum x = 91$$
$$\sum y = 113$$
$$\sum x^2 = 1983$$
$$\sum xy = 1744$$

$$b = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2}$$

$$b = \frac{5 * 1744 - (91 * 113)}{5 * 1983 - 91^2}$$

$$b = \frac{8720 - 10283}{9915 - 8281}$$

$$b = \frac{-1563}{1634} = -.9565$$

$$\overline{X} = 18.2$$
$$\overline{Y} = 22.6$$
$$b = -.957$$
$$a = \overline{Y} - b\overline{X}$$
$$a = 22.6 - (-.957 * 18.2)$$
$$a = 40.01$$

THE REGRESSION MODEL

The situation using the regression model is analogous to that of the interviewers, except instead of using interviewers, predictions are made by performing a linear transformation of the predictor variable. Rather than interviewers in the above example, the predicted value would be obtained by a linear transformation of the score. The prediction takes the form

$$Y' = a + bX$$

where a and b are parameters in the regression model.

Example Uses of Regression Models

Pregnancy

A woman in the first trimester of pregnancy has a great deal of concern about the environmental factors surrounding her pregnancy and asks her doctor about what to impact they might have on her unborn child. The doctor makes a "point estimate" based on a regression model that the child will have an IQ of 75. It is highly unlikely that her child will have an IQ of exactly 75, as there is always error in the regression procedure. Error may be incorporated into the information given the woman in the form of an "interval estimate." For example, it would make a great deal of difference if the doctor were to say that the child had a ninety-five percent chance of having an IQ between 70 and 80 in contrast to a ninety-five percent chance of an IQ between 50 and 100. The concept of error in prediction will become an important part of the discussion of regression models.
It is also worth pointing out that regression models do not make decisions for people. Regression models are a source of information about the world. In order to use them wisely, it is important to understand how they work.

TYPES OF REGRESSION ANALYSIS:

Regression analysis is generally classified into two kinds: simple and multiple. Simple
regression involves only two variables, one of which is dependent variable and the other
Is explanatory(independent) variable. The associated model in the case of simple
regression will be a simple regression model.

- A regression analysis may involve a linear model or a nonlinear model.

The term linear can be interpreted in two different ways:

1. Linear in variable
2. Linearity in the parameter

Regression Analysis: Model Assumptions

- Model assumptions are stated in terms of the random errors, ε, as follows:
  - the errors are normally distributed,
  - with mean = zero, and
  - constant variance σ^2_ε, that does not depend on the settings of the driver variables, and
  - the errors are independent of one another.
- This is often summarized symbolically as: ε is NID(0, σ^2_ε)

LINEAR REGRESSION

In linear regression, the model specification is that the dependent variable, y_i is a linear combination of the *parameters* (but need not be linear in the *independent variables*). For example, in simple linear regression for modeling *n* data points there is one independent variable: x_i, and two parameters, β_0 and β_1:

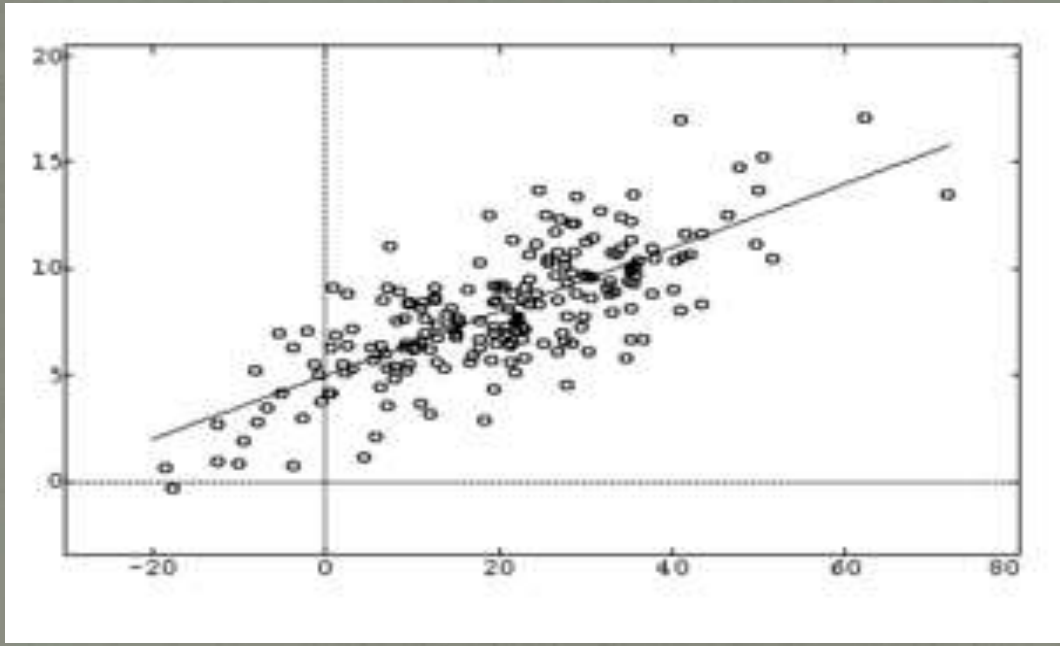


Fig: Illustration of linear regression on a data set

In the case of simple regression, the formulas for the least squares estimates are

$$Y = a + bX$$

$$b = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} \qquad a = \frac{\sum Y - b\sum X}{N}$$

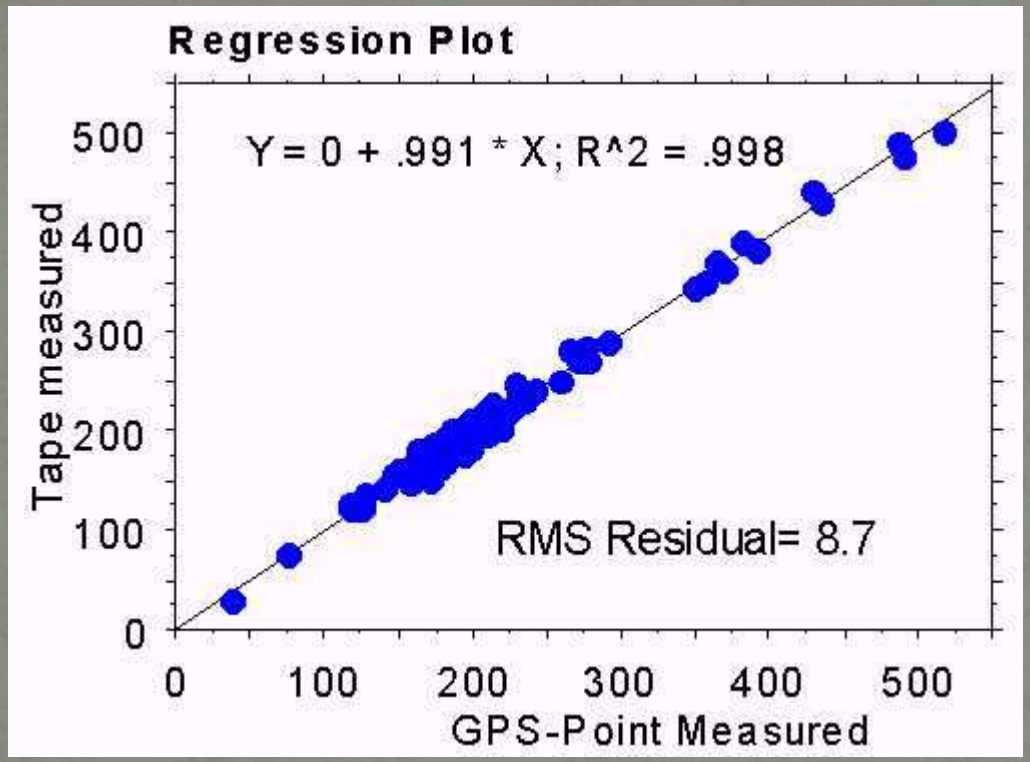
Where,

N = number of observations, or years

X = a year index (decade)

Y = population size for given census years

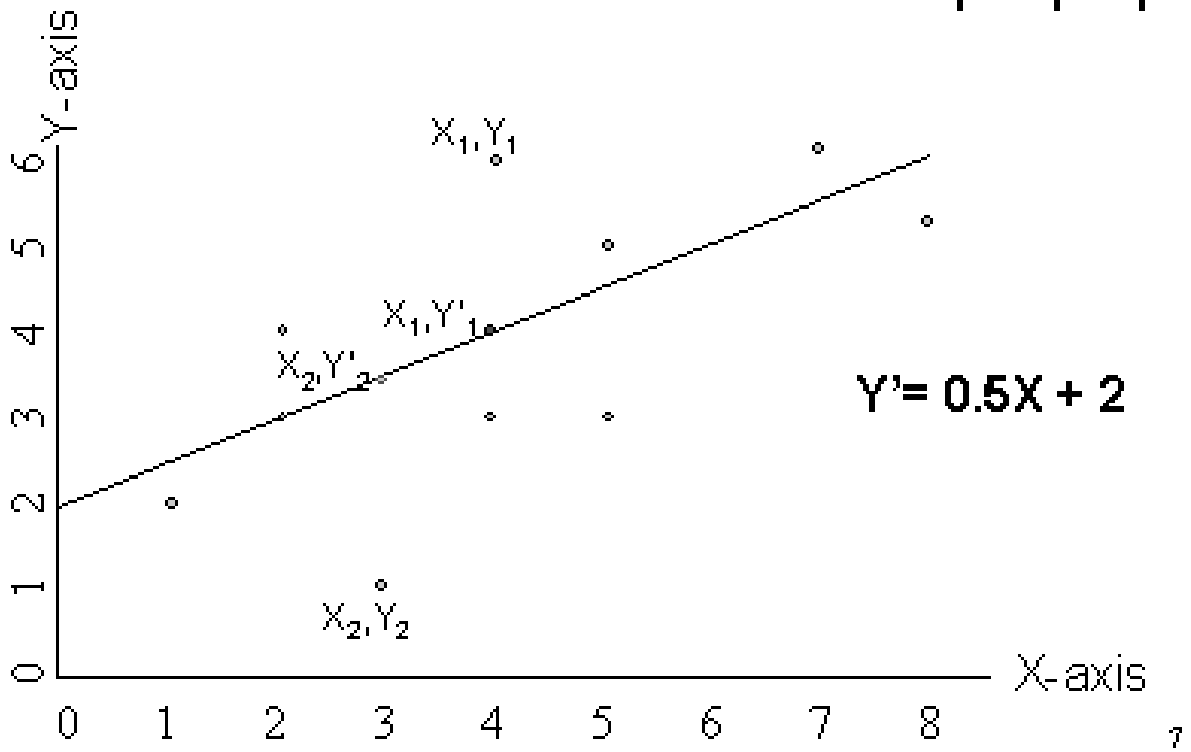
GRAPH:


Regression Plot
Y = 0 + .991 * X; R^2 = .998
RMS Residual= 8.7
Tape measured
GPS-Point Measured
0
100
200
300
400
500

Scatter plot

- regression

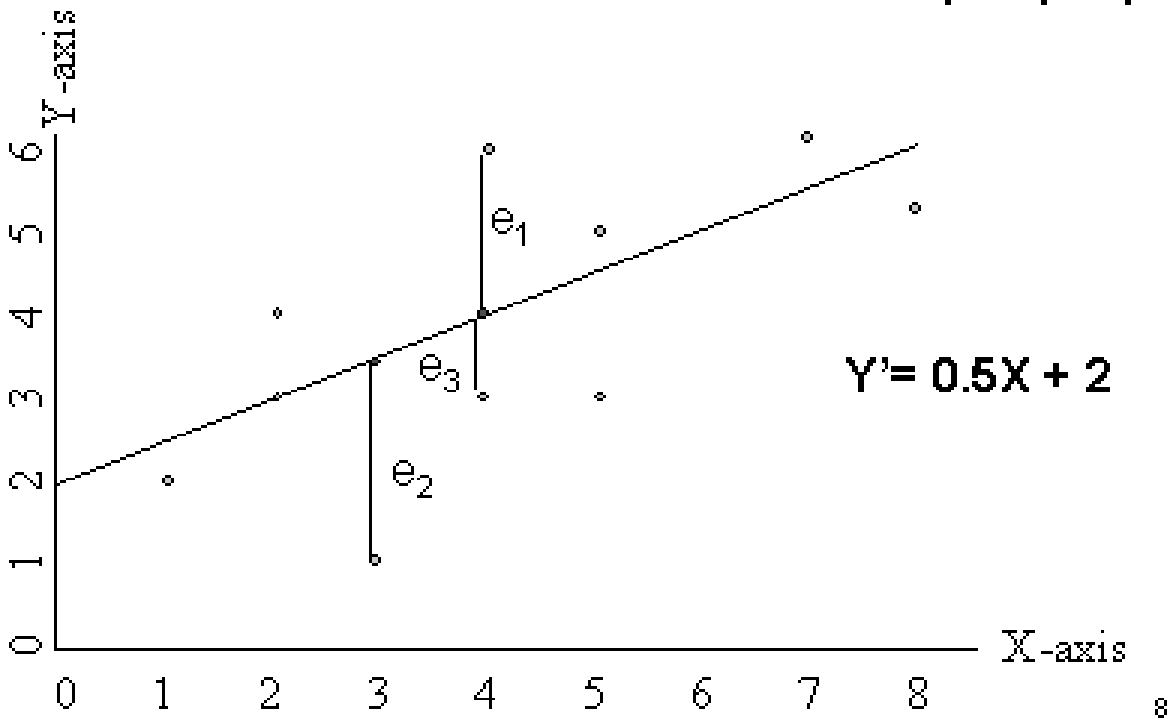
$e_i = Y_i - Y'_i$



Scatter plot

- **Error of estimate**: the distance in the Y direction between the predicted and actual Yscores

$e_i = Y_i - Y'_i$



Linear relations

Example

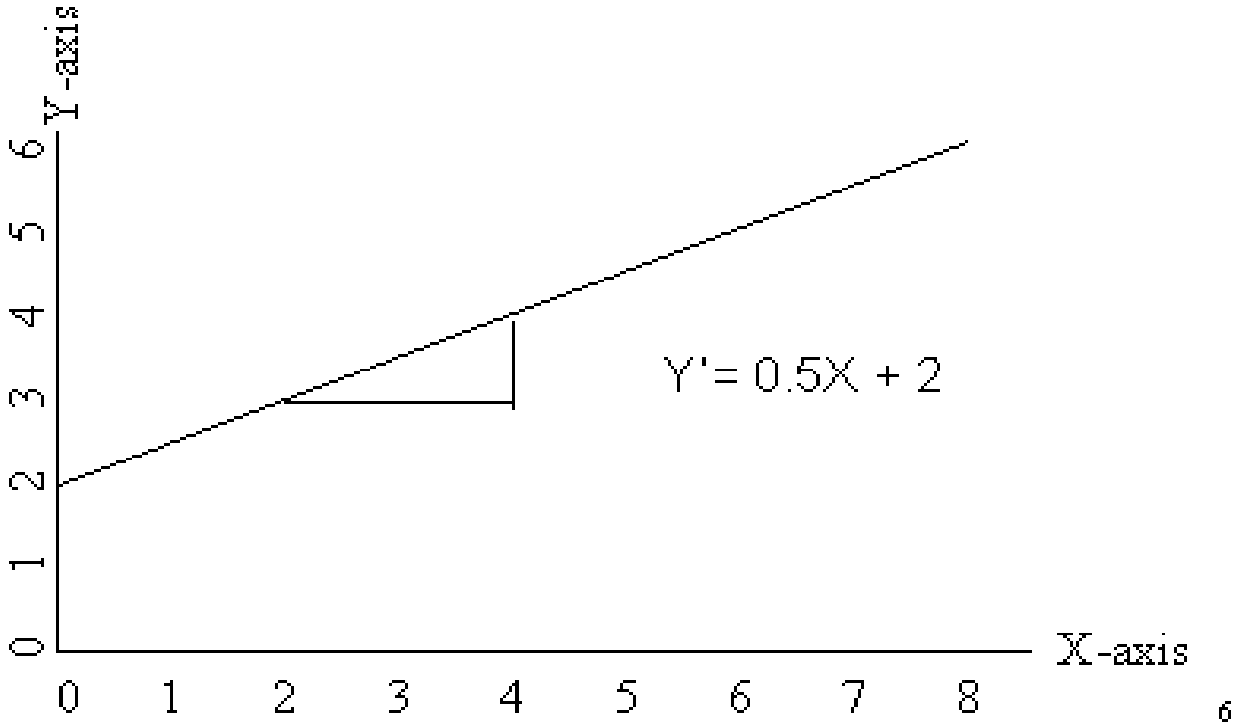
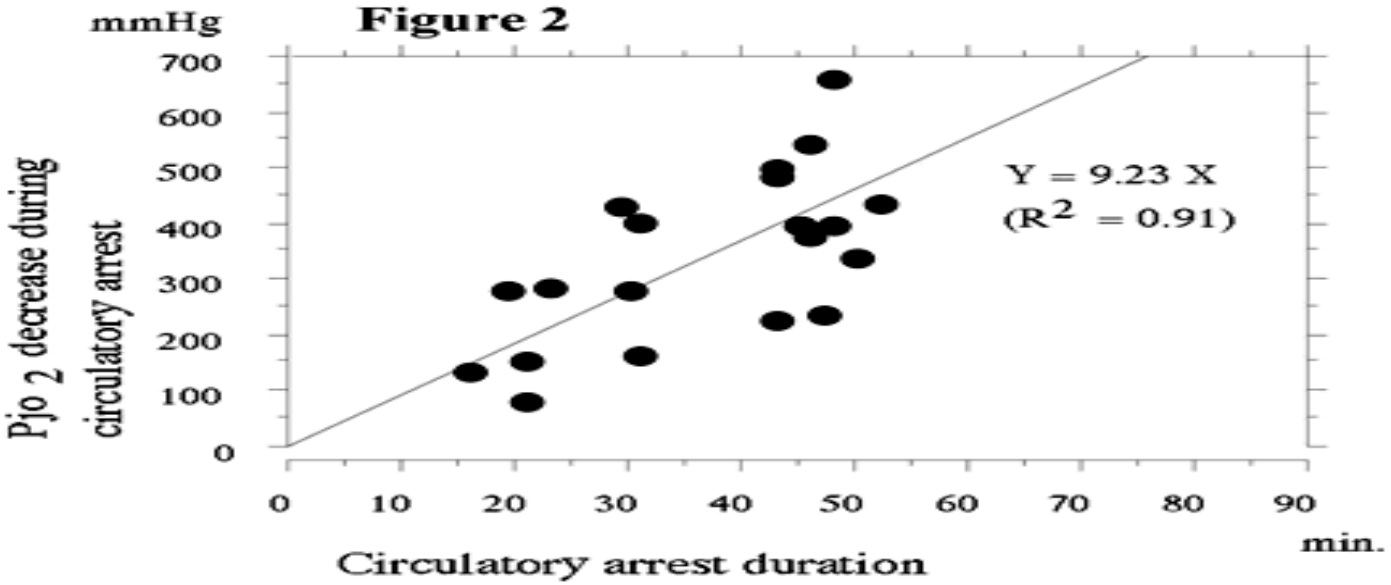
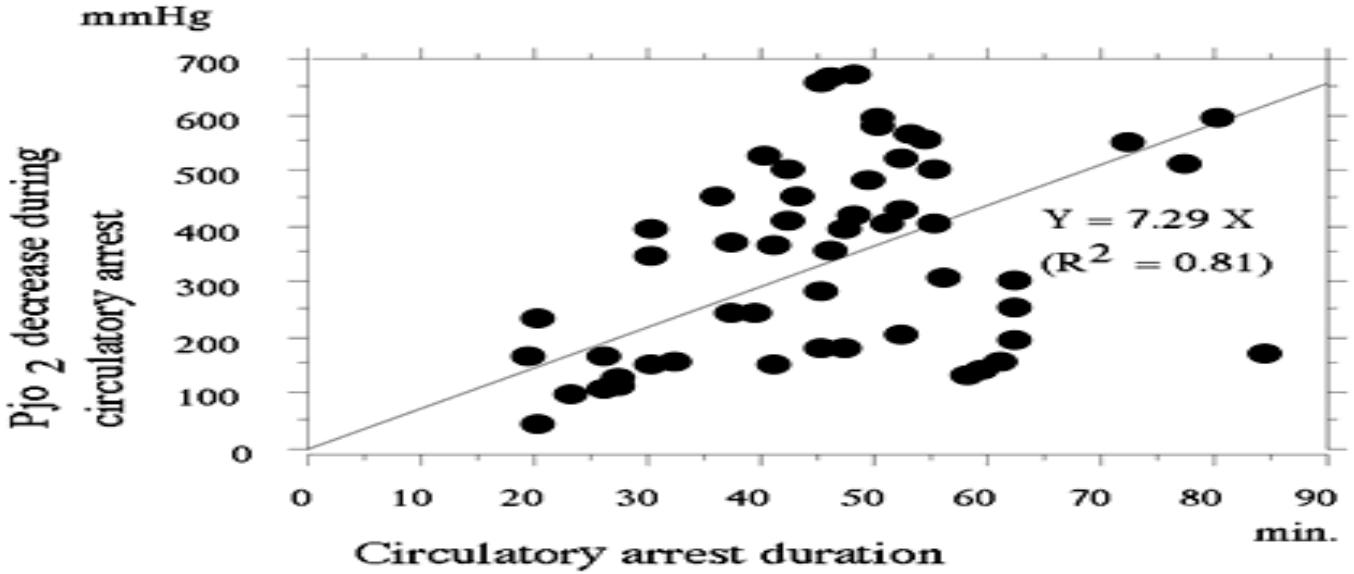


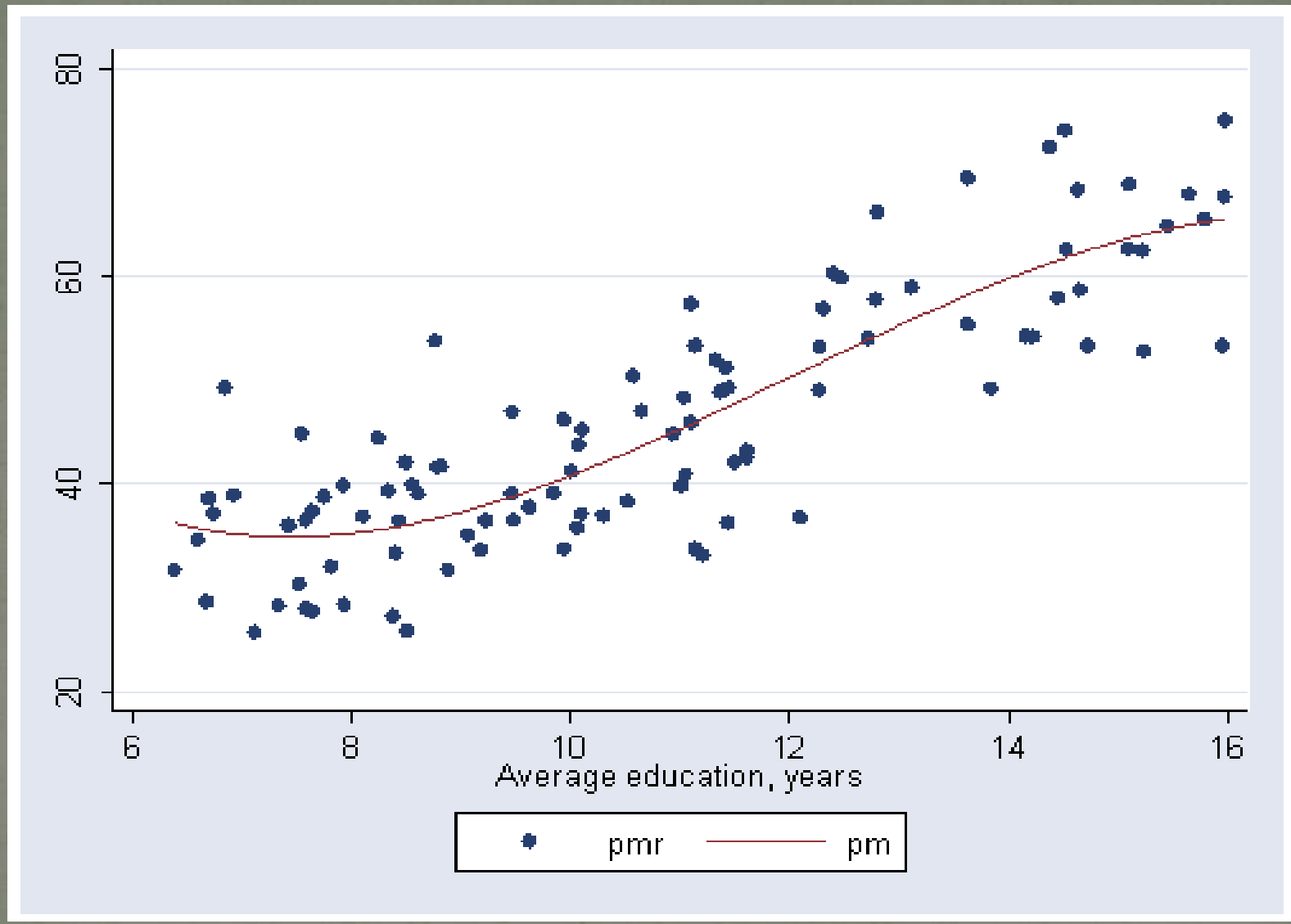
Figure 2





The correlation between the duration of circulatory arrest and the decrease in Pjo2 during circulatory arrest.

This shows a tendencey toward decreasing slope of the correlation line with increasing thiopental dose.


80
60
40
20
6
8
10
12
14
16
Average education, years
pmr
pm

Thank You....