

Regression Analysis

Prepared by Dr. Mohamed Sleem

Subject code : FNT 631
Data Analysis

Lecture 6



Regression Analysis

Lecture 6

Subject code : FNT 631
Data Analysis

Prepared by
Dr. Mohamed Sleem

www.midocean.ae

Answer Review Questions of Chapter 4

Discuss the different types of sampling methods mentioned in the document. Provide examples of situations where each type of sampling method would be most appropriate and explain the reasons for their suitability.

- ❑ **Simple Random Sampling:** Each individual or item in the population frame is equally likely to be selected, ensuring unbiased representation, like drawing names from a hat.
- ❑ **Systematic Sampling:** Every n th individual or item is selected from the population, providing a systematic approach to ensure coverage, such as surveying every 10th person in a town of 5000.
- ❑ **Stratified Random Sampling:** The population is divided into subgroups based on characteristics (e.g., age, income), and a random sample is taken from each subgroup to ensure representation, useful in surveys needing diverse subgroup inclusion.
- ❑ **Cluster Sampling:** Sampling is done by selecting clusters (e.g., city blocks), making it cost-effective when a complete list of the population is not available or too expensive, like surveying all households in randomly chosen city blocks.

Answer Review Questions of Chapter 4

Explain the concept of point estimates and confidence interval estimates in the context of statistical analysis. How do these methods help in assessing the value of unknown population parameters? Illustrate your answer with examples from the document.

- 1. Point Estimates:** A single number used to estimate a population parameter. For example, the sample mean is a point estimate of the population mean. If you calculate the average height of a sample of students, this average (sample mean) serves as a point estimate for the average height of all students (population mean).
- 2. Confidence Interval Estimates:** A range of values believed to contain the population parameter with a specified probability. For instance, a 95% confidence interval for the mean height of students might be 150 cm to 160 cm. This means we are 95% confident that the true mean height of the population falls within this range.

Point estimates provide a single best guess of the population parameter, while confidence intervals give a range that likely includes the parameter, thus providing more information about the precision of the estimate. The width of the confidence interval depends on the sample size and variability in the data. Larger samples and less variability lead to narrower intervals, indicating more precise estimates.

Answer Review Questions of Chapter 4

Describe the importance of confidence intervals in statistical inference. What factors affect the width of a confidence interval, and how does the sample size influence the confidence interval? Provide a detailed explanation using examples provided in the document.

1. **Sample Size:** Larger sample sizes generally lead to narrower confidence intervals. This is because larger samples provide more information about the population, reducing uncertainty. For example, a sample size of 100 might yield a confidence interval for the mean that is 5 units wide, while a sample size of 1000 might reduce this width to 2 units.
2. **Variability in Data:** Higher variability in the data results in wider confidence intervals. If the data points are spread out, there is more uncertainty about the population parameter.
3. **Confidence Level:** Higher confidence levels result in wider intervals. For example, a 99% confidence interval will be wider than a 95% confidence interval for the same data, as it provides greater assurance that the interval includes the true parameter.

Objectives

- **Objective 1: Understand the Basics of Regression Analysis**
 - Learn the fundamental concepts of regression analysis.
 - Understand the relationship between dependent and independent variables.
 - Differentiate between simple and multiple regression models.
- **Objective 2: Apply Regression Analysis Techniques**
 - Utilize regression techniques to analyze data.
 - Interpret the coefficients and statistics of regression models.
 - Use software tools to perform regression analysis.

Objectives Cont.

- **Objective 3: Evaluate Regression Model Accuracy**
 - Assess the goodness-of-fit for regression models.
 - Understand and calculate R-squared and adjusted R-squared.
 - Evaluate model assumptions and detect potential violations..
- **Objective 4: Implement Regression Models in Business Contexts**
 - Apply regression analysis to solve business problems.
 - Use regression models for forecasting and decision making.
 - Interpret and communicate regression analysis results effectively.

What is Regression?

Relationships

In science, we frequently measure two or more variables on the same individual (case, object, etc). We do this to explore the nature of the relationship among these variables. There are two basic types of relationships.

- Cause-and-effect relationships.
- Functional relationships.

Function: a mathematical relationship enabling us to predict what values of one variable (**Y**) correspond to given values of another variable (**X**).

- **Y:** is referred to as the *dependent variable*, the *response variable* or the *predicted variable*.
- **X:** is referred to as the *independent variable*, the *explanatory variable* or the *predictor variable*.

Examples

- The time needed to fill a soft drink vending machine
- The tensile strength of wrapping paper
- Percent germination of begonia seeds
- The mean litter weight of test rats
- Maintenance cost of tractors
- The repair time for a computer
- The number of cases needed to fill the machine
- The percent of hardwood in the pulp batch
- The intensity of light in an incubator
- The litter size
- The age of the tractor
- The number of components which have to be changed

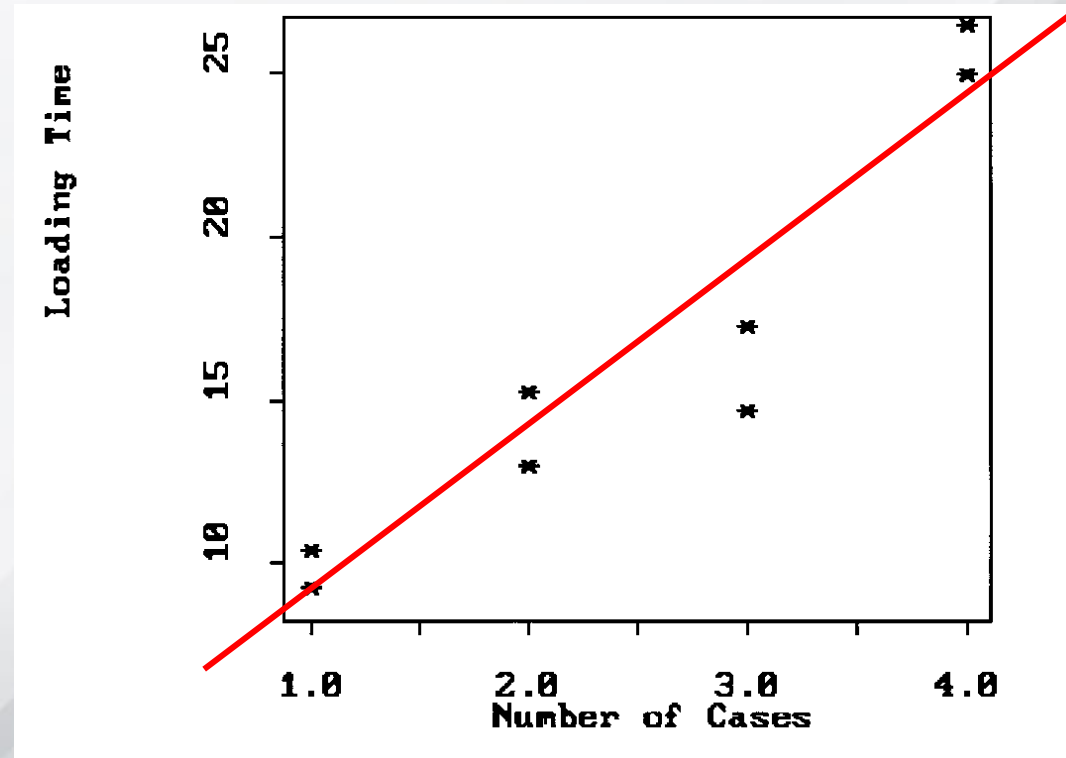
In each case, the statement can be read as; ***Y is a function of X.***

Two kinds of explanatory variables:

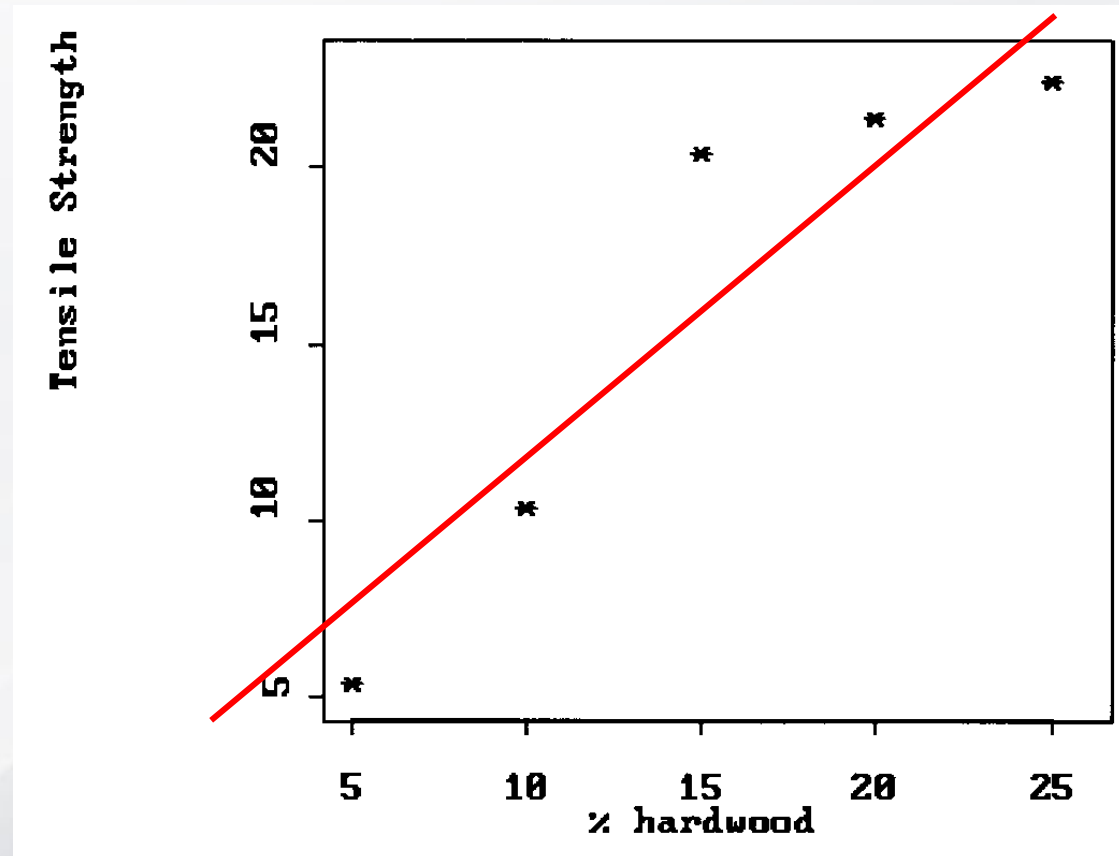
Those we can control

Those over which we have little or no control.

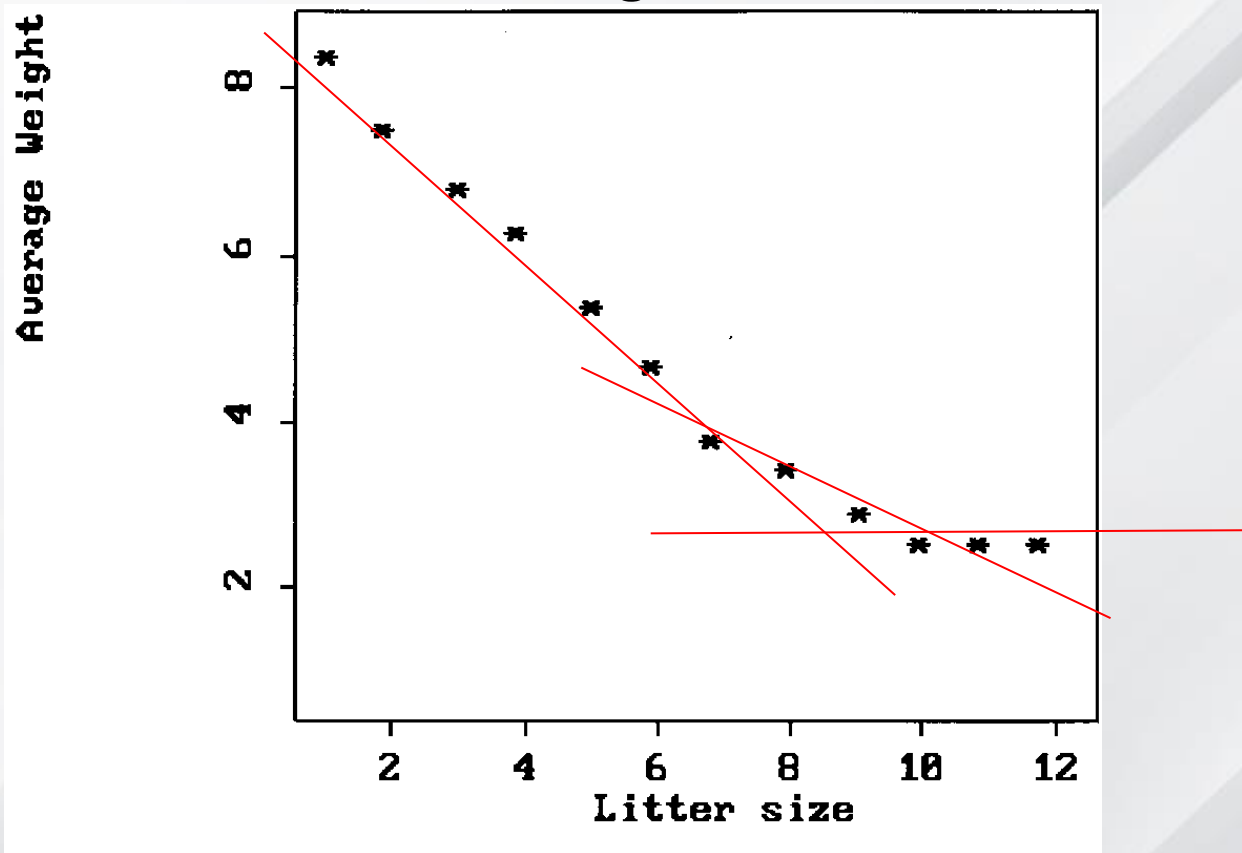
An operations supervisor measured how long it takes one of her drivers to put 1, 2, 3 and 4 cases of soft drink into a soft drink machine. In this case the levels of the explanatory variable, X are $\{1, 2, 3, 4\}$, and she controls them. She might repeat the measurement a couple of times at each level of X . A scatter plot of the resulting data might look like:



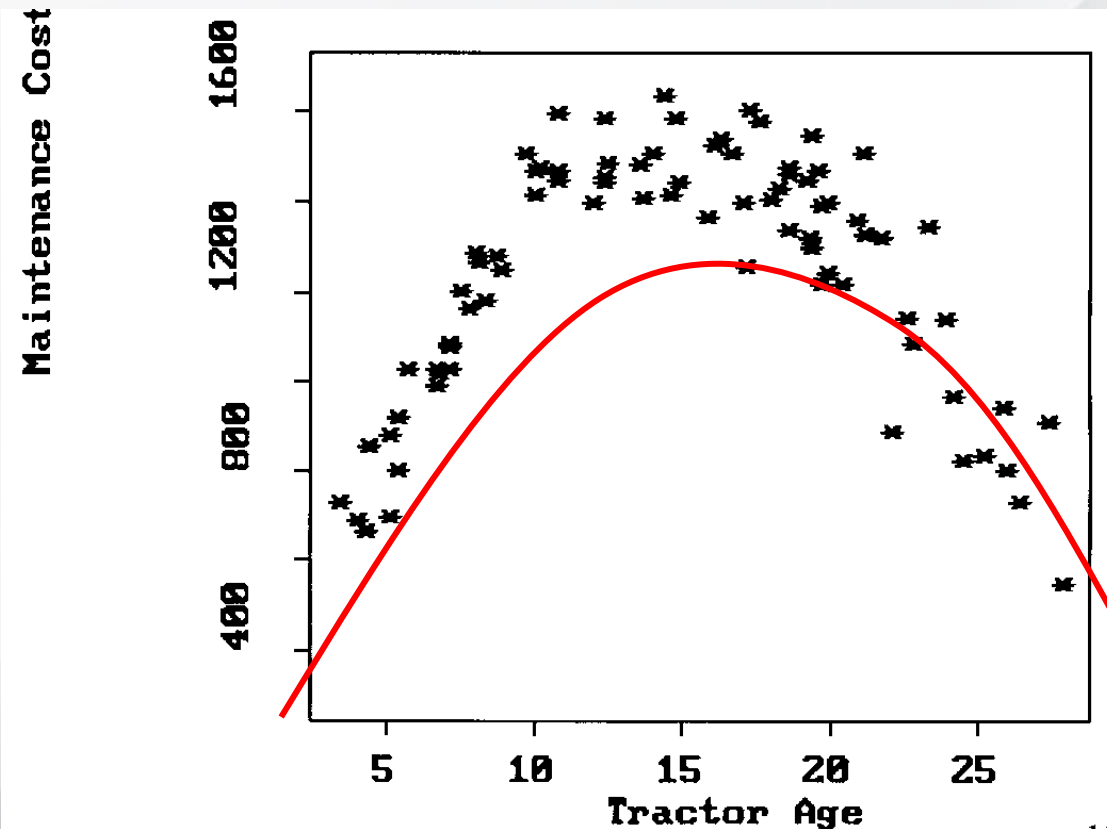
A forestry graduate student makes wrapping paper out of different percentages of hardwood then measure its tensile strength. He has the freedom to choose at the beginning of the study to have only five percentages to work with, say {5%, 10%, 15%, 20%, and 25%}. A scatter plot of the resulting data might look like:



A farm manager is interested in the relationship between litter size and average litter weight (average newborn piglet weight). She examines the farm records over the last couple of years and records the litter size and average weight for all births. A plot of the data pairs looks like the following:



A farm operations student is interested in the relationship between maintenance cost and age of farm tractors. He performs a telephone interview survey of the 52 commercial potato growers in Putnam County, FL. One part of the questionnaire provides information on tractor age and 1995 maintenance cost (fuel, lubricants, repairs, etc). A plot of these data might look like:



Questions needing answers.

- What is the *association* between Y and X?
- How can changes in Y be explained by changes in X?
- What are the *functional relationships* between Y and X?

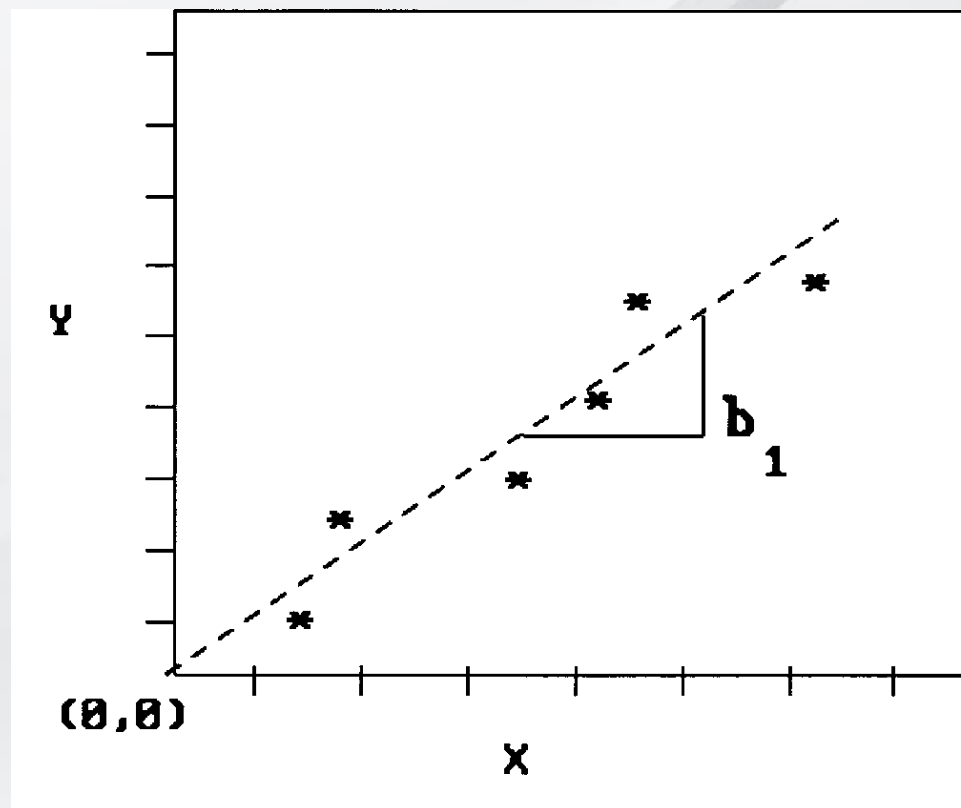
A functional relationship is symbolically written as:

Eq: 1 $Y = f(X)$

Example: A proportional relationship (e.g. fish weight to length).

$$Y = b_1 X$$

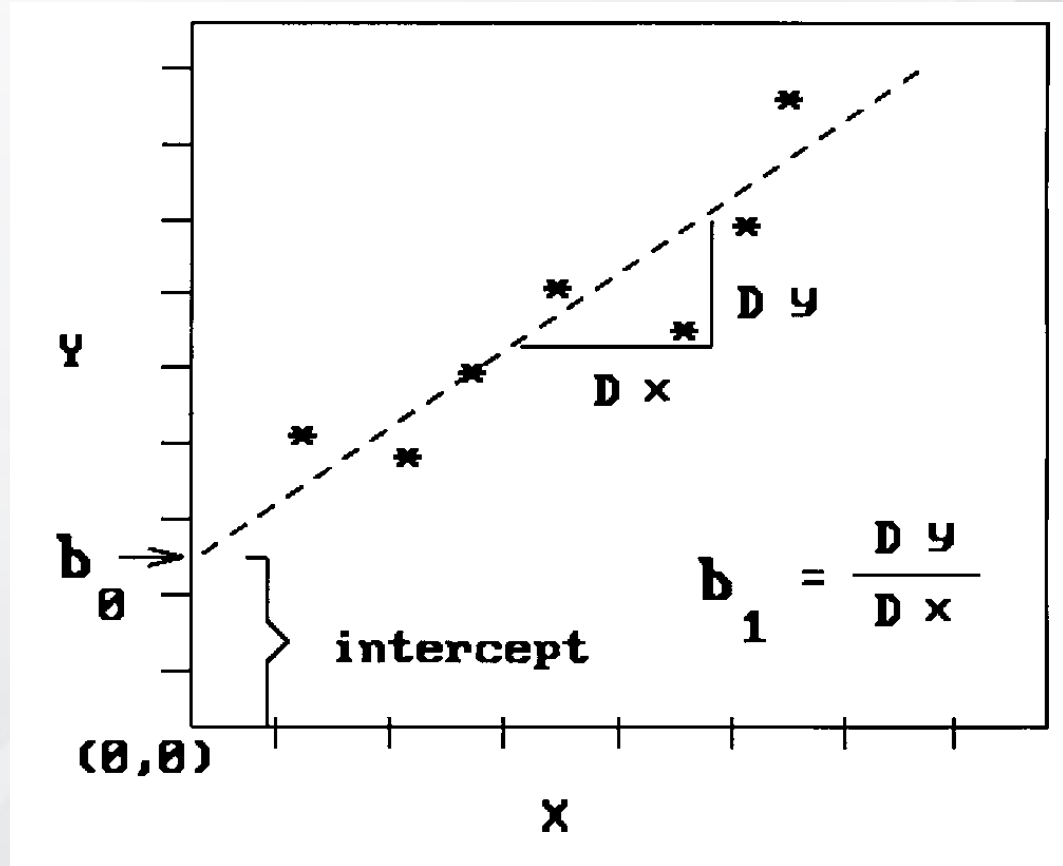
b_1 is the **slope** of the line.



Example: Linear relationship (e.g. Y=cholesterol versus X=age)

$$Y = b_0 + b_1 X$$

b_0 is the intercept,
 b_1 is the slope.



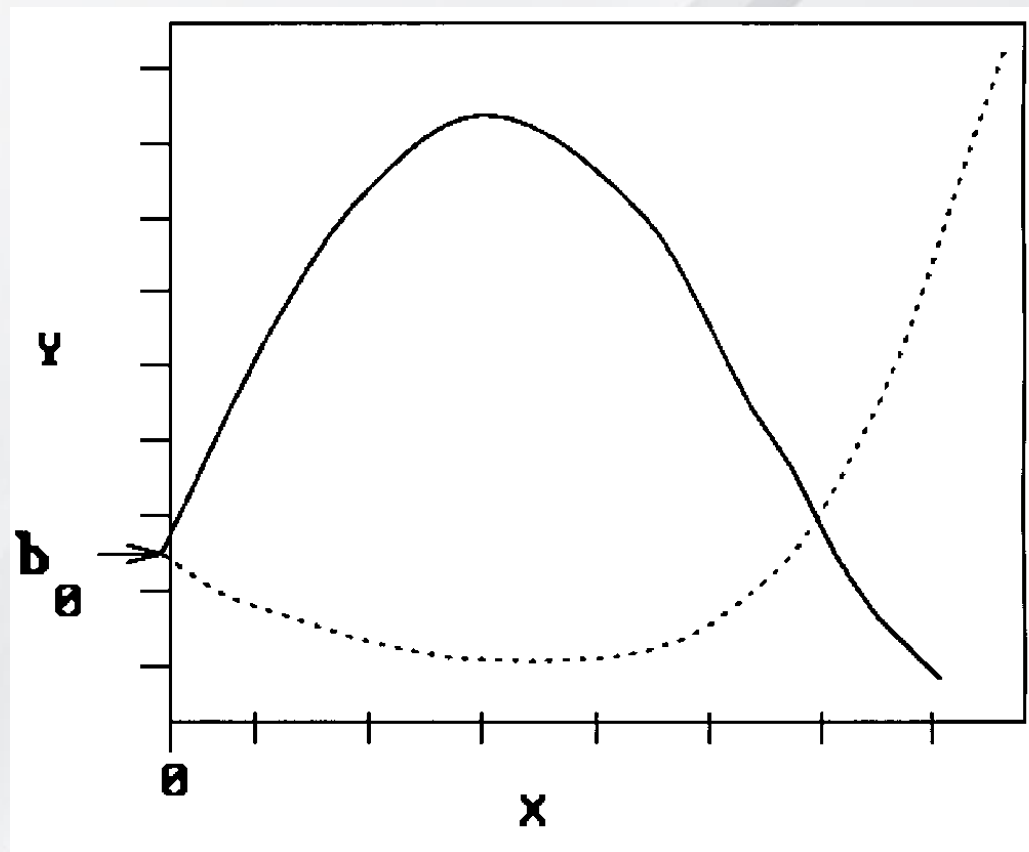
Example: Polynomial relationship (e.g. Y=crop yield vs. X=pH)

$$Y = b_0 + b_1X + b_2X^2$$

b_0 : intercept,

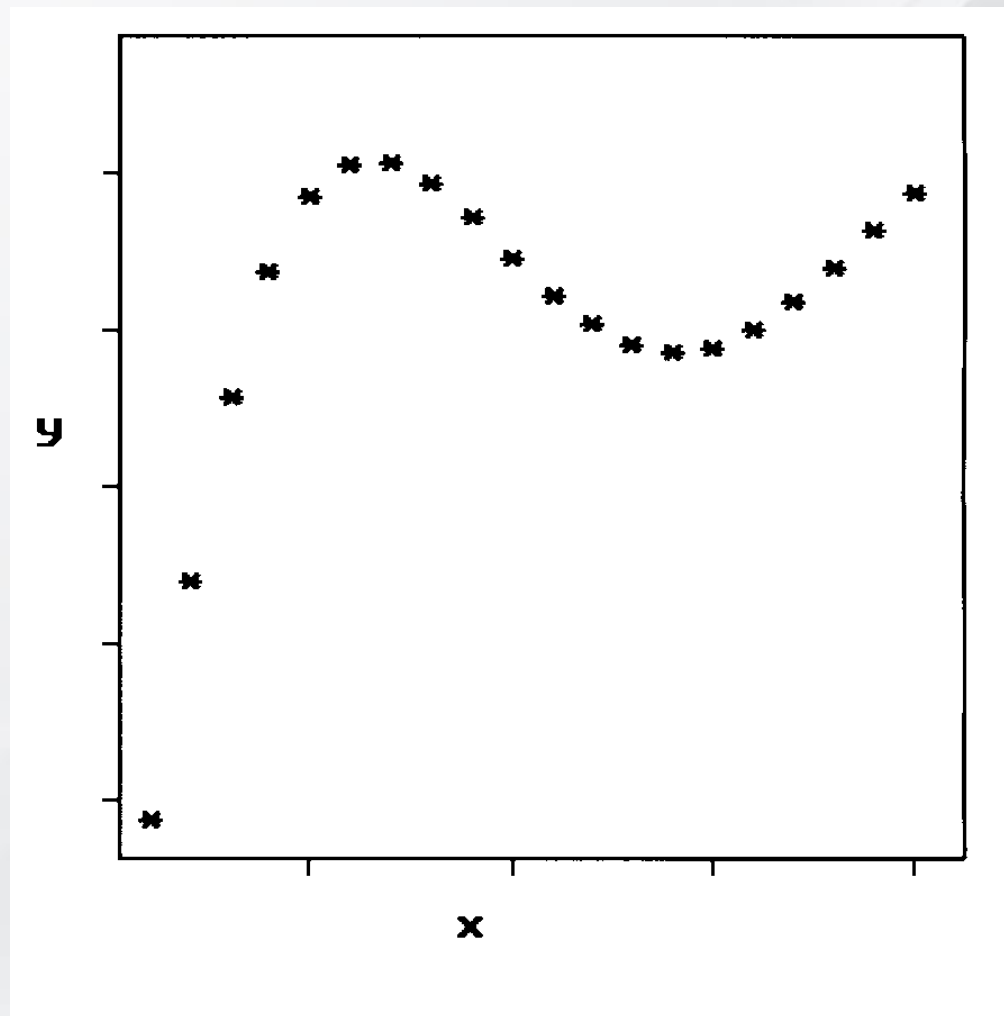
b_1 : linear coefficient,

b_2 : quadratic coefficient.



Nonlinear relationship:

$$Y = b_0 \sin(b_1 X + b_2 X^2)$$



Concerns:

- The proposed functional relationship will not fit exactly, i.e. something is either wrong with the data (**errors in measurement**), or the model is inadequate (**errors in specification**).
- The relationship is not truly known until we assign values to the **parameters** of the model.

The possibility of errors into the proposed relationship is acknowledged in the functional symbolism as follows:

Eq: 2
$$Y = f(X) + \varepsilon$$

ε is a random variable representing the result of both errors in model specification and measurement. As in AOV, the variance of ε is the **background variability** with respect to which we will assess the significance of the factors (explanatory variables).

The error term:

Another way to emphasize

Eq: 3 $\varepsilon = Y - f(X)$

or, emphasizing that $f(X)$ depends on unknown parameters.

Eq: 4 $Y = f(X \mid \beta_0, \beta_1) + \varepsilon$



What if we don't know the functional form of the relationship?

- Look at a scatter plot of the data for suggestions.
- Hypothesize about the nature of the underlying process. Often the hypothesized processes will suggest a functional form.

The straight line -- a conservative starting point.

Regression Analysis: the process of fitting a line to data.

Sir Francis Galton (1822-1911) -- a British anthropologist and meteorologist coined the term “regression”.

Regression towards mediocrity in hereditary stature - the tendency of offspring to be smaller than large parents and larger than small parents. Referred to as “regression towards the mean”.

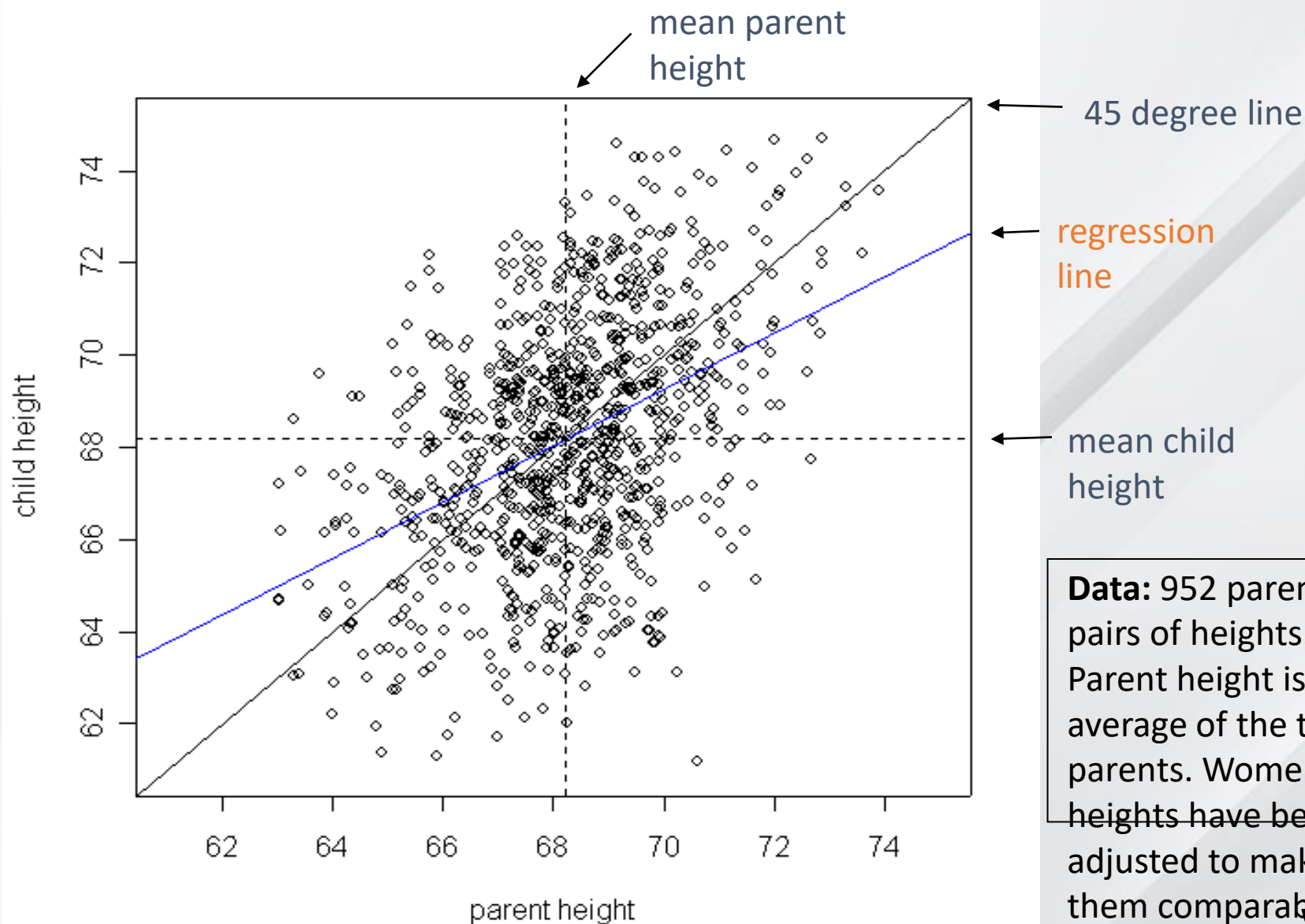
$$\hat{Y} = \bar{Y} + \frac{2}{3}(X - \bar{X})$$

Expected offspring height

Average sized offspring

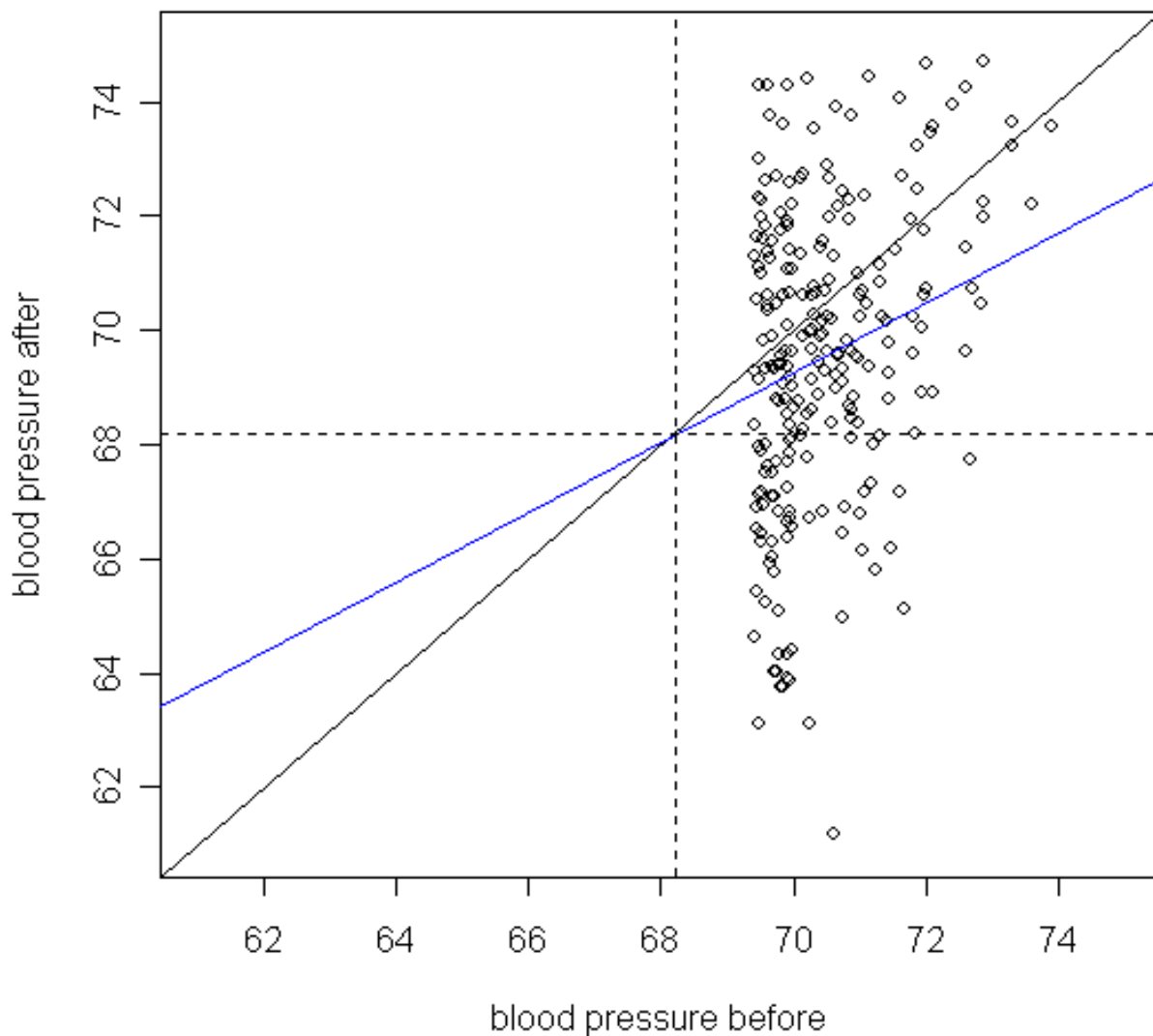
Adjustment for how far parent is from mean of parents

Regression to the Mean: Galton's Height Data



Data: 952 parent-child pairs of heights. Parent height is average of the two parents. Women's heights have been adjusted to make them comparable to men's.

Regression to the Mean is a Powerful Effect!

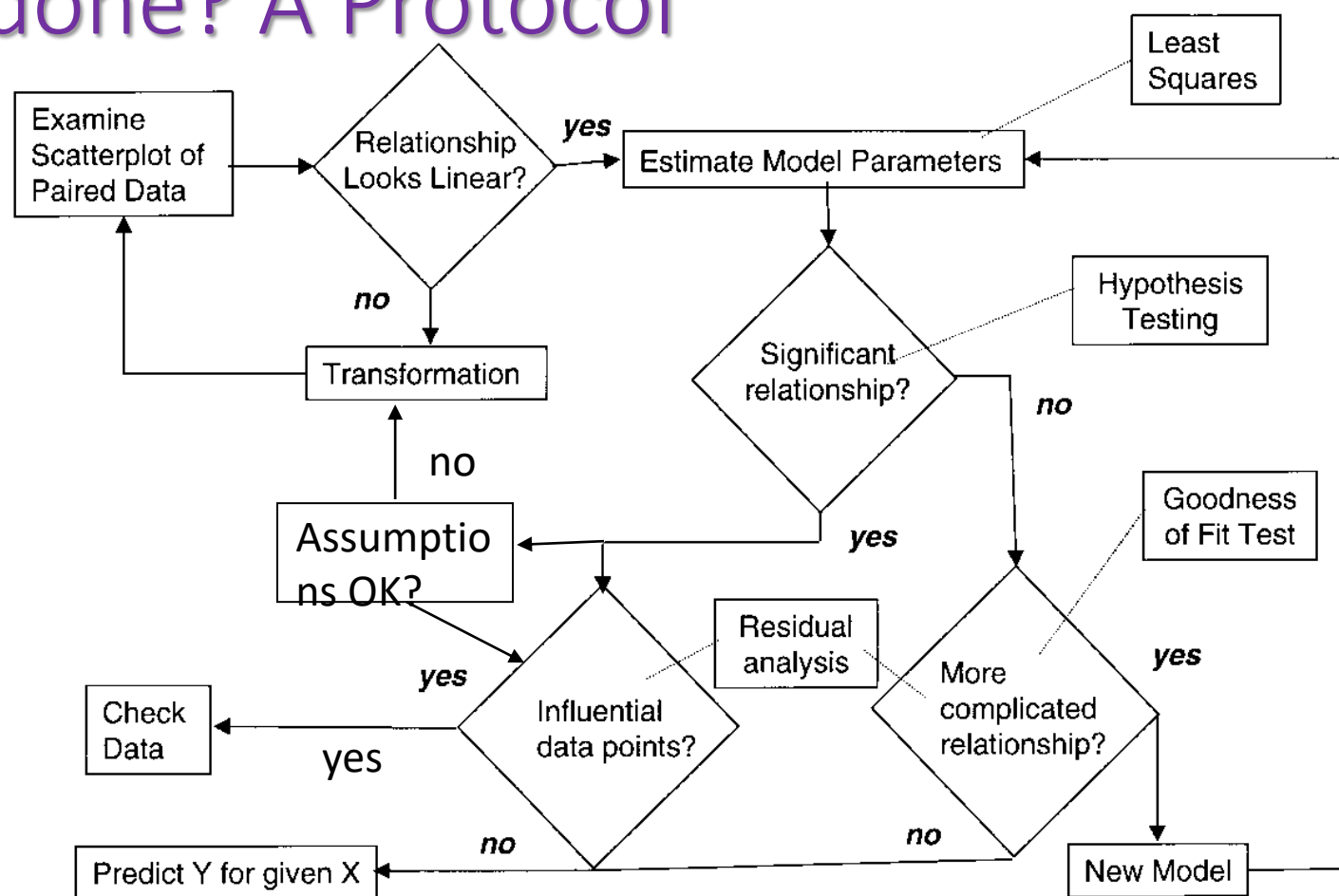


Same data, but suppose response is now blood pressure (bp) before & after (day 1, day 2).

If we track only those with elevated bp before (above 3rd quartile), we see an amazing improvement, even though no treatment took place!

This is the regression effect at work. If it is not recognized and taken into account, misleading results and biases can occur.

How is a Simple Linear Regression Analysis done? A Protocol



11-16

Steps in a Regression Analysis

1. **Examine the scatterplot of the data.**
 - Does the relationship look linear?
 - Are there points in locations they shouldn't be?
 - Do we need a transformation?
2. **Assuming a linear function looks appropriate, estimate the regression parameters.**
 - How do we do this? (Method of Least Squares)
3. **Test whether there really is a statistically significant linear relationship. Just because we assumed a linear function it does not follow that the data support this assumption.**
 - How do we test this? (F-test for Variances)
4. **If there is a significant linear relationship, estimate the response, Y , for the given values of X , and compute the residuals.**
5. **Examine the residuals for systematic inadequacies in the linear model as fit to the data.**
 - Is there evidence that a more complicated relationship (say a polynomial) should be considered; are there problems with the regression assumptions? (Residual analysis).
 - Are there specific data points which do not seem to follow the proposed relationship? (Examined using influence measures).

Simple Linear Regression - Example and Theory

SITUATION: A company that repairs small computers needs to develop a better way of providing customers typical repair cost estimates. To begin this process, they compiled data on repair times (in minutes) and the number of components needing repair or replacement from the previous week. The data, sorted by number of components are as follows:

Paired Observations (x_i, y_i)

	Number of components	Repair time
i	x_i	y_i
1	1	23
2	2	29
3	4	64
4	4	72
5	4	80
6	5	87
7	6	96
8	6	105
9	8	127
10	8	119
11	9	145
12	9	149
13	10	165
14	10	154

Assumed Linear Regression Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

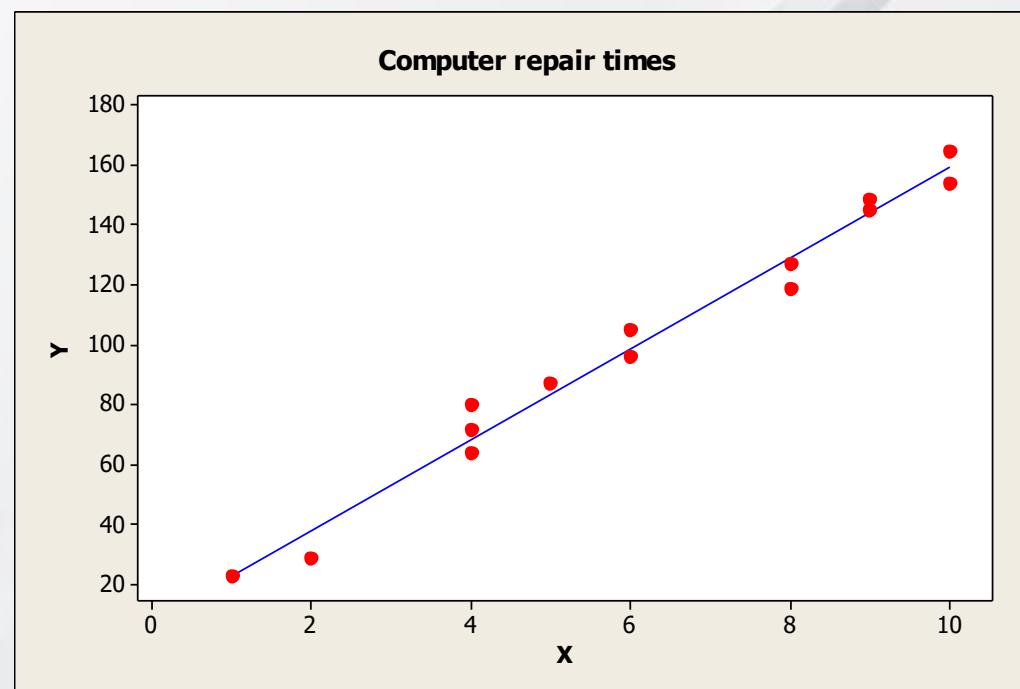
for $i = 1, 2, \dots, n$

Estimating the regression parameters

Objective: Minimize the difference between the observation and its prediction according to the line.

$$\begin{aligned}\varepsilon_i &= y_i - \hat{y}_i \\ &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\end{aligned}$$

$\hat{y}_i$ = predicted y value when $x = x_i$



We want the line which is best for all points. This is done by finding the values of β_0 and β_1 which minimizes some sum of errors. There are a number of ways of doing this. Consider these two:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n |\varepsilon_i|$$

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n \varepsilon_i^2$$

← *Sum of squared residuals*

The method of least squares produces estimates with statistical properties (e.g. sampling distributions) which are easier to determine.

Regression => least squares estimation

$$\hat{\beta}_0 \quad \hat{\beta}_1$$

Referred to as least squares estimates.

Parameter estimates:

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Easy to compute with a spreadsheet program.
Easier to do with a statistical analysis package.

Example:

$$\hat{\beta}_1 = 7.71$$

$$\hat{\beta}_0 = 15.20$$

$$\hat{y}_i = 15.20 + 7.71x_i \quad \text{Prediction}$$

Testing for a Statistically Significant Regression

H_0 : There is no relationship between Y and X.

H_A : There is a relationship between Y and X.

Which of two competing models is more appropriate?

Linear Model: $Y = \beta_0 + \beta_1 X + \varepsilon$

Mean Model: $Y = \mu + \varepsilon$

We look at the sums of squares of the prediction errors for the two models and decide if that for the linear model is **significantly smaller** than that for the mean model.

Sums of Squares About the Mean (TSS)

Sum of squares about the mean: sum of the prediction errors for the null (mean model) hypothesis.

$$TSS = S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

TSS is actually a measure of the variance of the responses.

Residual Sums of Squares

Sum of squares for error: sum of the prediction errors for the alternative (linear regression model) hypothesis.

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

SSE measures the variance of the residuals, the part of the response variation that is not explained by the model.

Regression Sums of Squares

Sum of squares due to the regression: difference between TSS and SSE, i.e. $SSR = TSS - SSE$.

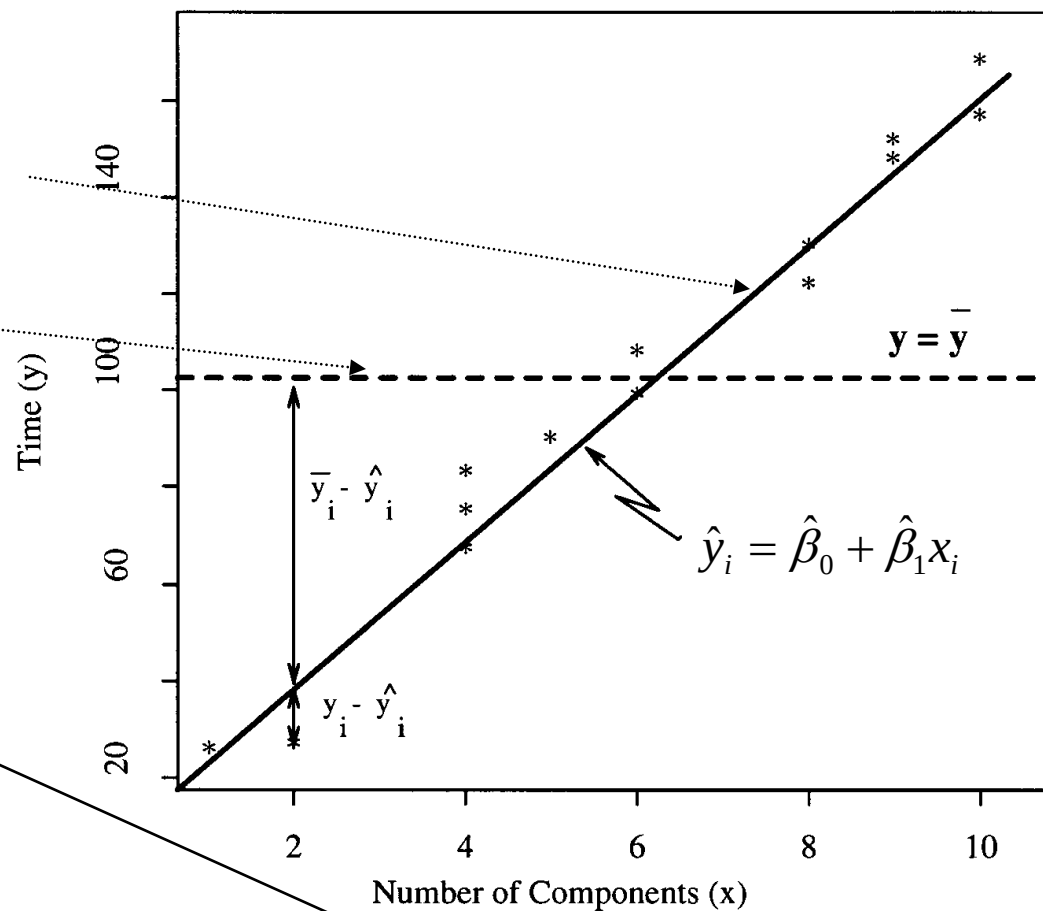
$$\begin{aligned} SSR &= \sum_{i=1}^n (y_i - \bar{y}_i)^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (\bar{y} - \hat{y}_i)^2 \end{aligned}$$

SSR measures how much variability in the response is explained by the regression.

Graphical View

Linear Model

Mean Model



$$TSS = SSR + SSE$$

Total
variability in
y-values

=

Variability
accounted for
by the
regression

+

Unexplained
variability

Mean Square Terms

Mean Square Total

$$\begin{aligned}\hat{\sigma}_T^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \frac{\text{TSS}}{n-1} \\ &= \text{MST}\end{aligned}$$

Sample variance of
the response, y:

$$\begin{aligned}\hat{\sigma}_R^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\ &= \frac{\text{SSR}}{1} \\ &= \text{MSR}\end{aligned}$$

Regression Mean Square:

Residual Mean Square

$$\begin{aligned}\hat{\sigma}_\varepsilon^2 &= \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \frac{\text{SSE}}{n-2} \\ &= \text{MSE}\end{aligned}$$

F Test for Significant Regression

Both MSE and MSR measure the same underlying variance quantity under the assumption that the null (mean) model holds.

$$\sigma_R^2 \approx \sigma_\varepsilon^2$$

Under the alternative hypothesis, the MSR should be much greater than the MSE.

$$\sigma_R^2 > \sigma_\varepsilon^2$$

Placing this in the context of a test of variance.

$$F = \frac{\sigma_R^2}{\sigma_\varepsilon^2} = \frac{\text{MSR}}{\text{MSE}}$$

Test Statistic

F should be near 1 if the regression is *not significant*, i.e. H_0 : mean model holds.

Assumptions

1. $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are independent of each other.
2. The ε_i are normally distributed with mean zero and have common variance σ^2 .

How do we check these assumptions?

- I. Appropriate graphs.
- II. Correlations (more later).
- III. Formal goodness of fit tests.

Analysis of Variance Table

We summarize the computations of this test in a table.

Source	Sums of Squares SSQ	Degrees of Freedom DF	Mean Squares MS	F
Regression	SSR	1	MSR	$F = \frac{MSR}{MSE}$
Error	SSE	n-2	MSE	
Total	SSM	n-1		

↑
TSS

Parameter Standard Error Estimates

Under the assumptions for regression inference, the least squares estimates themselves are random variables.

1. $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are independent of each other.
2. The ε_i are normally distributed with mean zero and have common variance σ^2 .

Using some more calculus and mathematical statistics we can determine the distributions for these parameters.

$$\hat{\beta}_0 \mapsto N\left(\beta_0, \sigma^2 \frac{\sum x_i^2}{nS_{XX}}\right) \quad \hat{\beta}_1 \mapsto N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right)$$

Testing regression parameters

The estimate of σ^2 is the mean square error: MSE

important

$$\hat{\sigma}^2 = MSE$$

Test $H_0: \beta_1=0$:

$$t_{\beta_1} = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{MSE}{S_{XX}}}}$$

Reject H_0 if:

$$|t_{\beta_1}| > t_{n-2, \alpha/2}$$

$(1-\alpha)100\%$ CI for β_1 :

$$\hat{\beta}_1 \pm t_{n-2, \alpha/2} \sqrt{\frac{MSE}{S_{XX}}}$$

Table 3: SAS Output

		Parameter Estimates			
		(1)	(2)	(3)	(4)
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	7.710997	4.11488737	1.874	0.0855
NCOMP	1	15.198210	0.60860005	24.972	0.0001

$\hat{\beta}_0$ P-values
 $\hat{\beta}_1$ $\sqrt{MSE/S_{XX}}$ $t_{\beta_1} = \frac{\hat{\beta}_1 - 0}{\sqrt{MSE/S_{XX}}}$

- Purpose: Regression analysis examines relationships between dependent and independent variables.
- Types: Includes simple regression (one independent variable) and multiple regression (multiple independent variables).
- Key Techniques:
 - Estimating model coefficients.
 - Testing for significance.
 - Analyzing residuals.
- Model Accuracy:
 - Assess using R-squared and adjusted R-squared.
 - Check for assumption violations.
- Applications:
 - Forecasting and predicting outcomes.
 - Solving business problems.
 - Supporting strategic decision-making.
- Tools: Software like Excel, SPSS, R, and Python for performing regression analysis.

Question 1:

What is the purpose of regression analysis?

Question 2:

What is the difference between simple regression and multiple regression?

Question 3:

What are some common applications of regression analysis?



Thank You