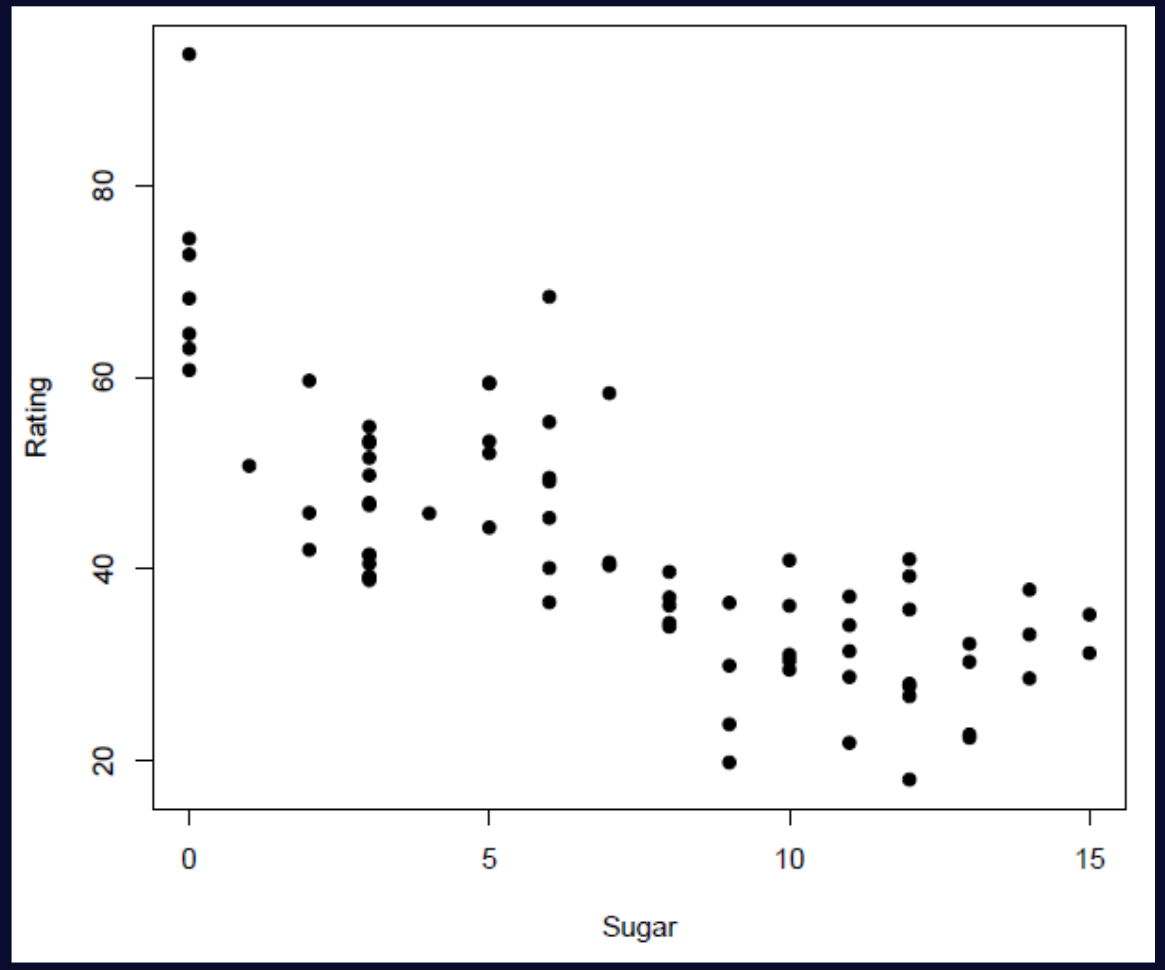
# Simple Linear Regression

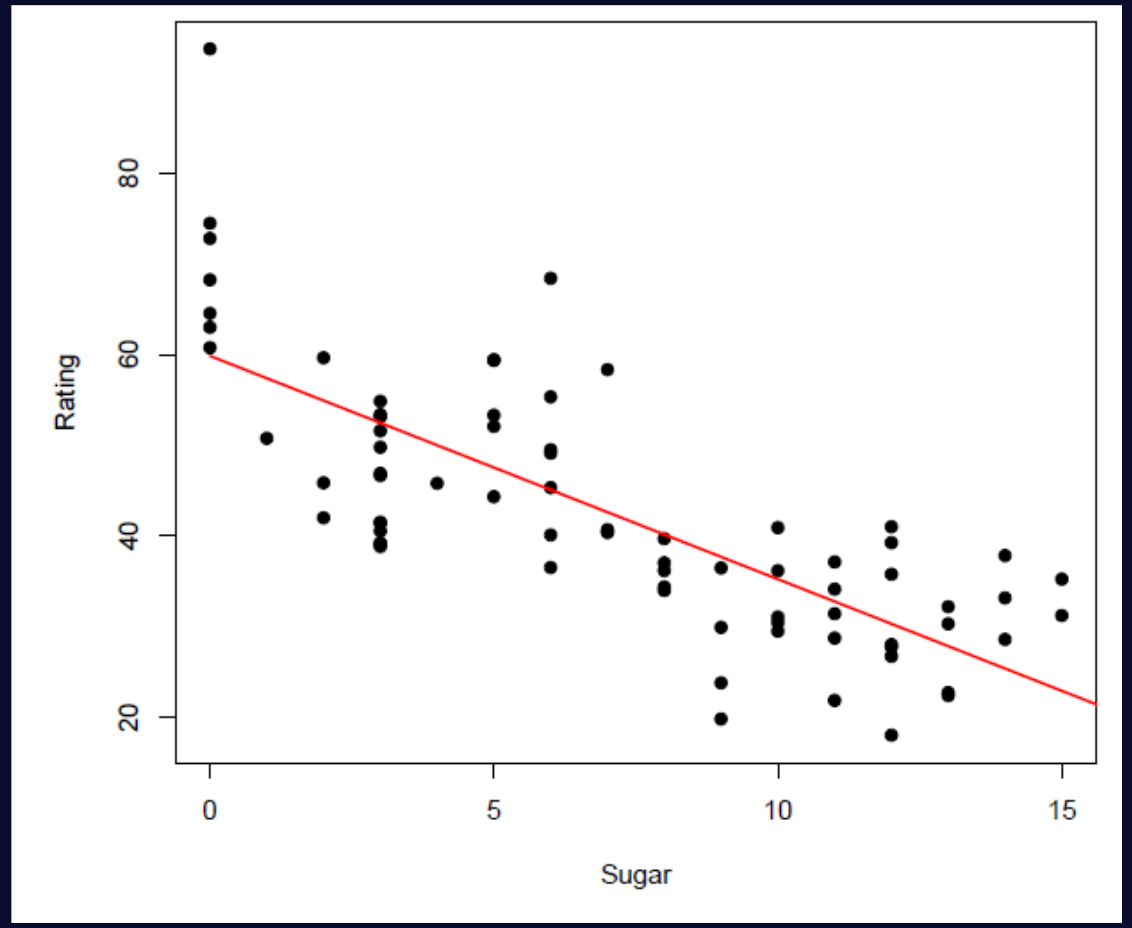
INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# What is linear regression?




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# What is linear regression?




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# What is linear regression?

- The simplest deterministic *mathematical* relationship between two variables *x* and *y* is a linear relationship:

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

- The objective of this section is to develop an equivalent *linear probabilistic model*.
- If the two (random) variables are probabilistically related, then for a fixed value of *x*, there is uncertainty in the value of the second variable.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# What is linear regression?

- So, we assume $Y = \beta_0 + \beta_1 x + \varepsilon$ where ε is a random variable.

- 2 variables are related linearly “on average” if for fixed x the actual value of Y differs from its expected value by a random amount (i.e. there is random error).

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Simple Linear Regression Model

Definition:

There are parameters $\beta_0, \beta_1$, and $\sigma^2$, such that for any fixed value of the independent variable x, the dependent variable is a random variable related to x through the model equation

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

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Simple Linear Regression Model

Definition

The quantity ε in the model equation is the “error” -- a random variable, assumed to be symmetrically distributed with

$$E(\varepsilon) = 0 \text{ and } V(\varepsilon) = \sigma_\varepsilon^2 = \sigma^2$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Simple Linear Regression Model

We also use the following notation:

- X: the independent, predictor, or explanatory variable (usually known). NOT RANDOM.
- Y: The dependent or response variable. For fixed x, Y will be random variable.
- ε: The random deviation or random error term. For fixed x, ε will be random variable.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Simple Linear Regression Model

Interpreting parameters:

- $\beta_0$ (the intercept of the true regression line):

  The average value of Y when x is zero.

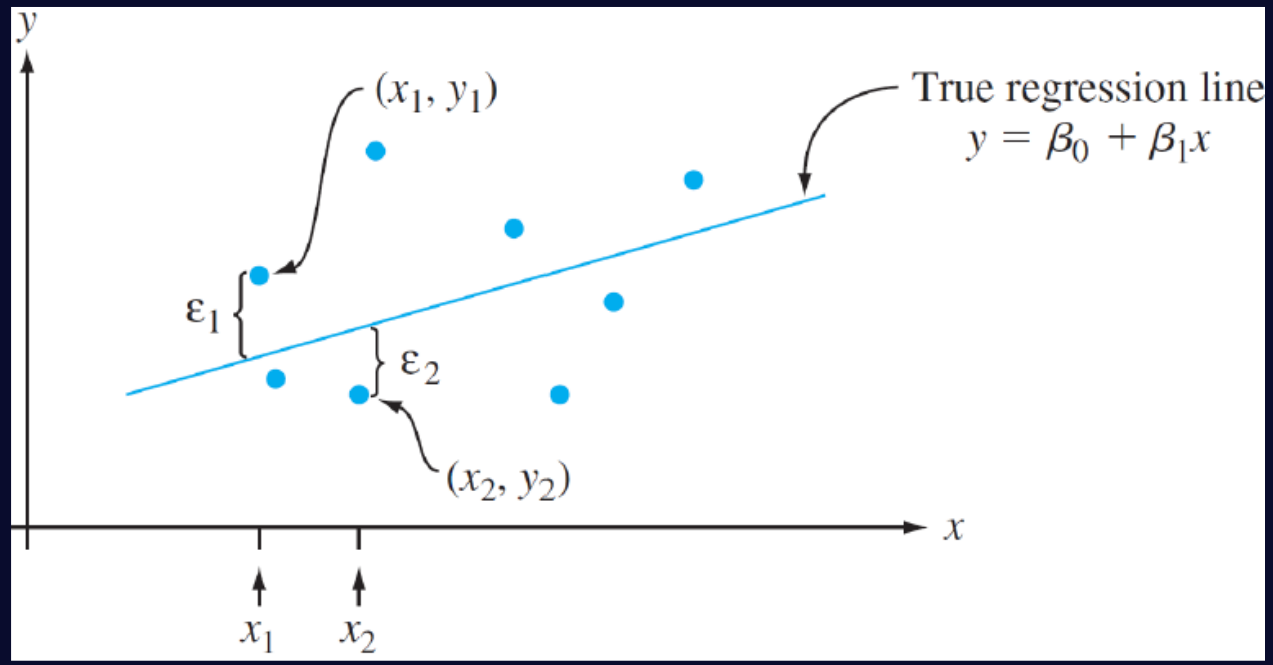
- $\beta_1$ (the slope of the true regression line):

  The expected (average) change in Y associated with a 1-unit increase in the value of x.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Simple Linear Regression Model

- The points $(x_1, y_1), \ldots, (x_n, y_n)$ resulting from n independent observations will then be scattered about the true regression line:

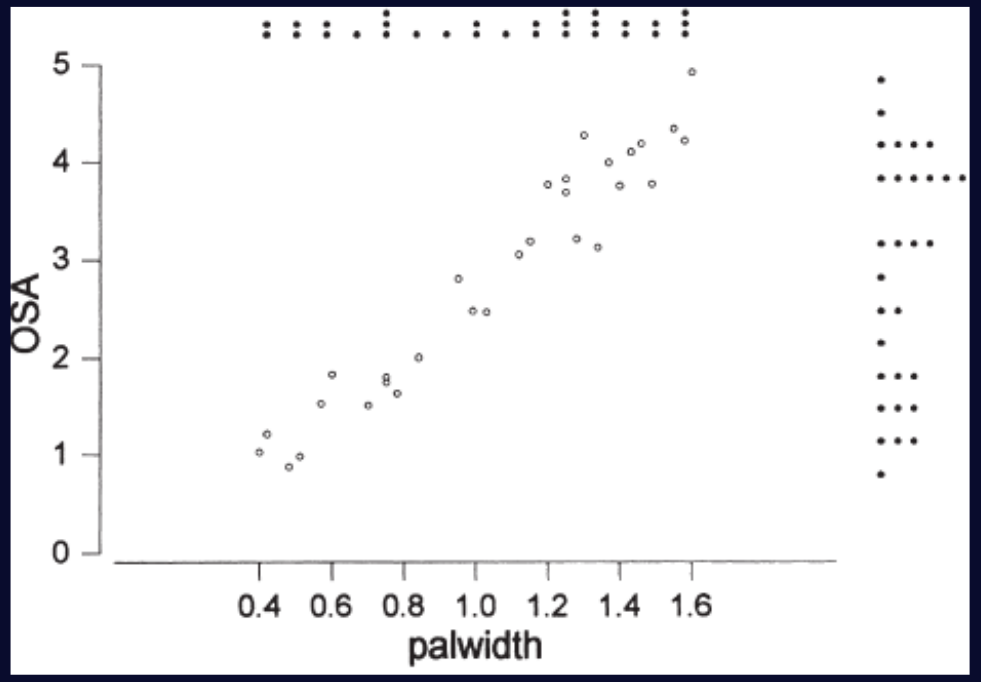



INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Simple Linear Regression Model

- How do we know simple linear regression is appropriate?

Scatterplots!




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Simple Linear Regression Model

- If we think of an entire population of (x, y) pairs, then $\mu_{Y|X^*}$ is the mean of all y values for which $x = X^*$, and $\sigma^2_{Y|x^*}$ is a measure of how much these values of y spread out about the mean value.

- Homoscedasticity: We assume the variance (amount of variability) of the distribution of *Y* values to be the same at each different value of fixed *x.* (*i.e.* homogeneity of variance assumption).

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Simple Linear Regression Model

- For example,

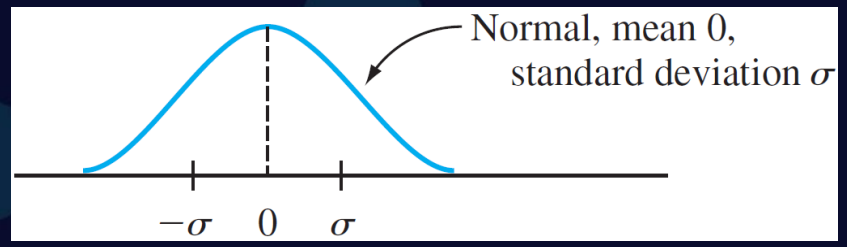
  x = age of a child  y = vocabulary size,

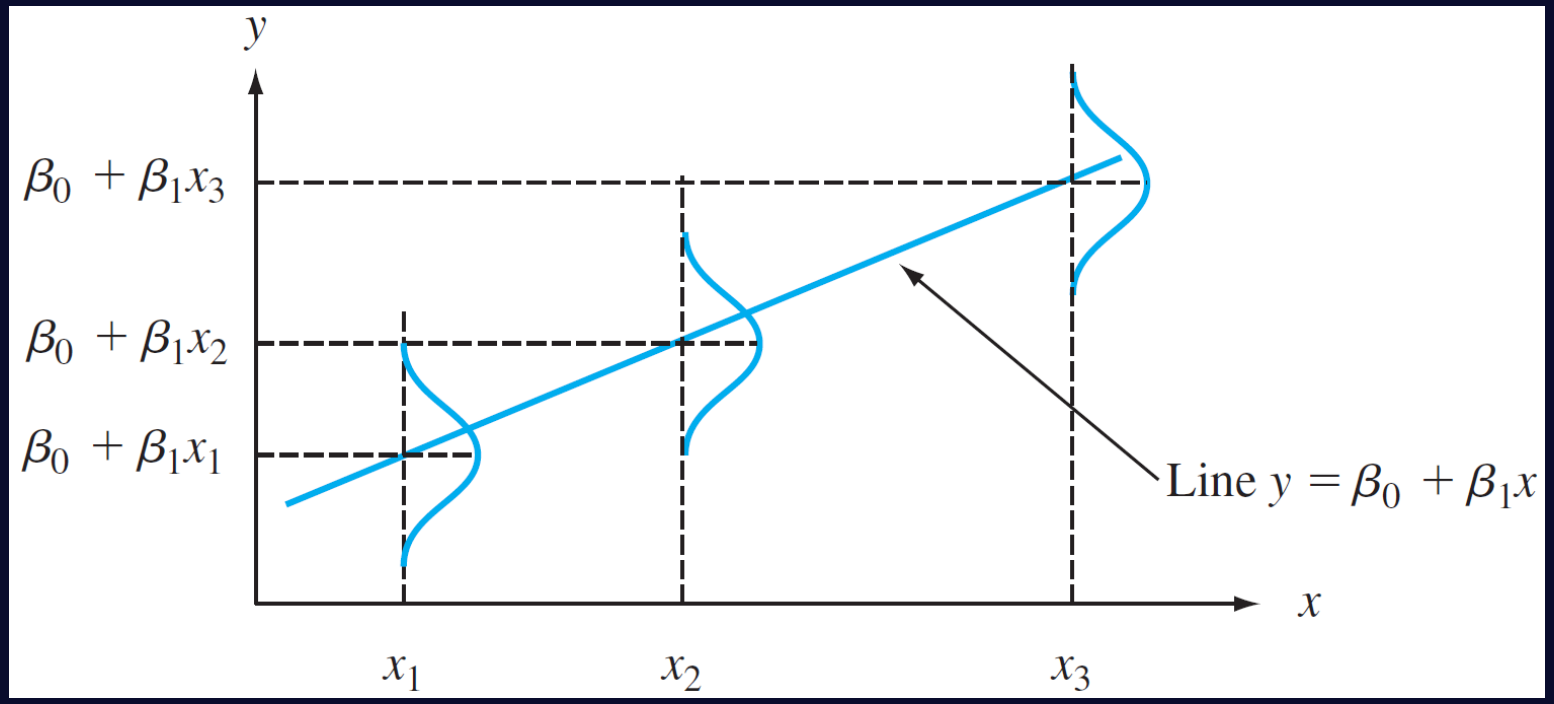
  then $\mu_{Y|5}$ is the average vocabulary size for all 5-year-old children in the population, and $\sigma^2_{Y|5}$ describes the amount of variability in vocabulary size for this part of the population.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# When errors are normally distributed…

- The variance parameter $\sigma^2$ determines the extent to which each normal curve spreads out about the regression line




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# When errors are normally distributed…

- When $\sigma^2$ is small, an observed point (x, y) will almost always fall quite close to the true regression line, whereas observations may deviate considerably from their expected values (corresponding to points far from the line) when $\sigma^2$ is large.

- Thus, this variance can be used to tell us how good the linear fit is
- But how do we define “good”?

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Simple Linear Regression Model

- The values of $\beta_0, \beta_1$ and $\sigma^2$ will almost never be known to an investigator.
- Instead, sample data consists of n observed pairs

$$(x_1, y_1), \ldots, (x_n, y_n)$$

from which the model parameters and the true regression line itself can be estimated. The data (pairs) are assumed to have been obtained independently of one another.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Simple Linear Regression Model

- Where,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ for } i = 1, 2, \ldots, n$$

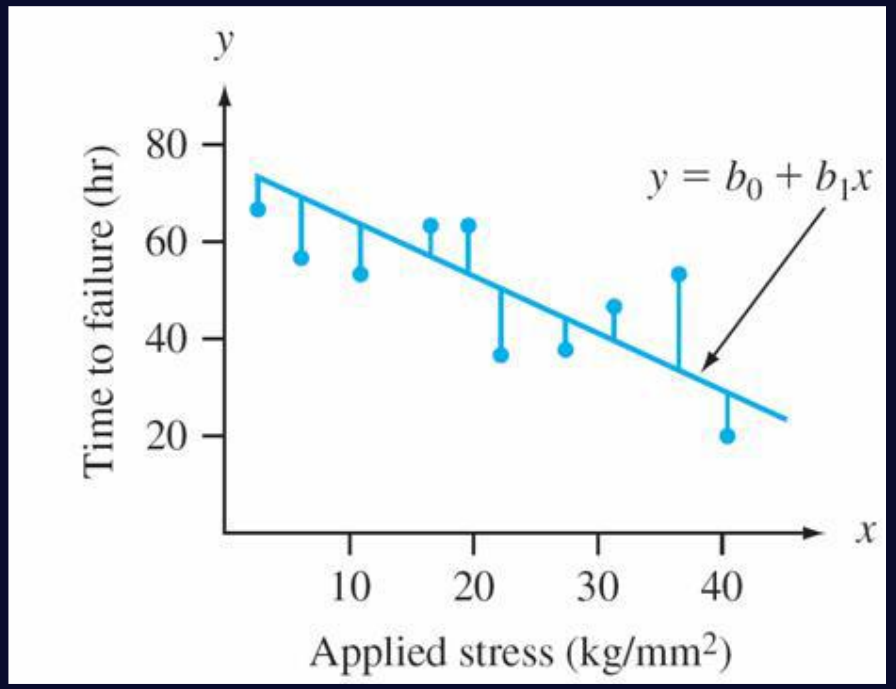
and the $n$ deviations $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent r.v.'s.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Estimating Model Parameters

- The “best fit” line is motivated by the principle of least squares, which can be traced back to the German mathematician Gauss (1777–1855):

- *A line provides the **best fit** to the data if the sum of the squared vertical distances (deviations) from the observed points to that line is as small as it can be.*




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Estimating Model Parameters

- The sum of squared vertical deviations from the points $(x_1, y_1), \ldots, (x_n, y_n)$ to the line is then

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

- The point estimates of $\beta_0, \beta_1$ , denoted by of $b_0, b_1$ are called the least squares estimates – they are those values that minimize f.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Estimating Model Parameters

- The fitted regression line or least squares line is then the line whose equation is $y = b_0 + b_1 x$
- The minimizing values of $b_0, b_1$ are found by taking partial derivatives of f($b_0, b_1$) with respect to both $b_0, b_1$. Equating them both to zero [analogously to f''(b) = 0 in univariate calculus], and solving the equations.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Estimating Model Parameters

- Solving a convex optimization problem:

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - b_0 - b_1 x_i)(-1) = 0$$

$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Estimating Model Parameters

- Solving the equations will give us the following:

$$b_1 = \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

- The least squares estimate of the slope coefficient $\beta_1$ of the true regression line is

$$S_{xy} = \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n \qquad S_{xx} = \Sigma x_i^2 - (\Sigma x_i)^2/n$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Estimating Model Parameters

- For the intersection

$$b_0 = \hat{\beta}_0 = \frac{\Sigma y_i - \hat{\beta}_1 \Sigma x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example

- The cetane number is a critical property in specifying the ignition quality of a fuel used in a diesel engine.
- Determination of this number for a biodiesel fuel is expensive and time-consuming.
- The article “Relating the Cetane Number of Biodiesel Fuels to Their Fatty Acid Composition: A Critical Study” (J. of Automobile Engr., 2009: 565–583) included the following data on x = iodine value (g) and y = cetane number for a sample of 14 biofuels (see next slide).

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example

- The iodine value (x) is the amount of iodine necessary to saturate a sample of 100 g of oil. The article's authors fit the simple linear regression model to this data, so let's do the same.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 132.0 | 129.0 | 120.0 | 113.2 | 105.0 | 92.0 | 84.0 | 83.2 | 88.4 | 59.0 | 80.0 | 81.5 | 71.0 | 69.2 |
| $y$ | 46.0 | 48.0 | 51.0 | 52.1 | 54.0 | 52.0 | 59.0 | 58.7 | 61.6 | 64.0 | 61.4 | 54.6 | 58.8 | 58.0 |

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example

| $x$ | 132.0 | 129.0 | 120.0 | 113.2 | 105.0 | 92.0 | 84.0 | 83.2 | 88.4 | 59.0 | 80.0 | 81.5 | 71.0 | 69.2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 46.0 | 48.0 | 51.0 | 52.1 | 54.0 | 52.0 | 59.0 | 58.7 | 61.6 | 64.0 | 61.4 | 54.6 | 58.8 | 58.0 |

$$\Sigma x_i = 1307.5, \qquad \Sigma y_i = 779.2,$$

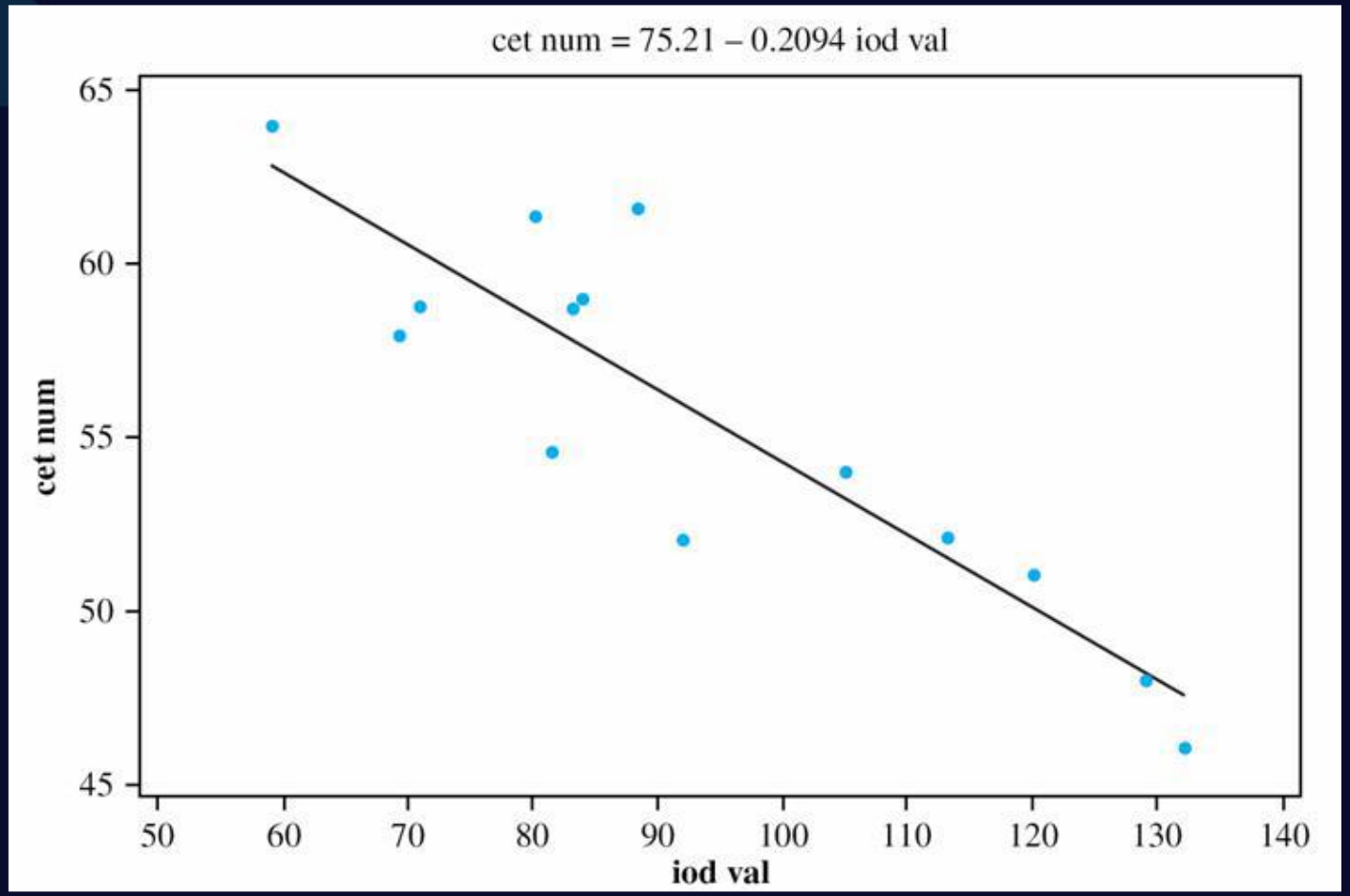
$$\Sigma x_i^2 = 128{,}913.93, \qquad \Sigma x_i y_i = 71{,}347.30,$$

$$S_{xx} = 128{,}913.93 - (1307.5)^2/14 = 6802.7693$$

$$S_{xy} = 71{,}347.30 - (1307.5)(779.2)/14 = -1424.41429$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Fitted Values

Fitted Values:

- The fitted (or predicted) values are obtained by substituting x1,…, xn into the equation of the estimated regression line:

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1, \hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_2, \ldots, \hat{y}_n = \hat{\beta}_0 + \hat{\beta}_1 x_n$$

- **Residuals:**

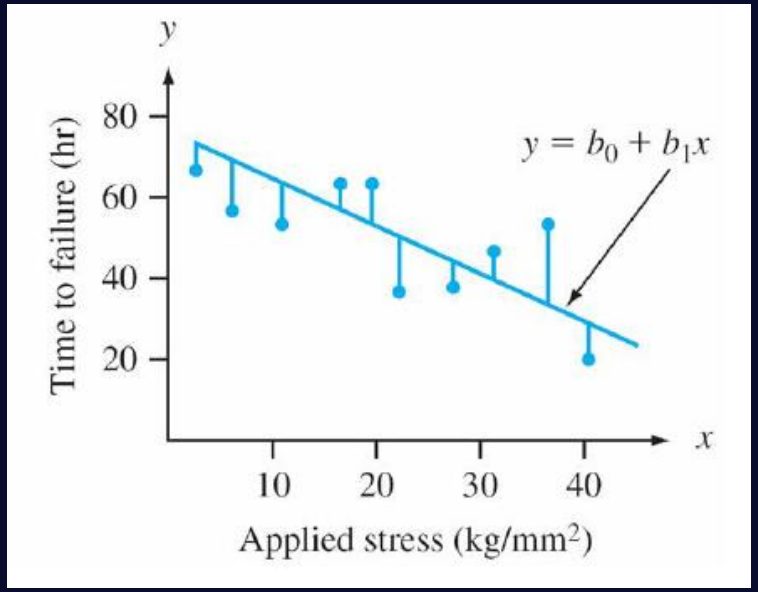
The differences between the observed and fitted *y* values. $y_1 - \hat{y}_1, y_2 - \hat{y}_2, \ldots, y_n - \hat{y}_n$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Sum of the residuals

- When the estimated regression line is obtained via the principle of least squares, the sum of the residuals will be zero:

$$\sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = n\overline{y} - n\hat{\beta}_0 - \hat{\beta}_1 n\overline{x} = n\hat{\beta}_0 - n\hat{\beta}_0 = 0$$




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example

- Suppose we have the following data on filtration rate (x) versus moisture content (y):

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 125.3 | 98.2 | 201.4 | 147.3 | 145.9 | 124.7 | 112.2 | 120.2 | 161.2 | 178.9 |
| $y$ | 77.9 | 76.8 | 81.5 | 79.8 | 78.2 | 78.3 | 77.5 | 77.0 | 80.1 | 80.2 |
| $x$ | 159.5 | 145.8 | 75.1 | 151.4 | 144.2 | 125.0 | 198.8 | 132.5 | 159.6 | 110.7 |
| $y$ | 79.9 | 79.0 | 76.7 | 78.2 | 79.5 | 78.1 | 81.5 | 77.0 | 79.0 | 78.6 |

$\Sigma x_i = 2817.9$, $\Sigma y_i = 1574.8$, $\Sigma x^2_i = 415{,}949.85$,

$\Sigma x_i\, y_i = 222{,}657.88$, and $\Sigma y^2_i = 124{,}039.58$,

$$S_{xx} = 18{,}921.8295,\ S_{xy} = 776.434$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example

- All predicted values (fits) and residuals appear in the accompanying table.

| Obs | Filtrate | Moistcon | Fit | Residual |
|---|---|---|---|---|
| 1 | 125.3 | 77.9 | 78.100 | −0.200 |
| 2 | 98.2 | 76.8 | 76.988 | −0.188 |
| 3 | 201.4 | 81.5 | 81.223 | 0.277 |
| 4 | 147.3 | 79.8 | 79.003 | 0.797 |
| 5 | 145.9 | 78.2 | 78.945 | −0.745 |
| 6 | 124.7 | 78.3 | 78.075 | 0.225 |
| 7 | 112.2 | 77.5 | 77.563 | −0.063 |
| 8 | 120.2 | 77.0 | 77.891 | −0.891 |
| 9 | 161.2 | 80.1 | 79.573 | 0.527 |
| 10 | 178.9 | 80.2 | 80.299 | −0.099 |
| 11 | 159.5 | 79.9 | 79.503 | 0.397 |
| 12 | 145.8 | 79.0 | 78.941 | 0.059 |
| 13 | 75.1 | 76.7 | 76.040 | 0.660 |
| 14 | 151.4 | 78.2 | 79.171 | −0.971 |
| 15 | 144.2 | 79.5 | 78.876 | 0.624 |
| 16 | 125.0 | 78.1 | 78.088 | 0.012 |
| 17 | 198.8 | 81.5 | 81.116 | 0.384 |
| 18 | 132.5 | 77.0 | 78.396 | −1.396 |
| 19 | 159.6 | 79.0 | 79.508 | −0.508 |
| 20 | 110.7 | 78.6 | 77.501 | 1.099 |

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Fitted Values

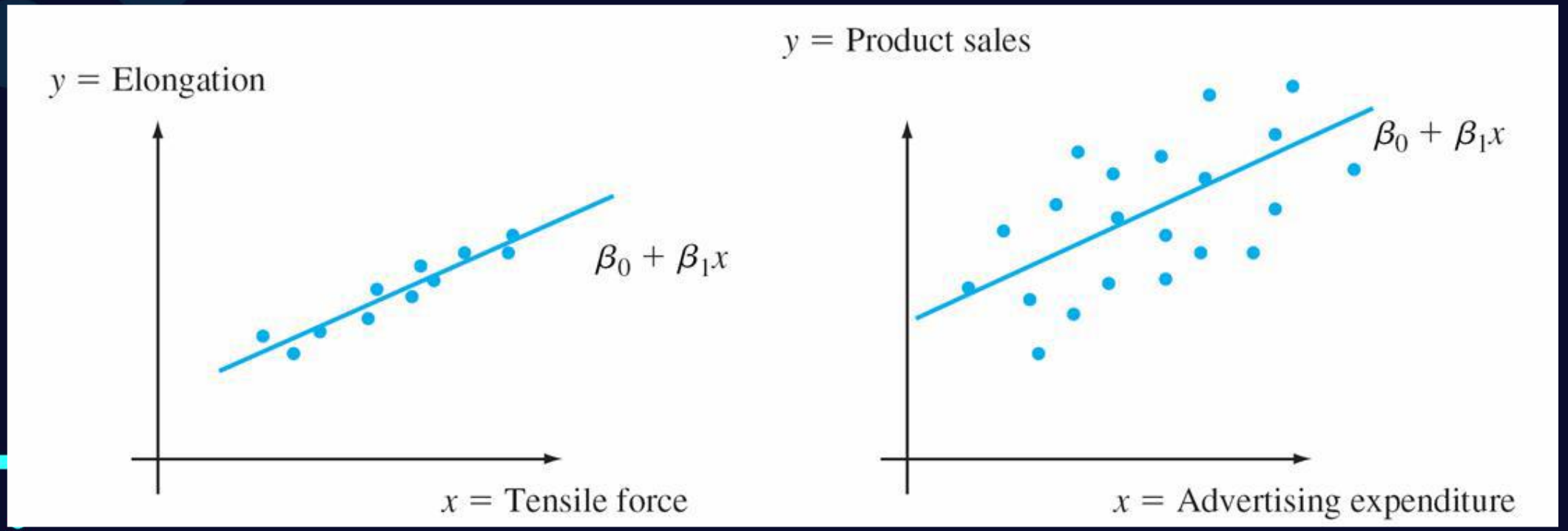
- We interpret the fitted value as the value of y that we would predict or expect when using the estimated regression line with x = xi; thus is the estimated true mean for that population when x = xi (based on the data).
- The residual is a positive number if the point lies above the line and a negative number if it lies below the line.
- The residual can be thought of as a measure of deviation and we can summarize the notation in the following way:

$$Y_i - \hat{Y}_i = \hat{\epsilon}_i$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Estimating $\sigma^2$ and and $\sigma$

- The parameter $\sigma^2$ determines the amount of spread about the true regression line.




המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Estimating $\sigma^2$ and and $\sigma$

- An estimate of $\sigma^2$ will be used in confidence interval (CI) formulas and hypothesis-testing procedures presented later.
- Many large deviations (residuals) suggest a large of $\sigma^2$ , whereas deviations all of which are small in magnitude suggest that $\sigma^2$ is small.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Estimating $\sigma^2$ and and $\sigma$

- The error sum of squares (equivalently, residual sum ofsquares), denoted by SSE, is

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

and the estimate of $\sigma^2$ which will be denoted as $s^2$ can be calculated using the below

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{\sum (y - \hat{y}_i)^2}{n-2} = \frac{1}{n-2} \sum_{i=1}^{n} \hat{e}_i^2$$

(Note that that the homoscedasticity assumption comes into play here)

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Estimating $\sigma^2$ and and $\sigma$

- The divisor n – 2 in the estimator is the number of degrees of freedom (df) associated with SSE.
- This is because to obtain $s^2$, the two parameters β0 and β1 must first be estimated, which results in a loss of 2 df.
- Replacing each yi in the formula for $s^2$ by the r.v. Yi gives the estimator for $s^2$.
- It can be shown that the r.v. $s^2$ is an unbiased estimator for $\sigma^2$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example

- The residuals for the filtration rate–moisture content data were calculated previously.
- The corresponding error sum of squares is

$$\text{SSE} = (-.200)^2 + (-.188)^2 + \cdots + (1.099)^2 = 7.968$$

- The estimate of $\sigma^2$ is then $s2 = 7.968/(20 - 2) = .4427$, and the *estimated standard deviation* is

$$\hat{\sigma} = s = \sqrt{.4427} = .665$$

- Roughly speaking, .665 is the *magnitude of a typical deviation from the estimated regression line*—some points are closer to the line than this and others are further away.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Estimating $\sigma^2$ and and $\sigma$

- Computation of SSE from the defining formula involves much tedious arithmetic, because both the predicted values and residuals must first be calculated.
- Use of the following shortcut formula does not require these quantities.

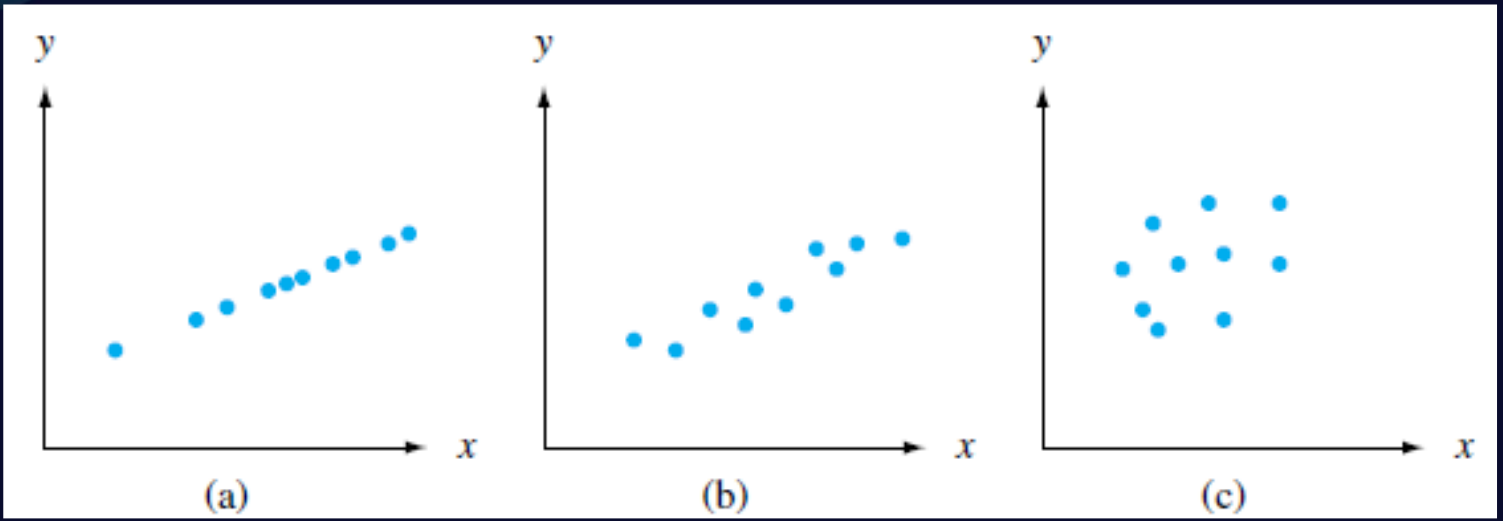
$$\text{SSE} = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$$

- This expression results from substituting into squaring the summand, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ carrying through the sum $\sum (y_i - \hat{y}_i)^2$ to the resulting three terms, and simplifying

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Coefficient of Determination

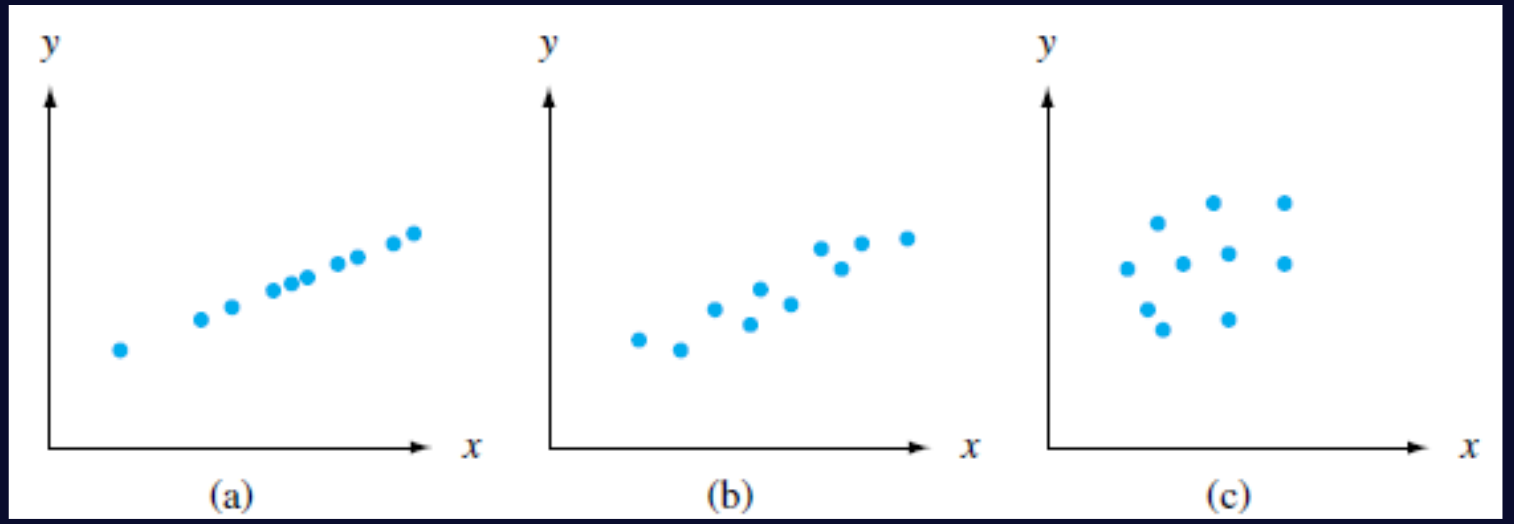
Using the linear model to explain y variation:
(a) data for which all variation is explained;;
(b) data for which most variation is explained;;
(c) data for which little variation is explained




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Coefficient of Determination

(a) data for which all variation is explained - The points in the first plot all fall exactly on a straight line. In this case, all (100%) of the sample variation in y can be attributed to the fact that x and y are linearly related in combination with variation in x.

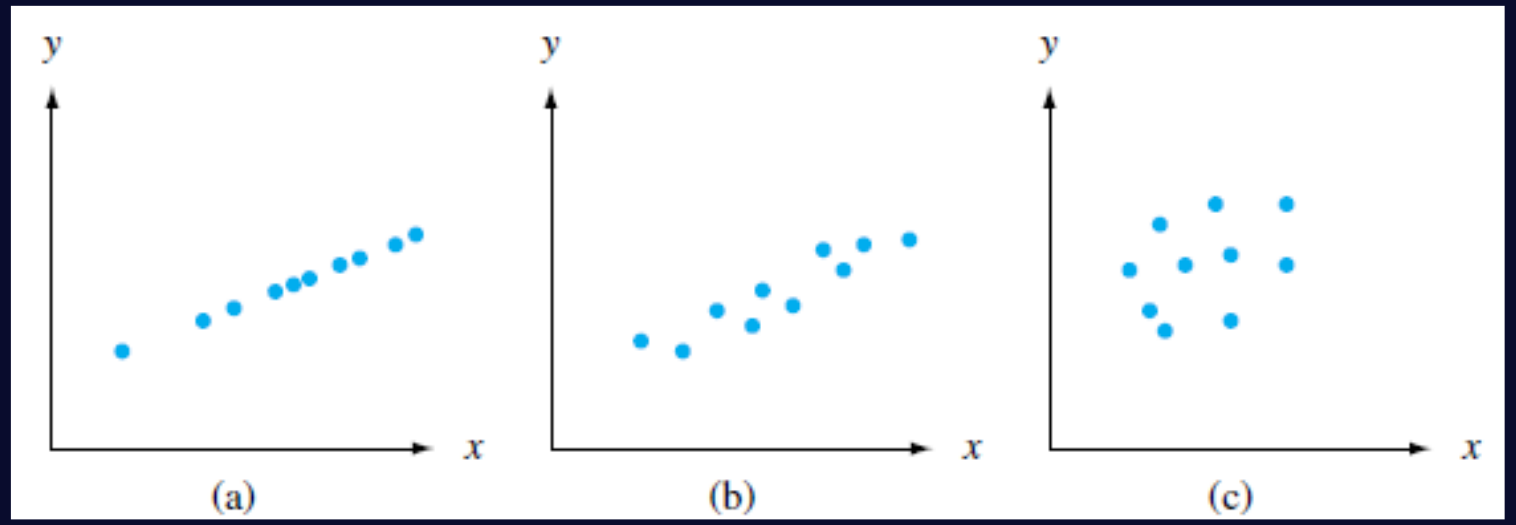



INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Coefficient of Determination

(b) data for which most variation is explained - The points in the second plot do not fall exactly on a line, but compared to overall y variability, the deviations from the least squares line are small.

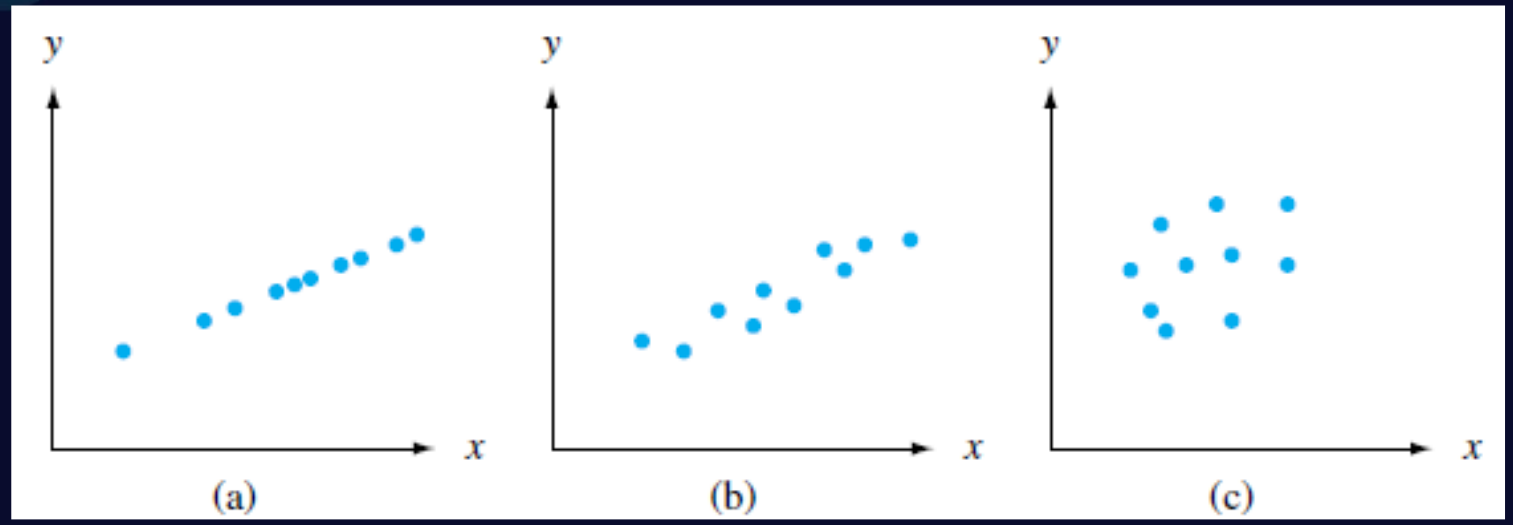
It is reasonable to conclude in this case that much of the observed y variation can be attributed to the approximate linear relationship between the variables postulated by the simple linear regression model.




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Coefficient of Determination

(c) data for which little variation is explained - When the scatter plot looks like that in the third plot, there is substantial variation about the least squares line relative to overall y variation, so the simple linear regression model fails to explain variation in y by relating y to x.




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Coefficient of Determination

- The error sum of squares SSE can be interpreted as a measure of how much variation in y is left unexplained by the model—that is, how much cannot be attributed to a linear relationship.
- In the first plot SSE = 0, and there is no unexplained variation, whereas unexplained variation is small for second, and large for the third plot.
- A quantitative measure of the total amount of variation in observed y values is given by the total sum of squares

$$\text{SST} = S_{yy} = \sum(y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2/n$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Coefficient of Determination

- Total sum of squares is the sum of squared deviations about the sample mean of the observed y values – when no predictors are taken into account.
- Thus, the same number y is subtracted from each $y_i$ in SST, whereas SSE involves subtracting each different predicted value from the corresponding observed $y_i$.
- The SST in some sense is as bad as SSE can get if there is no regression model (i.e., slope is 0) then

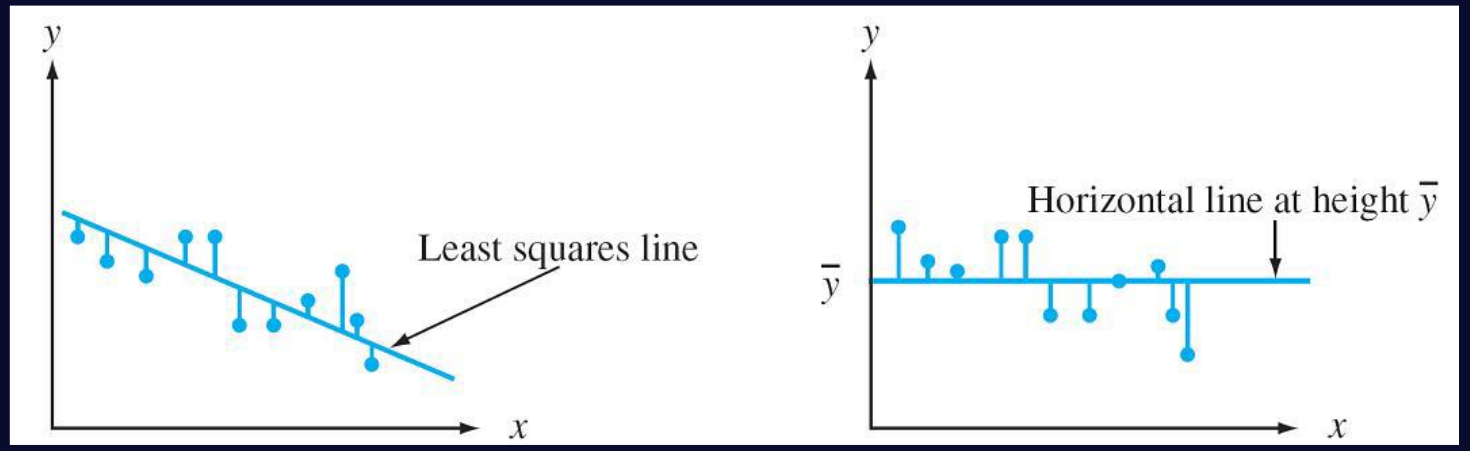
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad \Rightarrow \quad \hat{y} = \hat{\beta}_0 + \underbrace{\hat{\beta}_1}_{=0} \overline{x} = \hat{\beta}_0 = \overline{y}$$

Which motivates the definition of the SST.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Coefficient of Determination

- Just as SSE is the sum of squared deviations about the least squares line $y = \hat{\beta}_0 + \hat{\beta}_1 x$, SST is the sum of squared deviations about the horizontal line at height as pictured below:




Sums of squares illustrated: (a) SSE = sum of squared deviations about the least squares line;;
(b) SST = sum of squared deviations about the horizontal line

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Coefficient of Determination

- The sum of squared deviations about the least squares line is smaller than the sum of squared deviations about any other line, i.e. SSE < SST unless the horizontal line itself is the least squares line.
- The ratio SSE/SST is the proportion of total variation that cannot be explained by the simple linear regression model, and $r^2$ = 1 – SSE/SST (a number between 0 and 1) is the proportion of observed y variation explained by the model.
- Note that if SSE = 0 as in case (a), then $r^2$ = 1.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Coefficient of Determination

- The coefficient of determination, denoted by $r^2$ , is given by

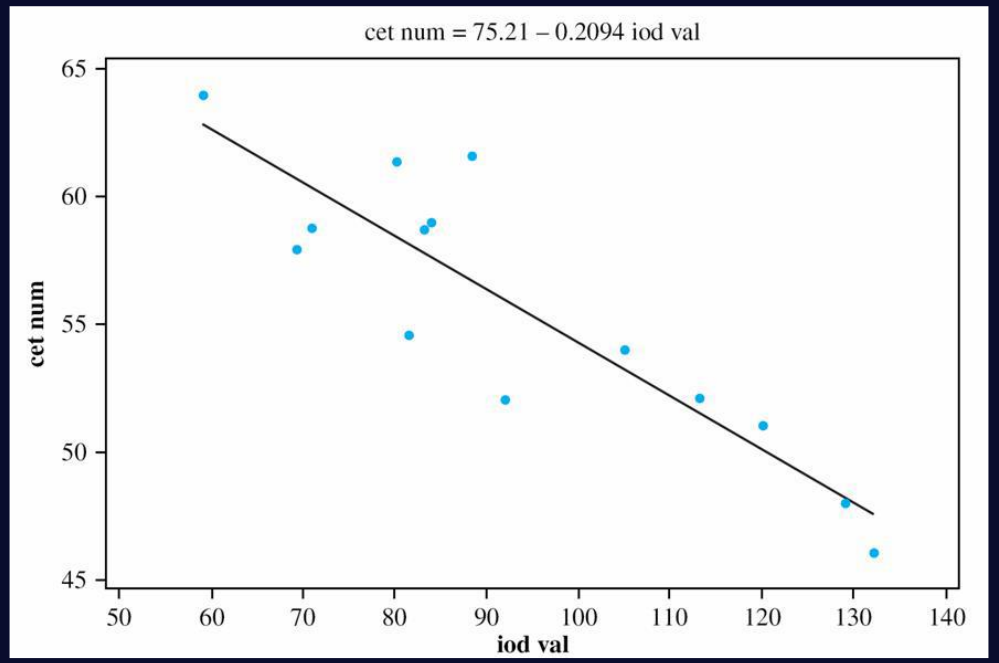
$$r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{SSE}{S_{yy}}$$

- It is interpreted as the proportion of observed y variation that can be explained by the simple linear regression model (attributed to an approximate linear relationship between y and x).
- The higher the value of r2, the more successful is the simple linear regression model in explaining y variation.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Coefficient of Determination

- The scatter plot of the iodine value–cetane number data in the previous example implies a reasonably high $r^2$ value.




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Coefficient of Determination

- The coefficient of determination for the previous example is
- Then

$$r^2 = 1 - \text{SSE/SST} = 1 - (78.920)/(377.174) = .791$$

- That is, 79.1% of the observed variation in cetane number is attributable to (can be explained by) the simple linear regression relationship between cetane number and iodine value.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Coefficient of Determination

- The coefficient of determination can be written in a slightly different way by introducing a third sum of squares—regression sum of squares, SSR—given by

$$SSR = \Sigma(\hat{y}_i - \bar{y})^2 = SST - SSE.$$

- *Regression sum of squares* is interpreted as the amount of total variation that *is* explained by the model. Then we have

$$r^2 = 1 - SSE/SST = (SST - SSE)/SST = SSR/SST$$

- the ratio of explained variation to total variation.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

- Before we move on to checking whether our models are significant. We will have a statistics refresher

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Statistics Refresher

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Illustrative Example: “Body Weight”

- The problem:

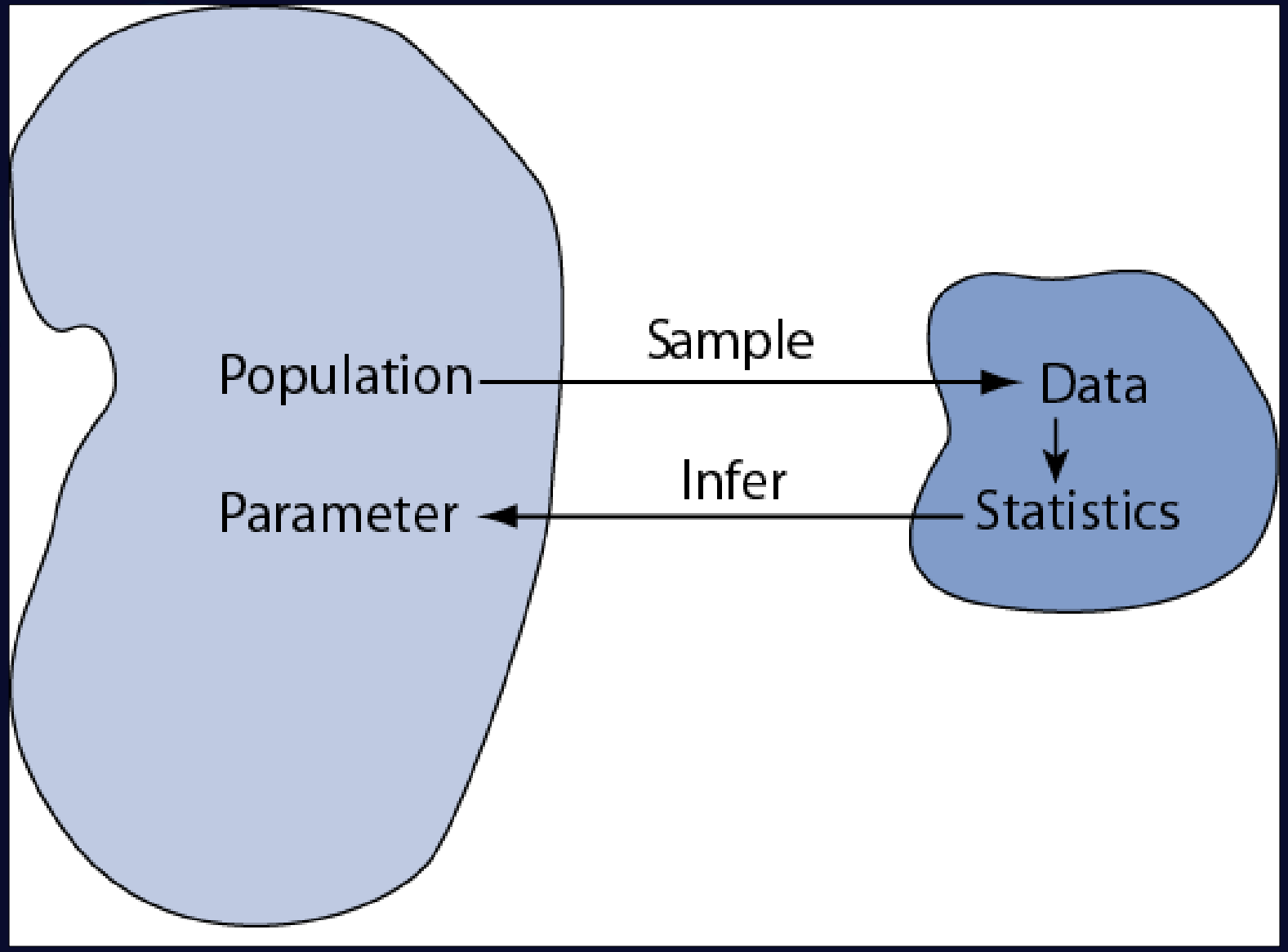
In the 1970s, 20–29 year old men in the U.S. had a mean $\mu$ body weight of 170 pounds. Standard deviation $\sigma$ was 40 pounds. We test whether mean body weight in the population is bigger now.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Definitions

- **Population** ≡ all possible values
- **Sample** ≡ a portion of the population
- **Statistical inference** ≡ generalizing from a sample to a population with calculated degree of certainty
- Two forms of statistical inference
  - **Hypothesis testing (Today)**
  - **Estimation (previous tutorial)**
- **Parameter** ≡ a characteristic of population, e.g., population mean $\mu$
- **Statistic** ≡ calculated from data in the sample, e.g., sample mean $\bar{X}$ $\bar{x}$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Distinctions Between Parameters and Statistics

| | Parameters | Statistics |
|---|---|---|
| Source | Population | Sample |
| Random | No | Yes |
| Calculated | No | Yes |

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Hypothesis Testing Steps

The procedure is broken into four steps:

1. Null and alternative hypotheses
2. Test statistic
3. P-value and interpretation
4. Significance Level

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Hypothesis Testing

- Confront two competing theories
  - **NULL** hypothesis $H_0$
    - any observed deviation from what we expect to see is due to chance variability
  - **ALTERNATIVE** hypothesis $H_a$
    - 'claim', or theory you wish to test

- The null hypothesis, $H_0$, is assumed true, until enough evidence goes against it
  - We then refute it and believe the alternative, $H_a$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Hypothesis Testing - cont.

- Null hypothesis $H_{0:}\ \mu = 170$ (“no difference”)
- Alternative hypothesis $H_{a:}\ \mu > 170$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Test Statistic

- A measure of how far the observed data is from what is expected assuming the null hypothesis $H_0$
  - Compute the value of a test statistic (TS) from the data
- The particular TS computed depends on the tested parameter
  - For example, to test the population mean, the TS is the sample mean (or standardized sample mean)
- The null hypothesis, $H_0$, is rejected if the TS falls in a user-specified rejection region

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Test Statistics - cont.

This is an example of a mean test when $\sigma$ is known.
Use this statistic to test the problem:

$$z_{stat} = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$$

where $\mu_0 \equiv$ population mean assuming $H_0$ is true

and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Illustrative Example: *z* statistic

- For the illustrative example, $\mu_0 = 170$
- We know $\sigma = 40$
- Take sample size of $n = 64$. Therefore

- If we found a sample mean of 173, then

$\sigma_{\bar{x}} = ?$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{64}} = 5$$

$z_{\text{stat}} = ?$

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{173 - 170}{5} = 0.60$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Illustrative Example: *z* statistic

If we found a sample mean of 185, then

$$z_{\text{stat}} = ?$$

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{185 - 170}{5} = 3.00$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# The Central Limit Theorem (CLT)

If $X_1, X_2, \ldots, X_n$ is a random sample of size $n$ taken from a population with mean $\mu$ and variance $\sigma^2$, and if $\overline{X}$ is the sample mean, the limiting form of the distribution of

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \qquad (3\text{-}42)$$

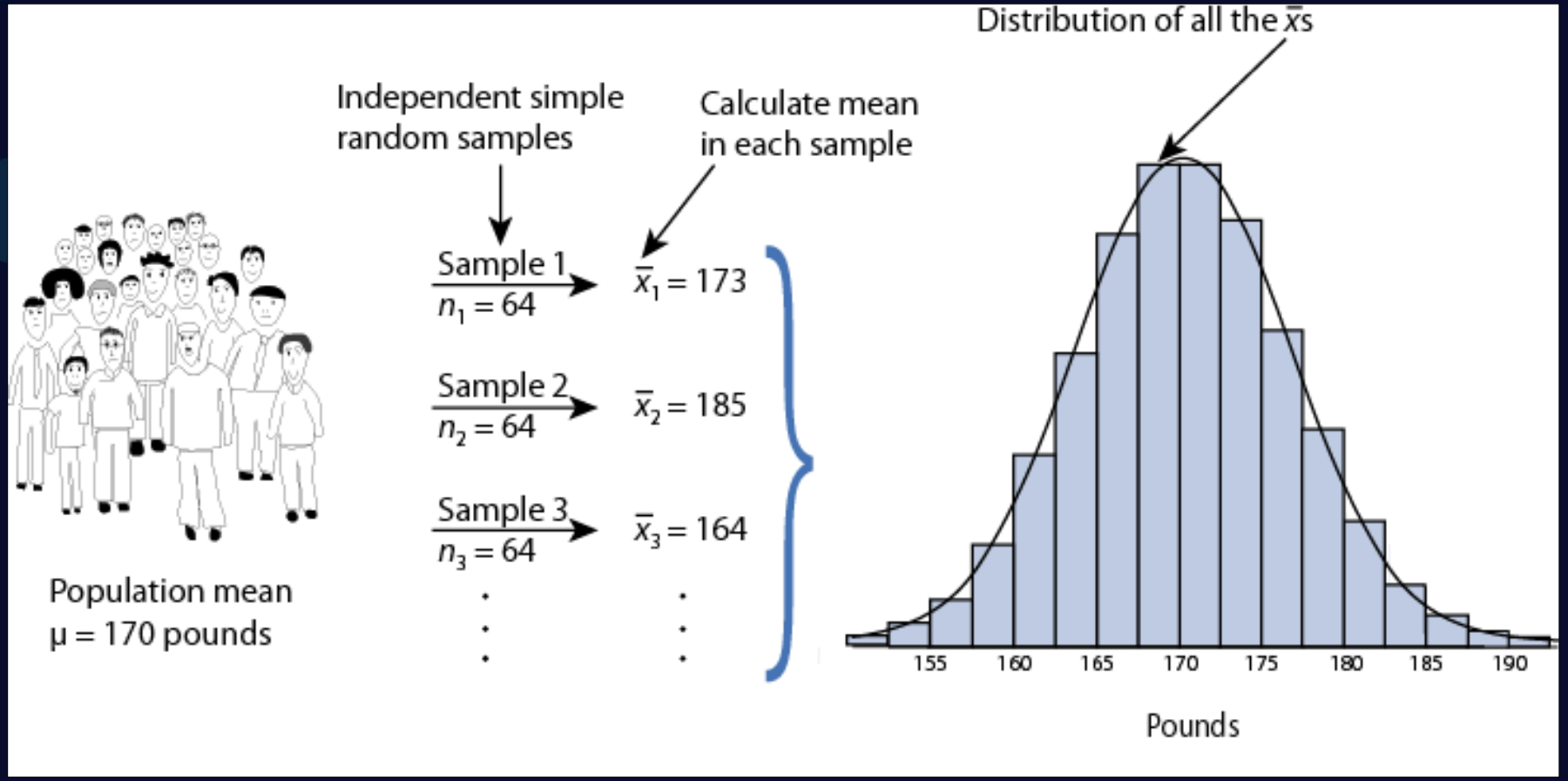
as $n \to \infty$, is the standard normal distribution.

- CLT Demo

http://onlinestatbook.com/stat_sim/sampling_dist/

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Reasoning Behind $z_{stat}$




Sampling distribution of $\bar{x}$ under $H_0$: $\bar{x} \sim N(170,5)$

$\mu = 170$ for $n = 64 \Rightarrow$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# *P*-value

- The *P*-value answers the question: What is the probability of the observed test statistic or one more extreme **when $H_0$ *is true*?**

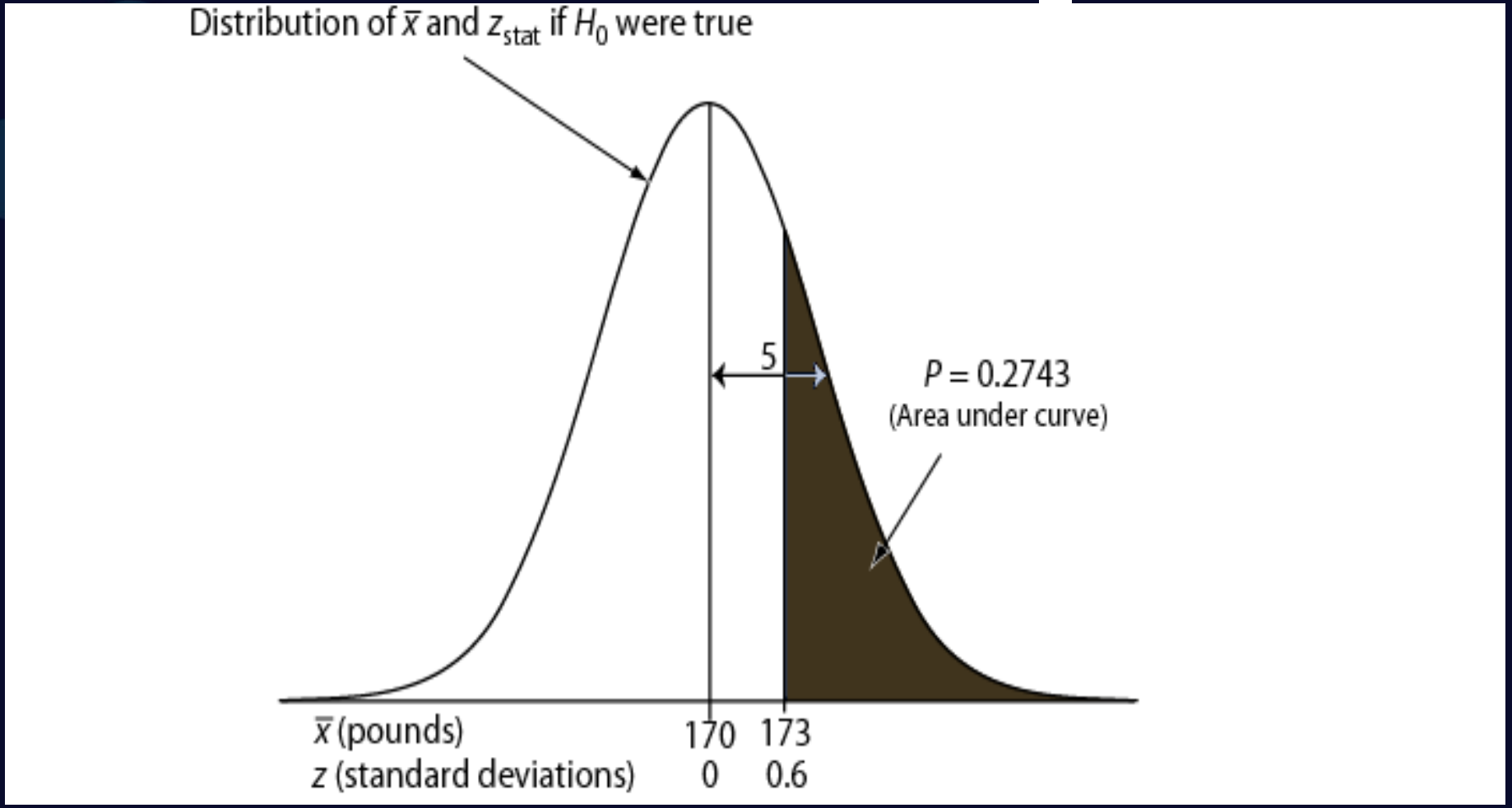
- This corresponds to the AUC in the tail of the Standard Normal distribution beyond the $z_{stat.}$

- Convert *z* statistics to *P*-value :

  For $H_a$: $\mu > \mu_0 \Rightarrow P = Pr(Z > z_{stat})$ = right-tail beyond $z_{stat}$

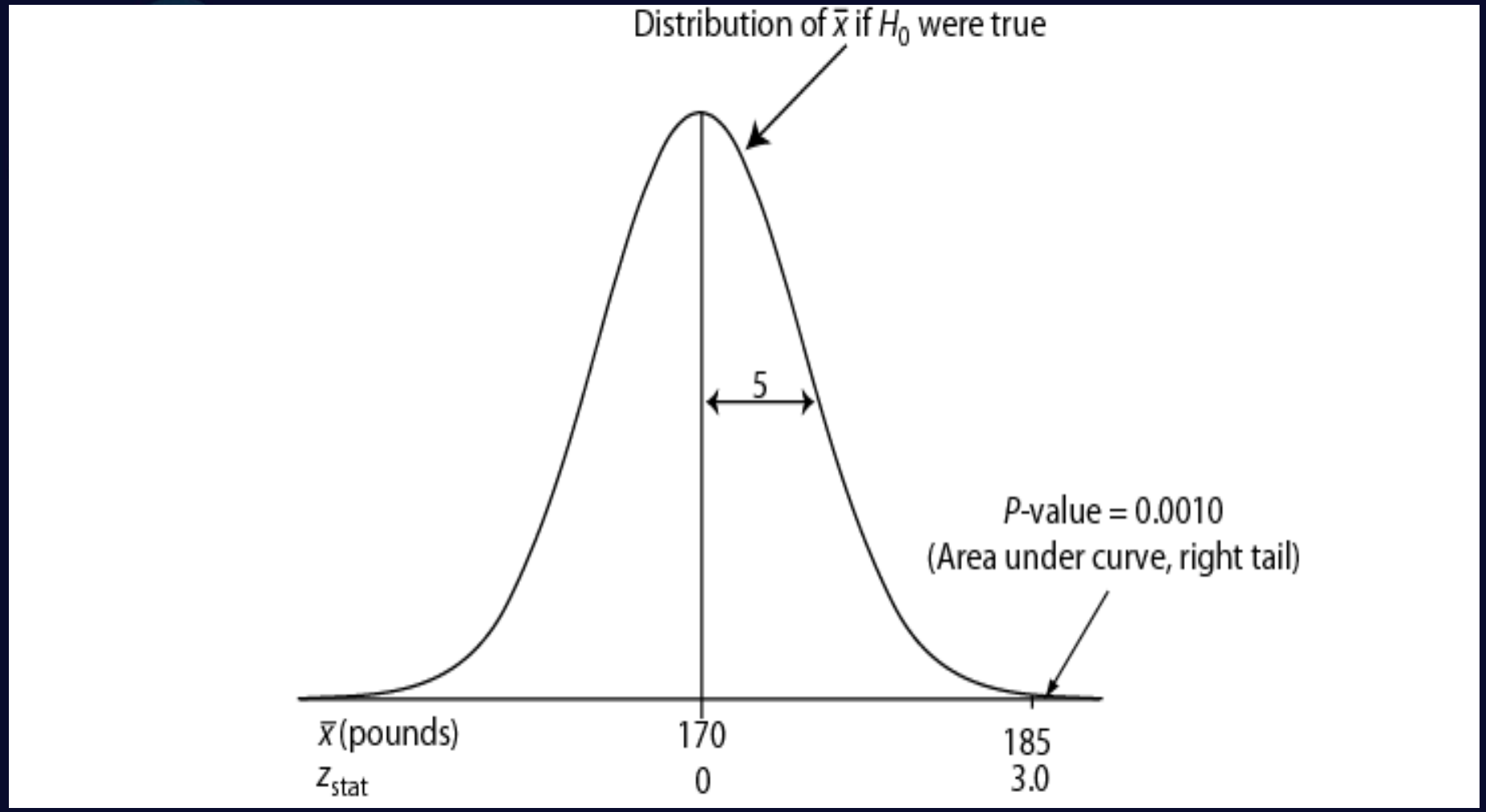
INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# P-value for $z_{stat}$ of 0.6




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# P-value for $z_{stat}$ of 3.0




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Interpretation

- *P*-value answer the question: What is the probability of the observed test statistic … **when $H_0$ *is true*?**

- Thus, smaller and smaller *P*-values provide stronger and stronger evidence against $H_0$

- Small *P*-value $\Rightarrow$ strong evidence

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Significance level (α) and *p-value*

- **Significance level (α)**

The degree of certainty required in order to *reject* the null hypothesis

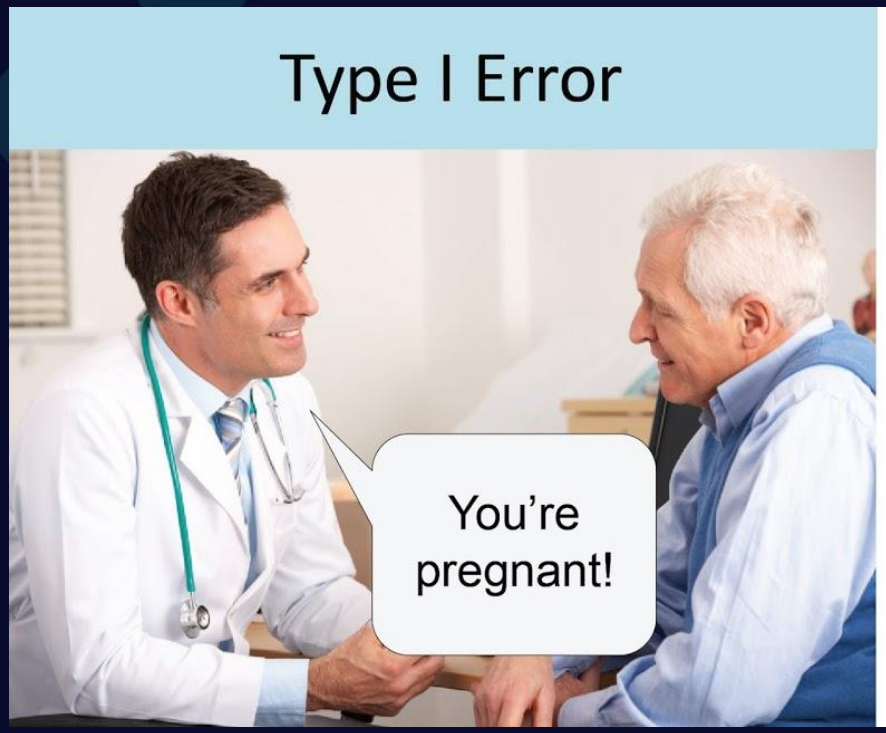
- A TS with a p-value less than some pre-specified false positive (or size) is said to be statistically significant at that level

| P-value | Wording |
|---|---|
| >0.05 | Not significant |
| 0.01 to 0.05 | Significant |
| 0.001 to 0.01 | Very significant |
| < 0.001 | Extremely significant |

commonly used p-values

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Error types




False Positive


False Negative

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Confusion Matrix

- **Type I error or false negative**
  - The chance of rejecting a NULL which is true is $\alpha$
- **Type II error or false positive**
  - The chance of not rejecting a NULL which is false is $\beta$

    - "rightfully" accept NULL
      - this is just $1 - \alpha$

    - "rightfully" reject the NULL
      - this is just $1 - \beta$, also called **power**

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Confusion Matrix – cont.

| Decision | | $H_0$ is in fact: False | $H_0$ is in fact: True |
|---|---|---|---|
| | Reject $H_0$ | ☺ Power, $1 - \beta$ | **Type I error, $\alpha$** |
| | Accept $H_0$ | **Type II error, $\beta$** | ☺ $1 - \alpha$ |

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

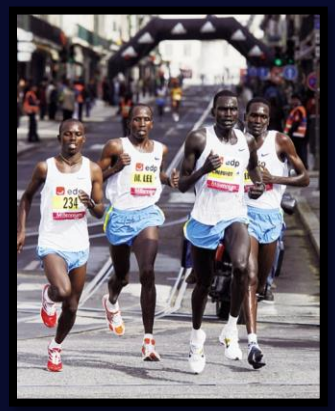
# T-test

- Z-distribution used in case that the variance is known. What happens in case the variance is unknown?

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Using Samples to Estimate Population Variability

- Acknowledge error
- Smaller samples, less spread



$$s = \sqrt{\frac{\Sigma(X - M)^2}{N - 1}}$$

*New!*

This correction will affect larger samples less so than it will affect smaller samples.

- *N = 65, N – 1 = 64 (change of 1.5%)*
- *N = 4, N – 1 = 3 (change of 25%)*

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Using Samples to Estimate Population Variability

- We have a new measure of standard deviation for a sample (as opposed to a population): *s*
  - *We need a new measure of standard error based on sample standard deviation:*

$$s_M = \frac{s}{\sqrt{N}}$$

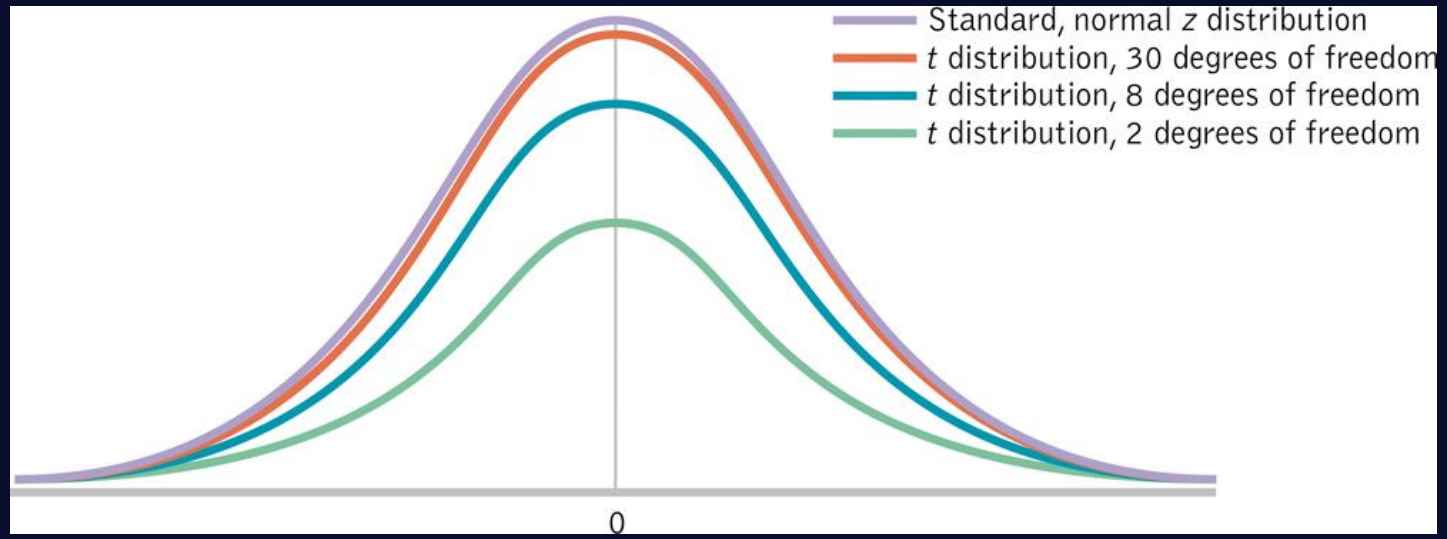
  - *Wait, what happened to "N-1"?*
  - *We already did that when we calculated s, don't correct again!*

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Student's $t$ Statistic

$$t = \frac{(M - \mu_M)}{s_M}$$

*Indicates the distance of a sample mean from a population mean in standard errors (like standard deviations)*




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Degrees of Freedom

- *Necessary when making estimates…*
- *The number of scores that are free to vary when estimating a population parameter from a sample*
  - $df = N - 1$ (for a Single-Sample *t* Test)

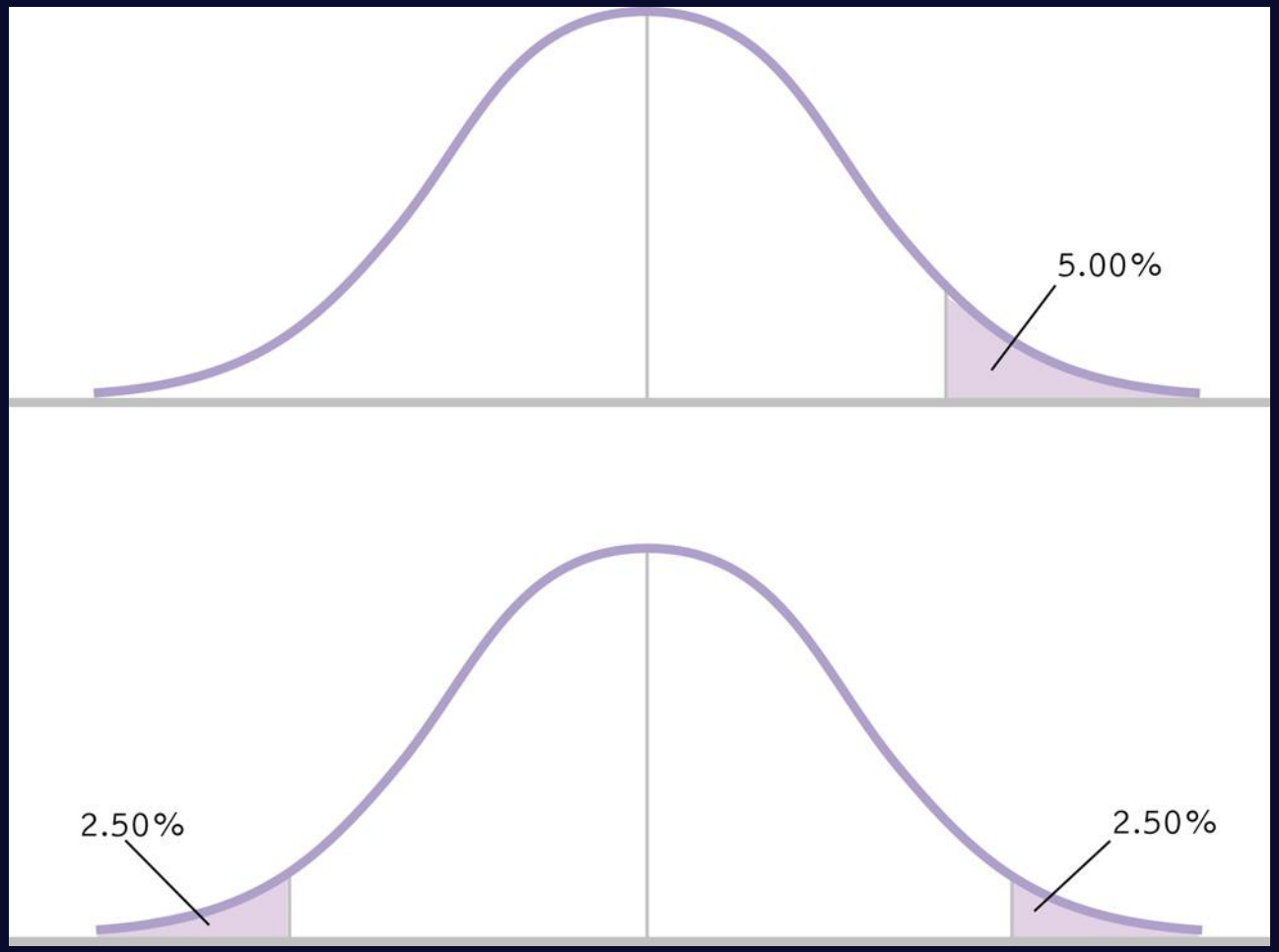
Example: I decide to ask 6 people how often they floss their teeth and record their average = 2 (times per week)

- Eventual goal: Estimate population parameters (population variability).
- How many scores are free to vary and can still produce an average of 2?

| | | |
|---|---|---|
| 3 | Free | 2 |
| 5 | Free | 1 |
| 1 | Free | 0 |
| 0 | Free | 0 |
| 2 | Free | 0 |
| **1** | **LOCKED** | **9** |
| **Average = 2** | | **Average = 2** |

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# One Tailed vs. Two Tailed Tests




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example: Attendance in Therapy Sessions

- Our Counseling center on campus is concerned that most students requiring therapy do not take advantage of their services.  Right now students attend only 4.6 sessions in a given year! Administrators are considering having patients sign a contract stating they will attend at least 10 sessions in an academic year.
- Question: Does signing the contract increase participation/attendance?

- We had 5 patients sign the contract and we counted the number of times they attended therapy sessions

| Number of Attended Therapy Sessions |
|---|
| 6 |
| 6 |
| 12 |
| 7 |
| 8 |


INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example: Attendance in Therapy Sessions

- Populations:
  - Pop 1: All clients who sign contract
  - Pop 2: All clients who do not sign contract
- Distribution:
  - One Sample mean: Distribution of means
- Test & Assumptions: Population mean is known but not standard deviation → single-sample *t* test
  1. Data are interval
  2. Probably not random selection
  3. Sample size of 5 is less than 30, therefore distribution might not be normal

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example: Attendance in Therapy Sessions

$H_0$: Clients who sign the contract will attend the same number of sessions as those who do not sign the contract.

$H_1$: Clients who sign the contract will attend a different number of sessions than those who do not sign the contract.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example: Attendance in Therapy Sessions

Determine characteristics of comparison distribution (distribution of sample means)

- Population: $\mu_M = \mu = 4.6$times
- Sample: $M =$ 7.8 times, $s =$ 2.490, $s_M =$ 1.114

| # of Sessions ($X$) | $X-M$ | $(X-M)^2$ |
|---|---|---|
| 6 | -1.8 | 3.24 |
| 6 | -1.8 | 3.24 |
| 12 | -4.2 | 17.64 |
| 7 | -0.8 | 0.64 |
| 8 | 0.2 | 0.04 |
| $M_X = 7.8$ | | $SS_X = 24.8$ |

$$s = \sqrt{\frac{\Sigma(X-M)^2}{N-1}} = \sqrt{\frac{24.8}{5-1}} = 2.490$$

$$s_M = \frac{s}{\sqrt{N}} = \frac{2.490}{\sqrt{5}} = 1.114$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example: Attendance in Therapy Sessions

$\mu_M = 4.6$, $s_M = 1.114$, $M = 7.8$, $N = 5$, $df = 4$

Determine critical value (cutoffs)

- In Behavioral Sciences, we use $p = .05$ (5%)
- Our hypothesis ("Clients who sign the contract will attend a different number of sessions than those who do not sign the contract.") is *nondirectional* so our hypothesis test is *two-tailed*.

*df* = 4 ⟶

Significance level = α

| Degrees of Freedom | .005 (1-tail) .01 (2-tails) | .01 (1-tail) .02 (2-tails) | .025 (1-tail) .05 (2-tails) | .05 (1-tail) .10 (2-tails) | .10 (1-tail) .20 (2-tails) | .25 (1-tail) .50 (2-tails) |
|---|---|---|---|---|---|---|
| 1 | 63.657 | 31.821 | 12.706 | 6.314 | 3.078 | 1.000 |
| 2 | 9.925 | 6.965 | 4.303 | 2.920 | 1.886 | .816 |
| 3 | 5.841 | 4.541 | 3.182 | 2.353 | 1.638 | .765 |
| 4 | 4.604 | 3.747 | 2.776 | 2.132 | 1.533 | .741 |
| 5 | 4.032 | 3.365 | 2.571 | 2.015 | 1.476 | .727 |
| 6 | 3.707 | 3.143 | 2.447 | 1.943 | 1.440 | .718 |
| 7 | 3.500 | 2.998 | 2.365 | 1.895 | 1.415 | .711 |

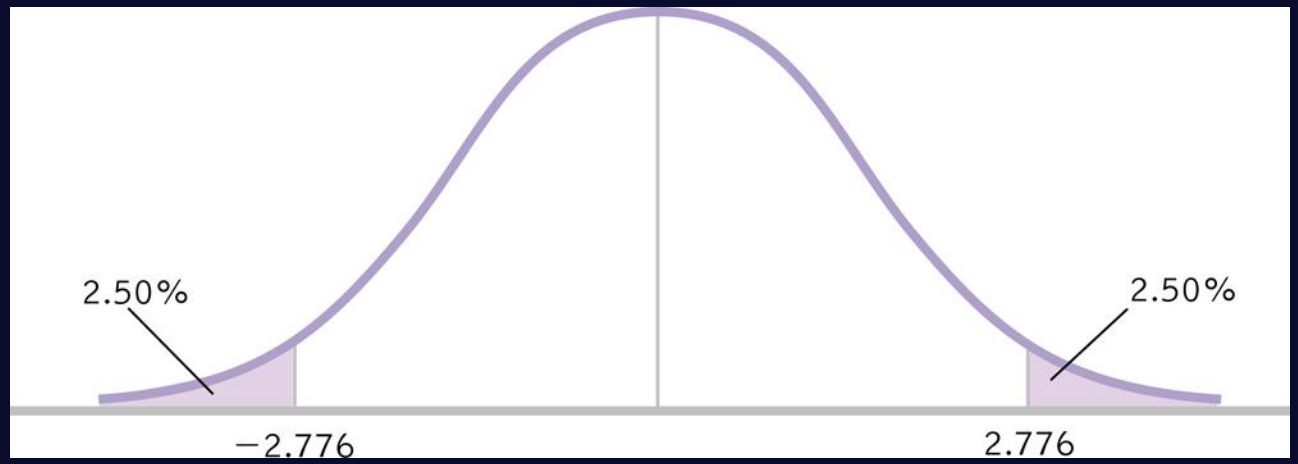
INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example: Attendance in Therapy Sessions

$\mu_M = 4.6, s_M = 1.114, M = 7.8, N = 5, df = 4$

4. Determine critical value (cutoffs)

$t_{crit} = \pm 2.76$



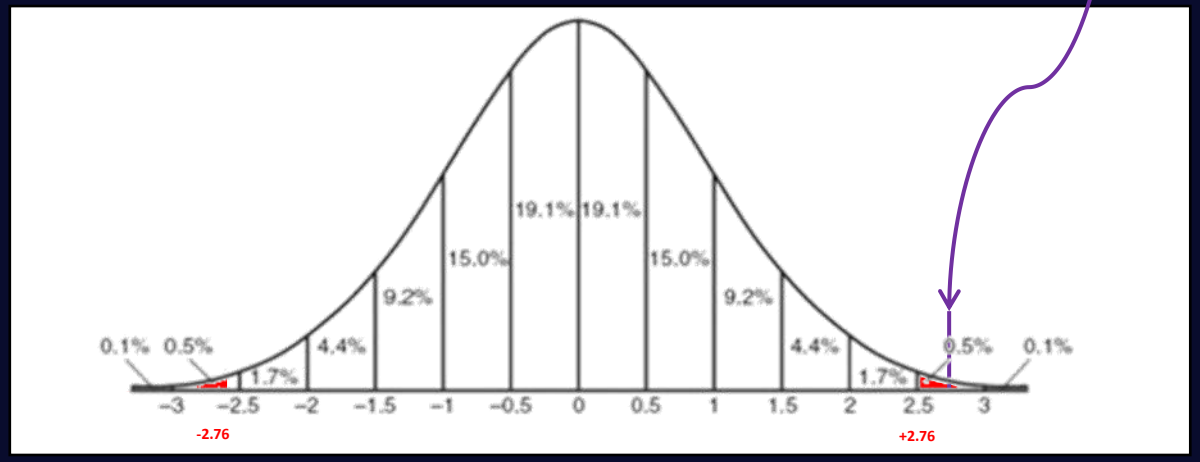

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example: Attendance in Therapy Sessions

$$\mu_M = 4.6,\ s_M = 1.114,\ M = 7.8,\ N = 5,\ df = 4$$

Calculate the test statistic

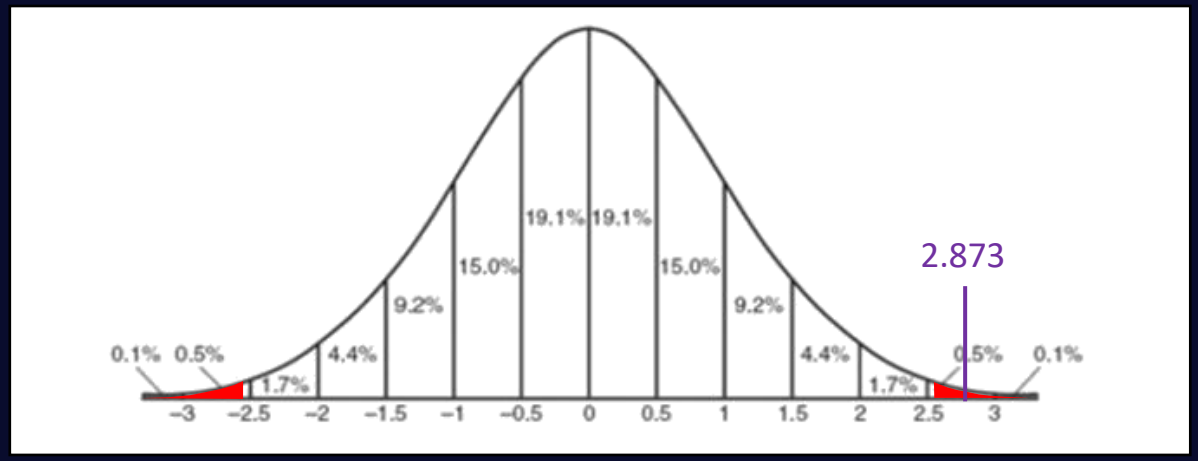
$$t = \frac{(M - \mu_M)}{s_M} = \frac{(7.8 - 4.6)}{1.114} = 2.873$$




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example: Attendance in Therapy Sessions

$\mu_M = 4.6$, $s_M = 1.114$, $M = 7.8$, $N = 5$, $df = 4$




Make a decision

$t = 2.873 > t_{crit} = \pm 2.776$, *reject the null hypothesis*

Clients who sign a contract will attend more sessions than those who do not sign a contract, $t(4) = 2.87$, $p < .05$.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

- Going back to linear regression…

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Coefficient of determination-common mistakes

- The coefficient of determination and the correlation coefficient  have to be the most often misused and misunderstood measures in the field of statistics.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

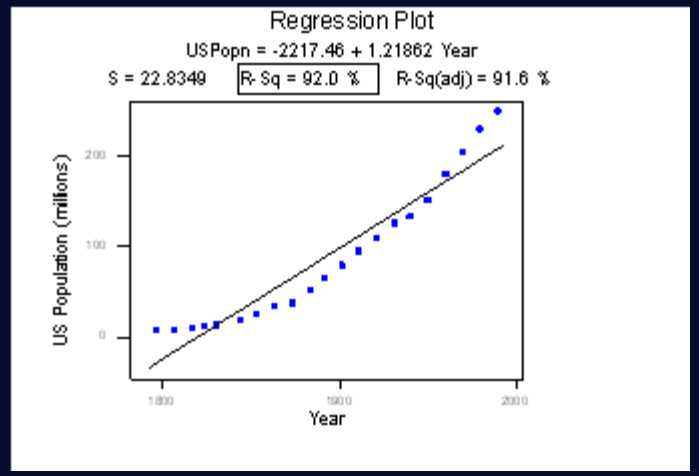
# Coefficient of determination-common mistakes

- The coefficient of determination and the correlation coefficient *r* quantify the strength of a *linear* relationship. It is possible that $r^2=0$ and $r = 0$, suggesting there is no linear relation between *x* and *y*, and yet a perfect non-linear relation exists.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Coefficient of determination-common mistakes

- A large value should not be interpreted as meaning that the estimated regression line fits the data well. Another function might better describe the trend in the data.




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

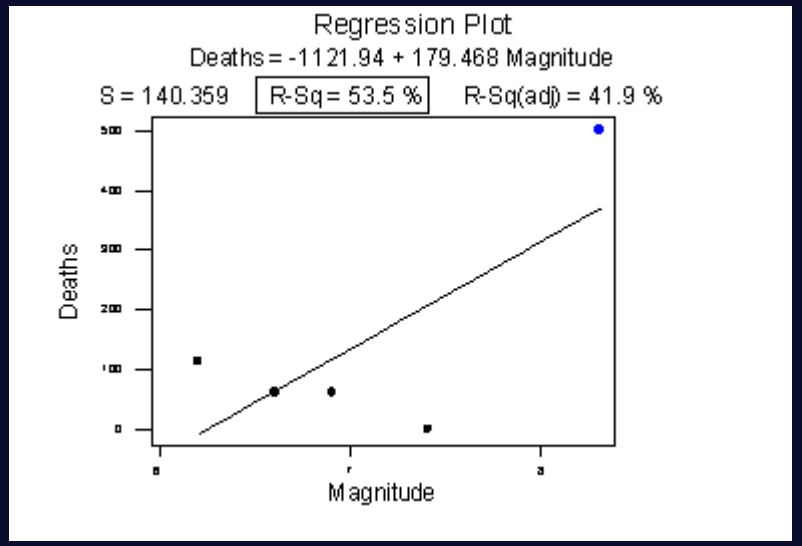
# Coefficient of determination-common mistakes

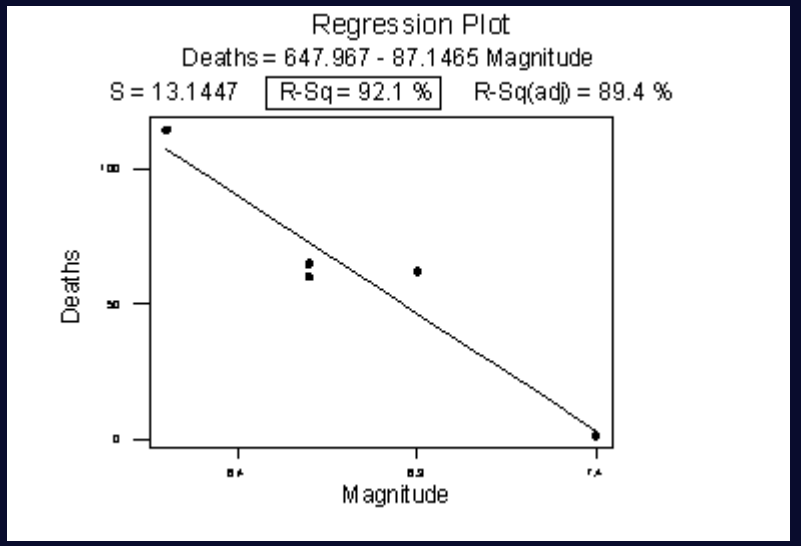
- The correlation of 0.959 and the value of 92.0% suggest a strong linear relationship between year and U.S. population.
- Indeed, only 8% of the variation in U.S. population is left to explain after taking into account the year in a linear way!
- The plot suggests, though, that a curve would describe the relationship even better. That is, the large value of 92.0% should not be interpreted as meaning that the estimated regression line fits the data well.
- It doesn't tell us that we could still do better.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Coefficient of determination-common mistakes

- The coefficient of determination and the correlation coefficient  can both be greatly affected by just one data point (or a few data points).




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Coefficient of determination-common mistakes

- Correlation does not imply causation.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Coefficient of determination-common mistakes

- A "statistically significant" value does not imply that the slope β is meaningfully different from 0.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Coefficient of determination-common mistakes

- A large value does not necessarily mean that a useful prediction of the response, or estimation of the mean response, can be made. It is still possible to get prediction intervals or confidence intervals that are too wide to be useful.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Model Evaluation

- Both the correlation coefficient and the coefficient of determination summarize the strength of a linear relationship in samples only.
- If we obtained a different sample, we would obtain different correlations, different values, and therefore potentially different conclusions.
- As always, we want to draw conclusions about populations, not just samples. To do so, we either have to conduct a hypothesis test or calculate a confidence interval.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Model Evaluation

- We are often interested in learning about the population intercept and the population slope.
- As you know, confidence intervals and hypothesis tests are two related, but different, ways of learning about the values of population parameters.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example

- Let’s investigate the relationship between skin cancer mortality and state latitude.
- y is the mortality rate (number of deaths per 10 million people) of white males due to malignant skin melanoma from 1950-1959.
- The predictor variable x is the latitude (degrees North) at the center of each of 49 states in the United States.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example

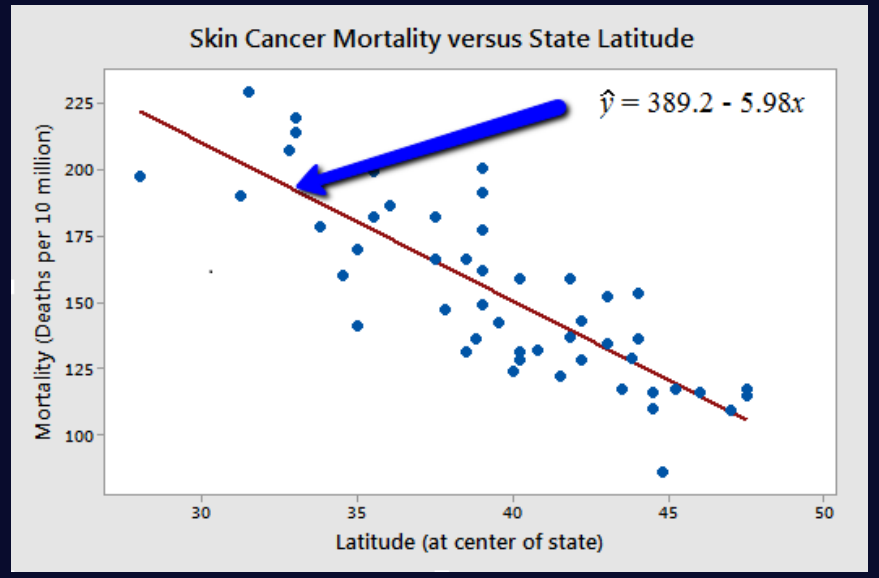
- Let’s investigate the relationship between skin cancer mortality and state latitude.
- y is the mortality rate (number of deaths per 10 million people) of white males due to malignant skin melanoma from 1950-1959.
- The predictor variable x is the latitude (degrees North) at the center of each of 49 states in the United States.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

| # | State | Latitude | Mortality |
|---|---|---|---|
| 1 | Alabama | 33.0 | 219 |
| 2 | Arizona | 34.5 | 160 |
| 3 | Arkansas | 35.0 | 170 |
| 4 | California | 37.5 | 182 |
| 5 | Colorado | 39.0 | 149 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 49 | Wyoming | 43.0 | 134 |

המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Is there a relationship between state latitude and skin cancer mortality?




Since the estimated slope of the line, b1, is -5.98, not 0, there is a relationship between state latitude and skin cancer mortality in the sample of 49 data points.

But, we want to know if there is a relationship between the population of all of the latitudes and skin cancer mortality rates. That is, we want to know if the population slope is unlikely to be 0.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# (1-$\alpha$)100% t-interval for the slope parameter

**Confidence Interval for $\beta_1$**

The formula for the confidence interval for $\beta_1$, in words, is:

**Sample estimate ± (t-multiplier × standard error)**

and, in notation, is:

$$b_1 \pm t_{(\alpha/2,n-2)} \times \left( \frac{\sqrt{MSE}}{\sqrt{\sum(x_i - \bar{x})^2}} \right)$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# (1-$\alpha$)100% t-interval for the slope parameter

- The resulting confidence interval not only gives us a range of values that is likely to contain the true unknown value.
- It also allows us to answer the research question "is the predictor x linearly related to y?" If the confidence interval for contains 0, then we conclude that there is no evidence of a linear relationship between the predictor x and the response y in the population.
- On the other hand, if the confidence interval for does not contain 0, then we conclude that there is evidence of a linear relationship between the predictor x and y in the population.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# $\alpha$ -level hypothesis test for the slope parameter

- We follow standard hypothesis test procedures in conducting a hypothesis test for the slope .

- **Null hypothesis** $H_0: \beta_1 =$ some number $\beta$
- **Alternative hypothesis** $H_A: \beta_1 \neq$ some number $\beta$

Most often we are interested in testing against 0.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

**Theorem**

The assumptions of the simple linear regression model imply that the standardized variable

$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$$

has a *t* distribution with $n - 2$ df (since $\sigma \approx s$).

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example

- Going back to the previous example
- By default, the test statistic is calculated assuming the user wants to test that the slope is 0. Dividing the estimated coefficient -5.9776 by the estimated standard error 0.5984 We will get a t statistic of -9.99
- Calculating the probability that a *t*-random variable with *n*-2 = 47 degrees of freedom would be larger than 9.99. We will get that the *P*-value is less than 0.001.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example

- Because the *P*-value is small (less than 0.001), we can reject the null hypothesis and conclude that does not equal 0.
- In other words, there is sufficient evidence, at the significance level of 0.05 level, to conclude that there is a linear relationship in the population between skin cancer mortality and latitude.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example

- Calculate a 95% confidence interval -

$$t_{(0.025,47)} = 2.0117$$

Then, the 95% confidence interval for the slope is

$$-5.9776 \pm 2.0117(0.5984)$$ (-7.2, -4.8)

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Example

- We can be 95% confident that the population slope is between -7.2 and -4.8. That is, we can be 95% confident that for every additional one-degree increase in latitude, the mean skin cancer mortality rate decreases between 4.8 and 7.2 deaths per 10 million people.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Factors affecting the width of a confidence interval

- We want our confidence intervals to be as narrow as possible. If we know what factors affect the length of a confidence interval for the slope, we can control them to ensure that we obtain a narrow interval.
- The factors can be easily determined by studying the formula for the confidence interval:

$$b_1 \pm t_{\alpha/2,n-2} \times \left( \frac{\sqrt{MSE}}{\sqrt{\sum (x_i - \bar{x})^2}} \right)$$

$$\text{Width} = 2 \times t_{\alpha/2,n-2} \times \left( \frac{\sqrt{MSE}}{\sqrt{\sum (x_i - \bar{x})^2}} \right)$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Factors affecting the width of a confidence interval

- As the confidence level decreases, the width of the interval decreases. Therefore, if we decrease our confidence level, we decrease the width of our interval.
- Clearly, we don't want to decrease the confidence level too much. Typically, confidence levels are never set below 90%.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Factors affecting the width of a confidence interval

- As MSE decreases, the width of the interval decreases. The value of MSE depends on only two factors — how much y vary naturally around the estimated regression line, and how well your regression function (line) fits the data. Clearly, we can't control the first factor all that much other than to ensure that you are not adding any unnecessary error in your measurement process.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Factors affecting the width of a confidence interval

- The more spread out the predictor x values, the narrower the interval. The quantity in the denominator summarizes the spread of the predictor x values. The more spread out the predictor values, the larger the denominator, and hence the narrower the interval. Therefore, we can decrease the width our interval by ensuring that our predictor values are sufficiently spread out.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Factors affecting the width of a confidence interval

- As the sample size increases, the width of the interval decreases. The sample size plays a role in two ways. First, recall that the t-multiplier depends on the sample size through n-2. Therefore, as the sample size increases, the t-multiplier decreases, the length of the interval decreases.
- Second, the denominator also depends on n. The larger the sample size, the more terms you add to this sum, the larger the denominator, the narrower the interval. Therefore, in general, we can ensure that your interval is narrow by having a large enough sample.

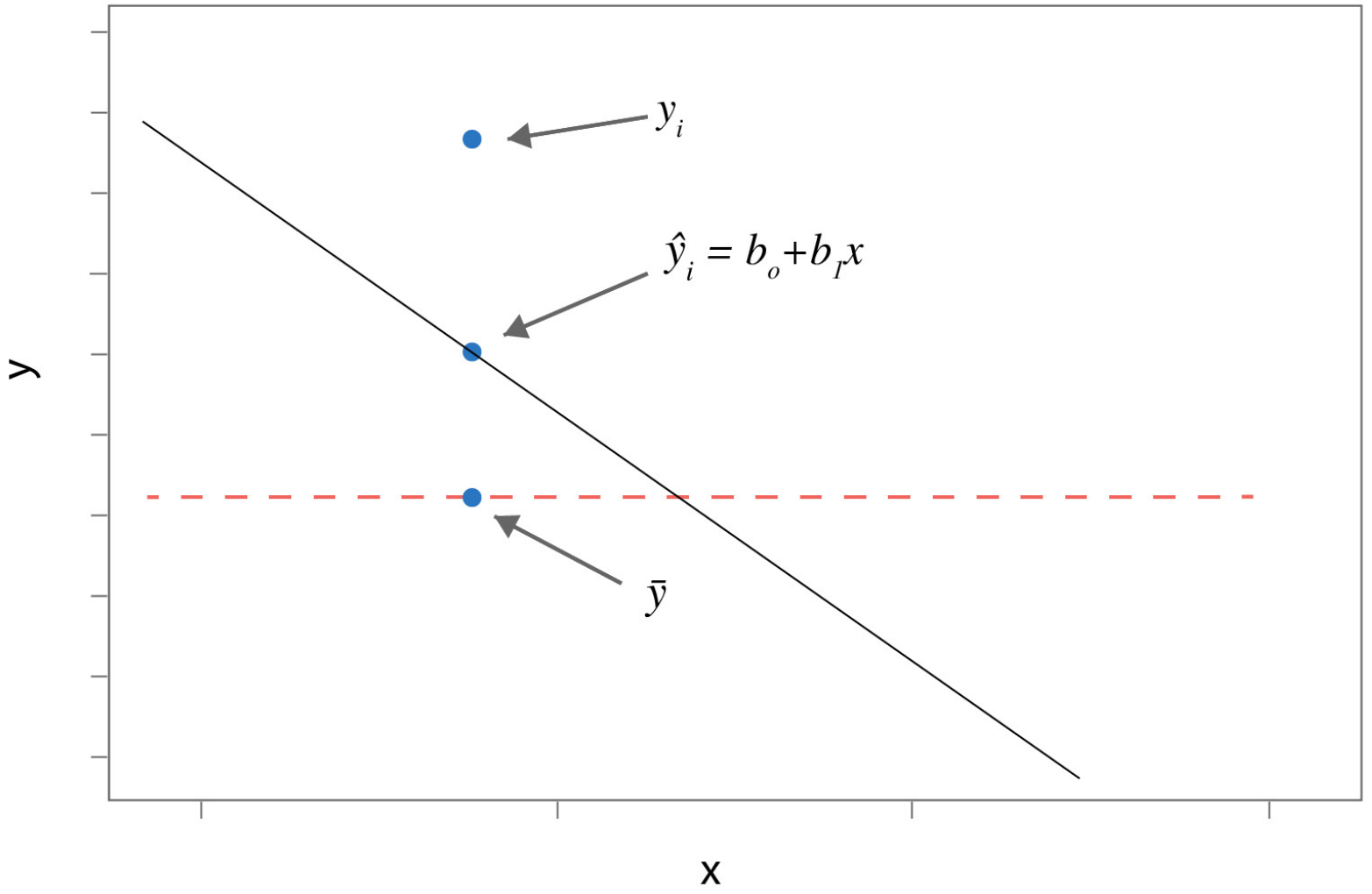
INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Analysis of Variance (ANOVA)

- Break down the total variation in y SST (total sum of squares) into two components:
- a component that is "due to" the change in x ("regression sum of squares (SSR)“)
- a component that is just due to random error ("error sum of squares (SSE)")

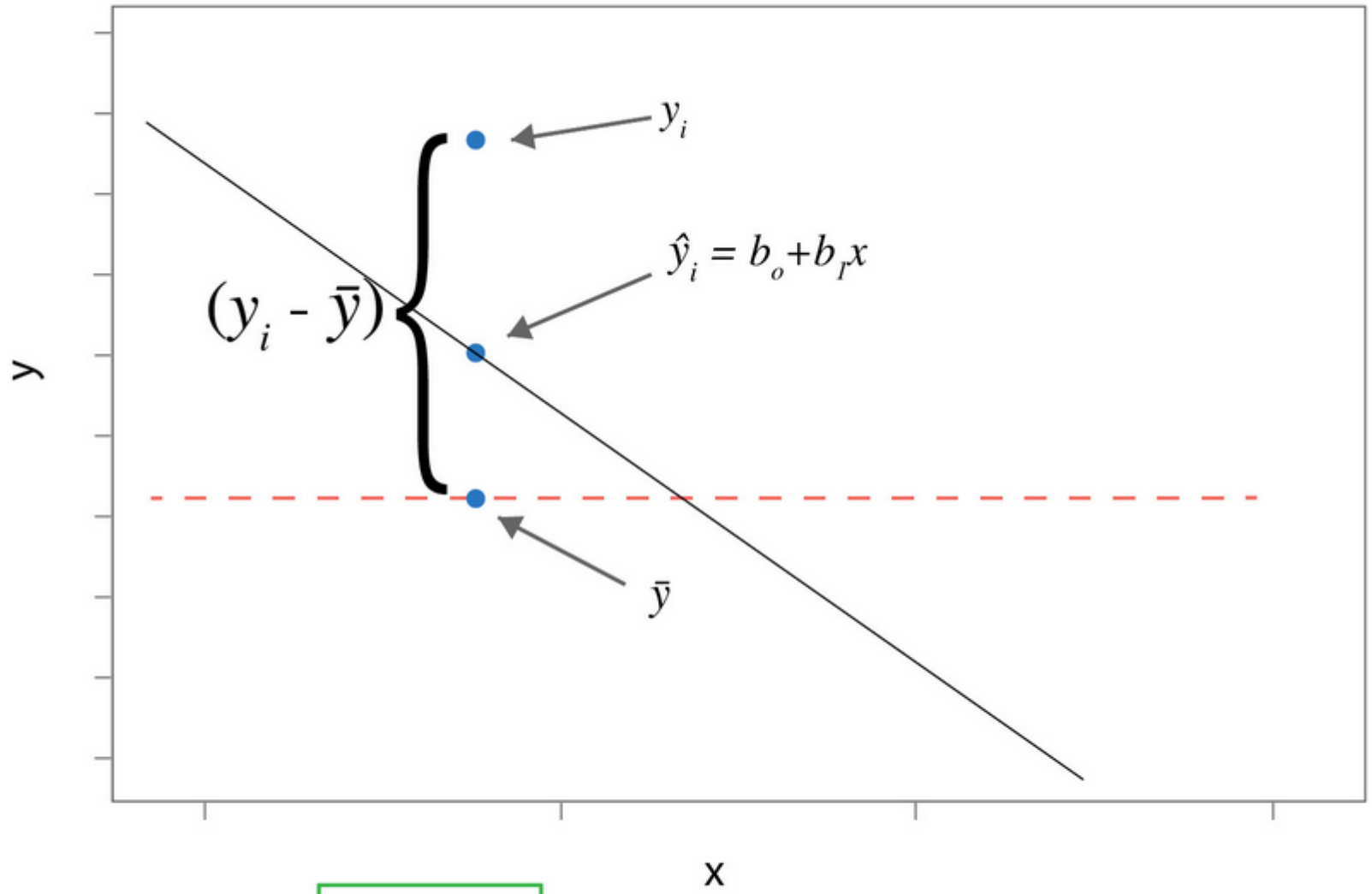
- If the regression sum of squares is a "large" component of the total sum of squares, it suggests that there is a linear association between the predictor x and the response y.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות



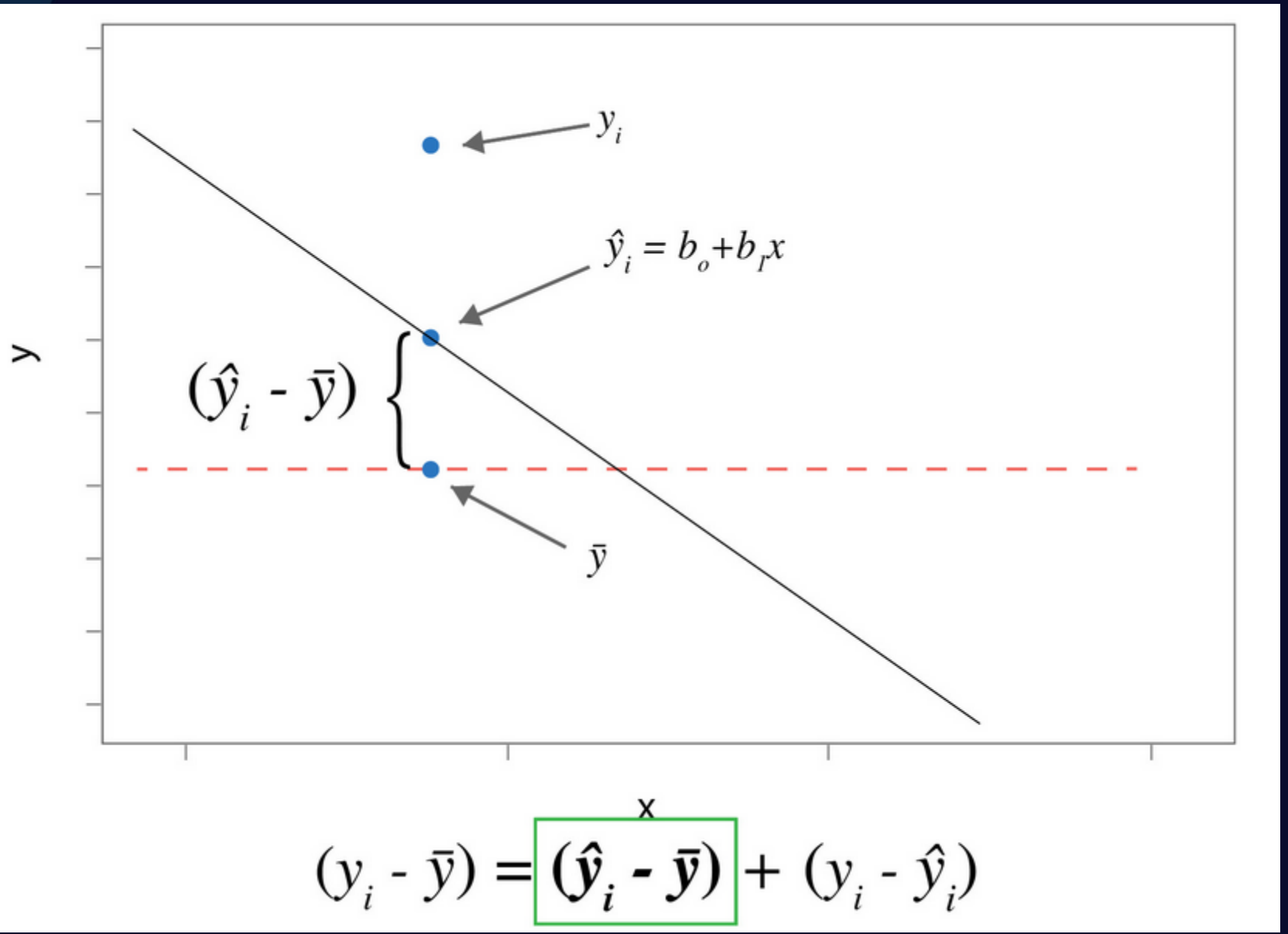

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות



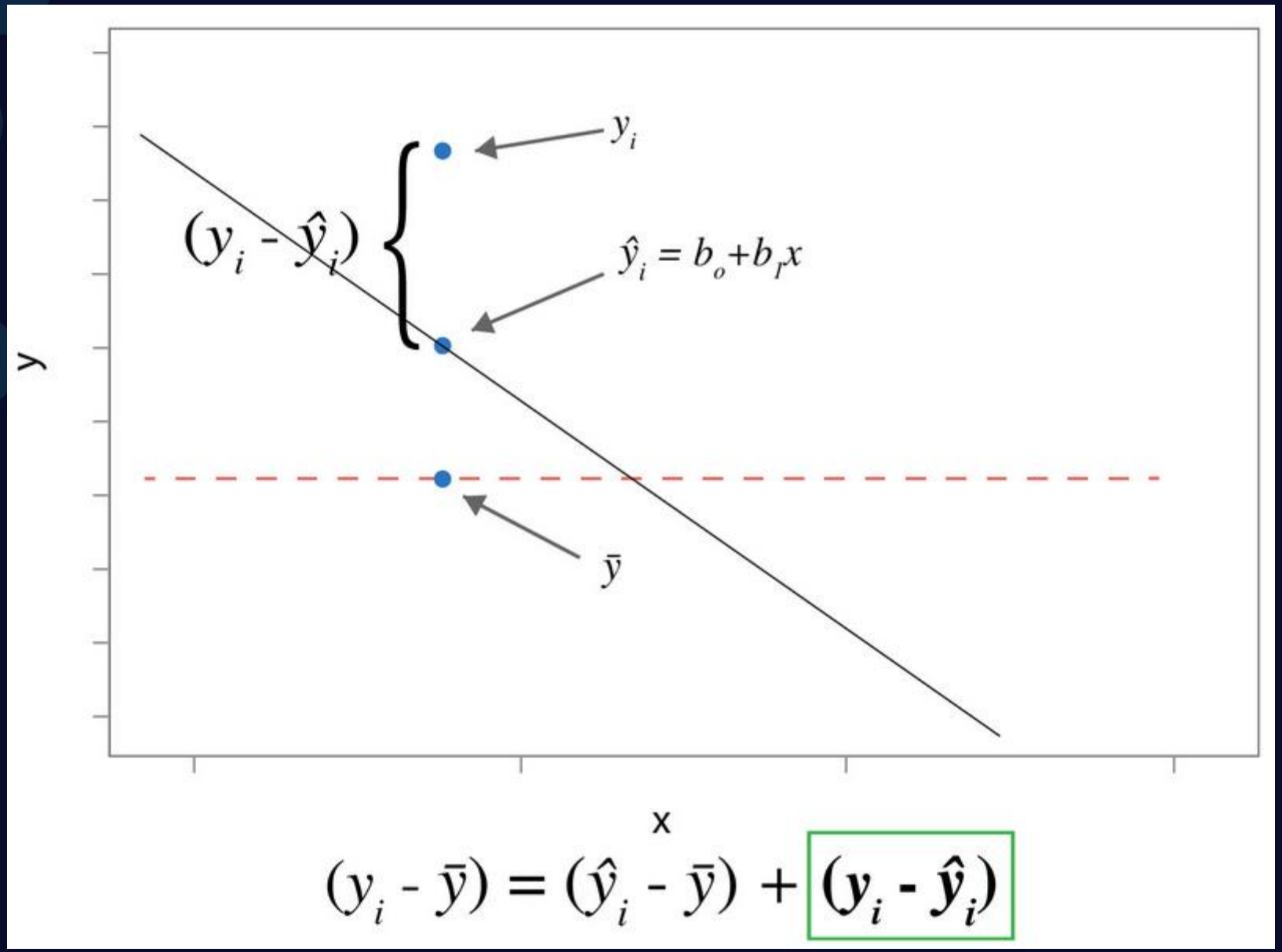

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות




$$(y_i - \bar{y}) = \boxed{(\hat{y}_i - \bar{y})} + (y_i - \hat{y}_i)$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות




$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + \boxed{(y_i - \hat{y}_i)}$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Analysis of Variance (ANOVA)

- The decomposition also valid for the sum of the squared distances

$$\underbrace{\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}_{\substack{\text{SST}\\ \text{Total Sum of Squares}}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\substack{\text{SSR}\\ \text{Regression of Sums}}} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y})^2}_{\substack{\text{SSE}\\ \text{Error Sum of Squares}}}$$

$$\text{SST} = \text{SSR} + \text{SSE}$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Analysis of Variance (ANOVA)

- The degrees of freedom associated with each of these sums of squares follow a similar decomposition.
  - You might recognize SST as being the numerator of the sample variance. Recall that the denominator of the sample variance is n-1. Therefore, n-1 is the degrees of freedom associated with SST.
  - Recall that the mean square error MSE is obtained by dividing SSE by n-2. Therefore, n-2 is the degrees of freedom associated with SSE.
- Then, we obtain the following breakdown of the degrees of freedom:

$$\underset{\substack{\text{degrees of freedom}\\ \text{associated with SST}}}{(n-1)} = \underset{\substack{\text{degrees of freedom}\\ \text{associated with SSR}}}{(1)} + \underset{\substack{\text{degrees of freedom}\\ \text{associated with SSE}}}{(n-2)}$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Analysis of Variance (ANOVA)

| Source of Variation | DF | SS | MS | F |
|---|---|---|---|---|
| **Regression** | 1 | $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ | $MSR = \frac{SSR}{1}$ | $F^* = \frac{MSR}{MSE}$ |
| **Residual error** | *n*-2 | $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ | $MSE = \frac{SSE}{n-2}$ | |
| **Total** | *n*-1 | $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ | | |

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Analysis of Variance (ANOVA)

- The ratio F* follows F-distribution, with 1 numerator degree of freedom and n-2 denominator degrees of freedom. For this reason, it is often referred to as the analysis of variance F-test.

- The null hypothesis is $H_0: \beta_1 = 0$
- The alternative hypothesis is $H_A: \beta_1 \neq 0$
- The test statistic is $F^* = \frac{MSR}{MSE}$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Assumptions

- Linear Function: The mean of Y at each value of the predictor  x  is a linear function of x
- Independent: The errors are independent.
- Normally Distributed: The errors at each value of the predictor x, are normally distributed.
- Equal variances: The errors at each value of the predictor x have equal variance.

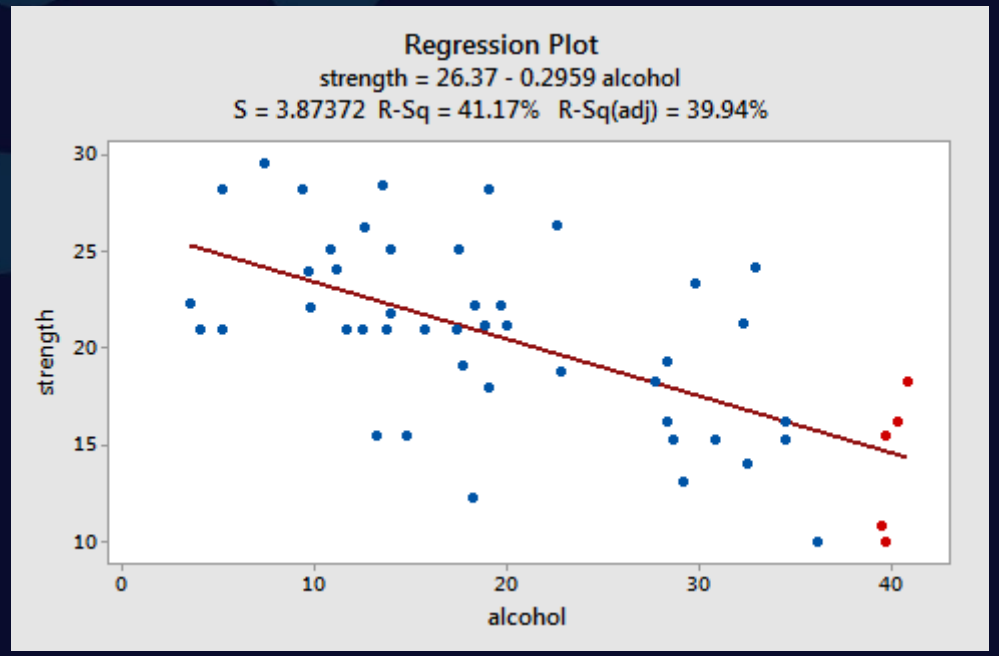
INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

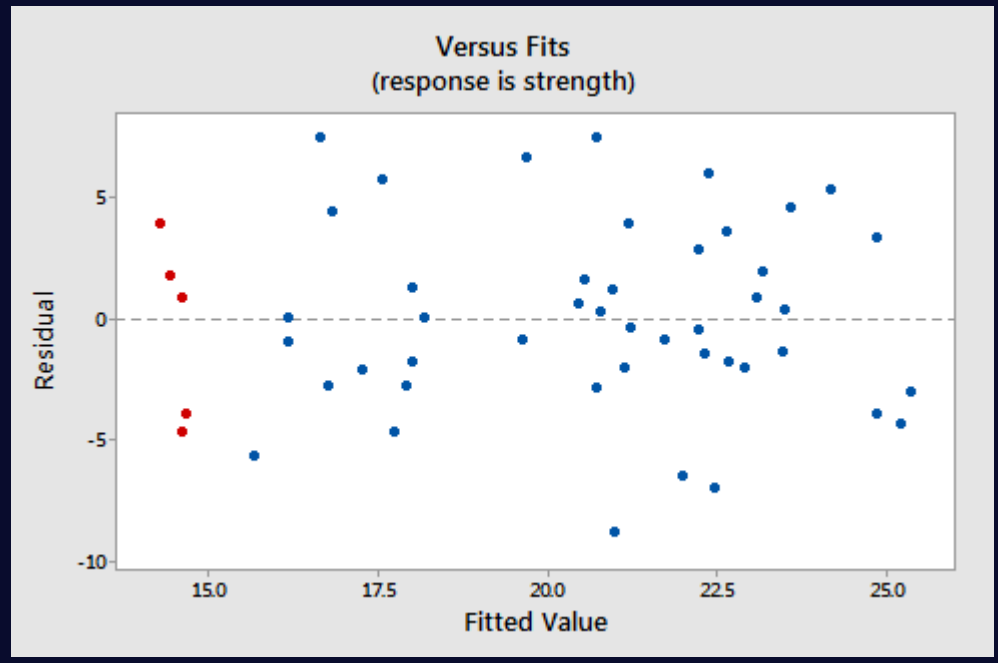
# Residuals Plot

- When conducting a residual analysis, a "residuals versus fits plot" is the most frequently created plot. It is a scatter plot of residuals on the y axis and fitted values (estimated responses) on the x axis. The plot is used to detect non-linearity, unequal error variances, and outliers.

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Residuals Plot




INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

# Residuals Plot

- The residuals "bounce randomly" around the residual = 0 line. This suggests that the assumption that the relationship is linear is reasonable.
- The residuals roughly form a "horizontal band" around the residual = 0 line. This suggests that the variances of the error terms are equal.
- No one residual "stands out" from the basic random pattern of residuals. This suggests that there are no outliers

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות