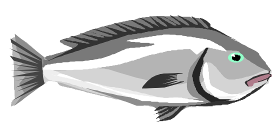
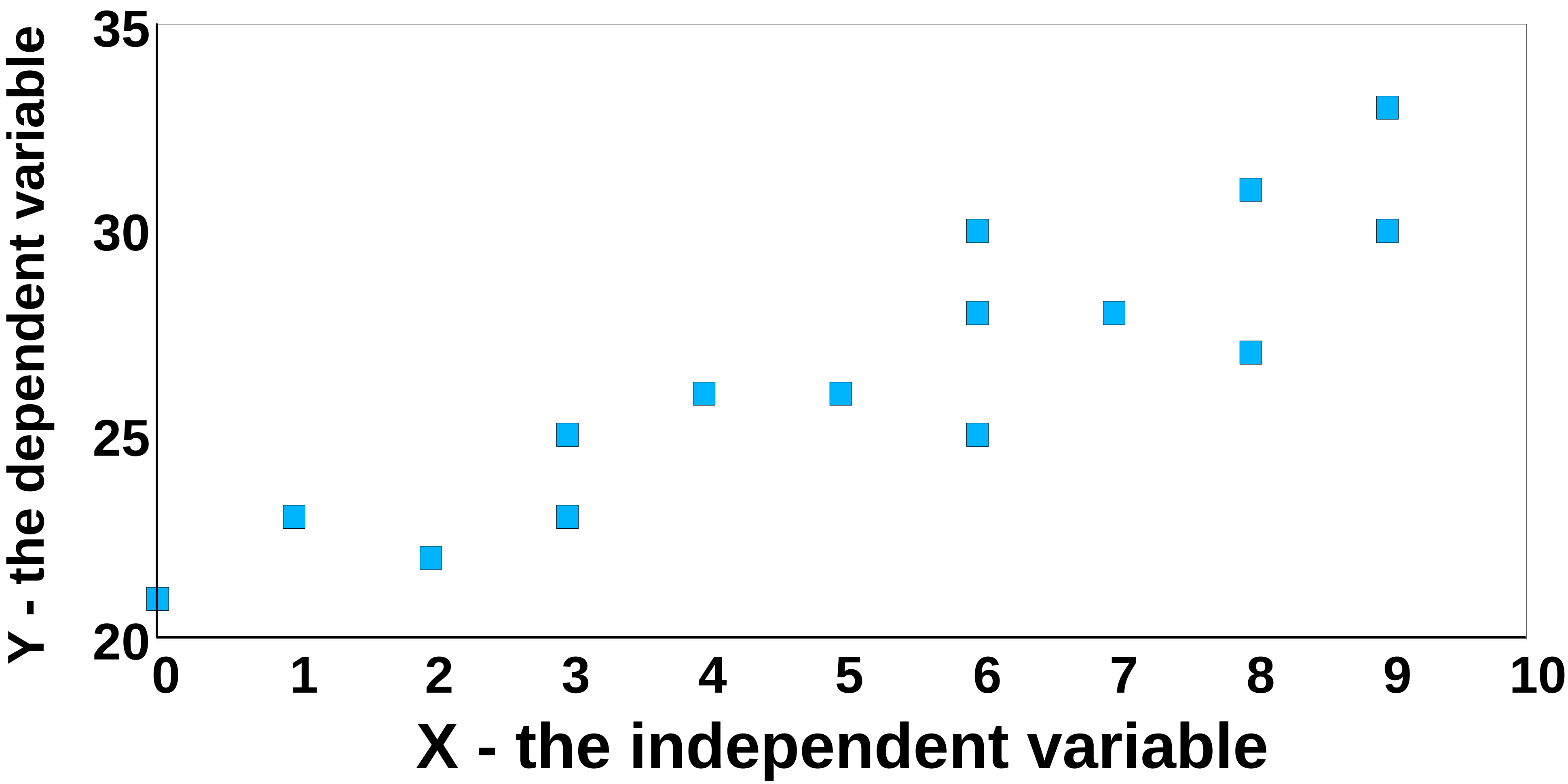
Statistical Techniques II
EXST7015

Simple Linear Regression



The objective

- **Given points plotted on two coordinates, Y and X, find the best line to fit the data.**



The concept

- **Data consists of paired observations with a presumed potential for the existence of some underlying relationship**
- **We wish to determine the nature of, and quantify, the relationship if it exists.**
  - **Note that we cannot prove that the relationship exists by using regression (e.g. we cannot prove cause and effect).**
  - **Regression can only show if a "correlation" exists, and provide an equation for the relationship.**

The concept *(continued)*

- **Given a data set consisting of paired, quantitative variables,**
- **and recognizing that there is variation in the data set,**
- **we will define,**
  - **POPULATION MODEL (SLR)**
    - $Y_{ij} = \beta_0 + \beta_1 X_i + \varepsilon_i$
- **This is the line we will fit.**

The concept *(continued)*

- **We must estimate the population equation for a straight line**
- **The Population Parameters estimated are**
  - **$\mu_{y.x}$ = the true population mean of Y at each value of X**
  - **β_0 = the true value of the Y intercept**
  - **β_1 = the true value of the slope, the change in Y per unit of X**
  - $\mu_{y.x} = \beta_0 + \beta_1 X_i$

Terminology

- **Dependent variable - variable to be predicted**
  - **Y = dependent variable (all variation occurs in Y)**
- **Independent variable - predictor or regressor variable**
  - **X = independent variable (X is measured without error)**

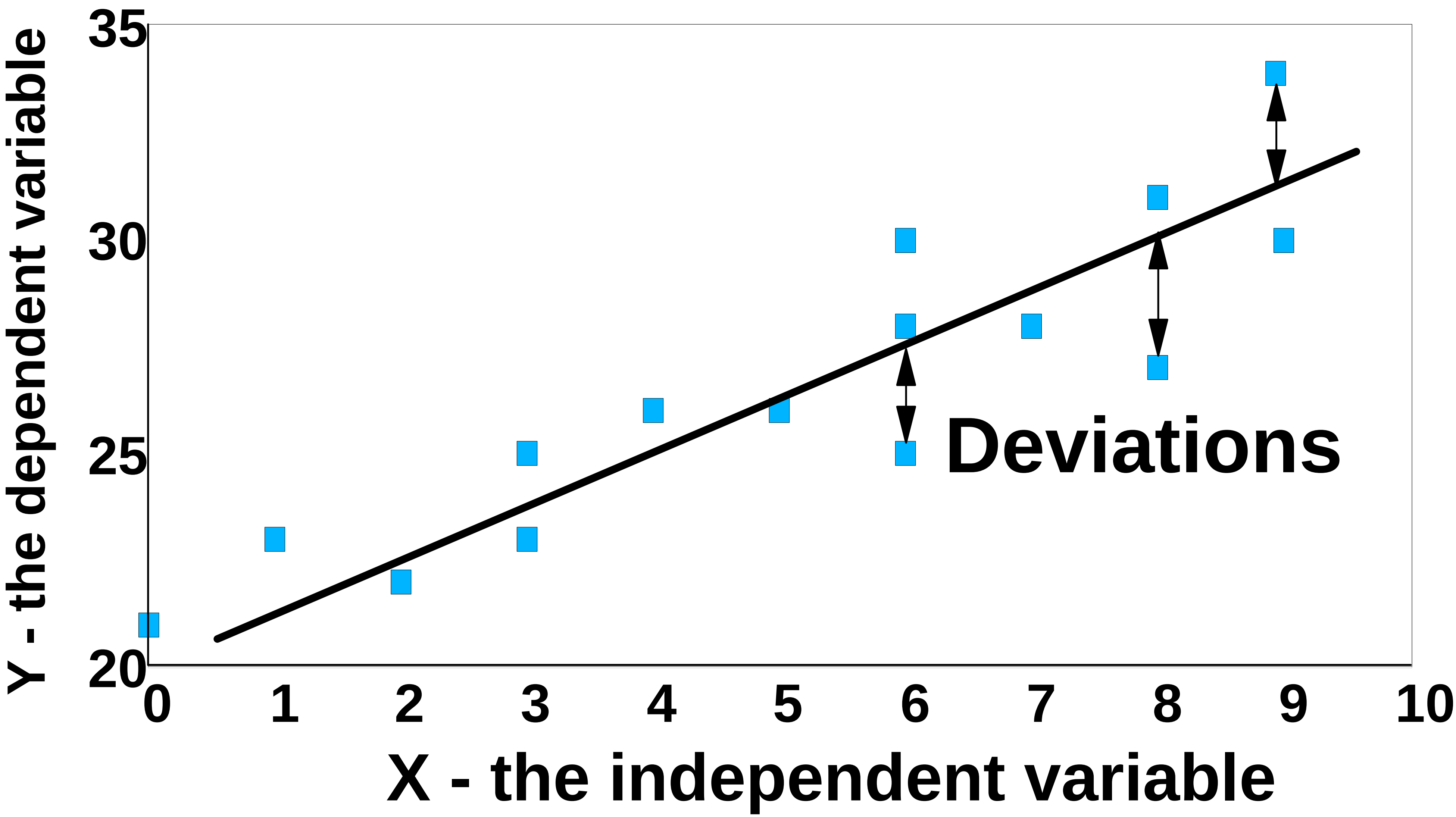
Terminology *(continued)*

- **Intercept - value of Y when X = 0, point where the regression line passes through the Y axis**
  - **NOTE: units are "Y" units**
- **Slope - the value of the change in Y for each unit increase in X**
  - **NOTE: units are "Y" units per "X" unit**

Terminology *(continued)*

- **Deviation - distance from an observed point to the regression line, also called a residual.**
- **Least squares regression line - the line that minimizes the squared distances from the line to the individual observations.**

Regression line

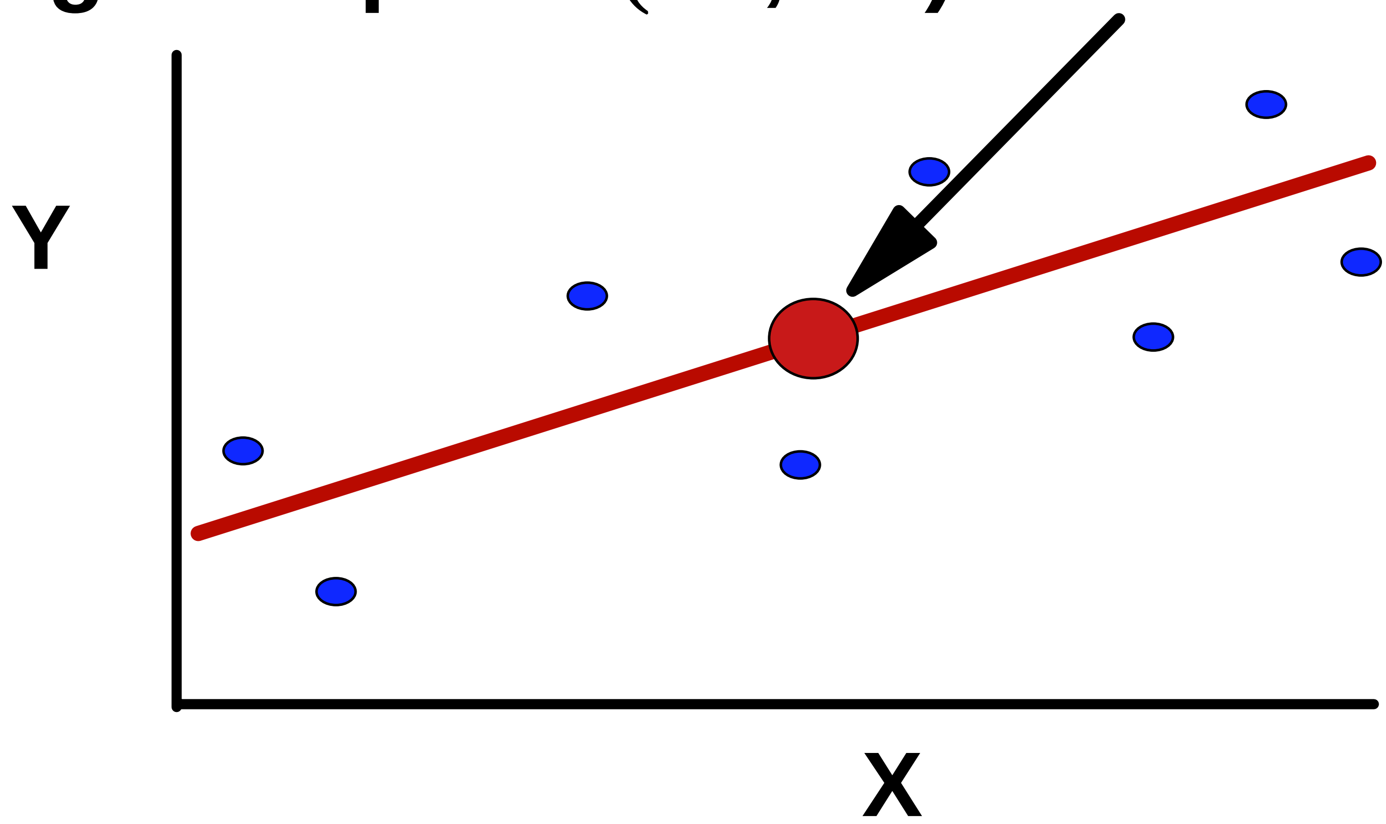


Regression calculations

- **All calculations for simple linear regression start with the same values. These are,**
- ΣX_i, ΣX_i^2, ΣY_i, ΣY_i^2, $\Sigma X_i Y_i$, n
- **Calculations for simple linear regression are first adjusted for the mean. These are called "corrected values". They are corrected for the MEAN by subtracting a "correction factor".**

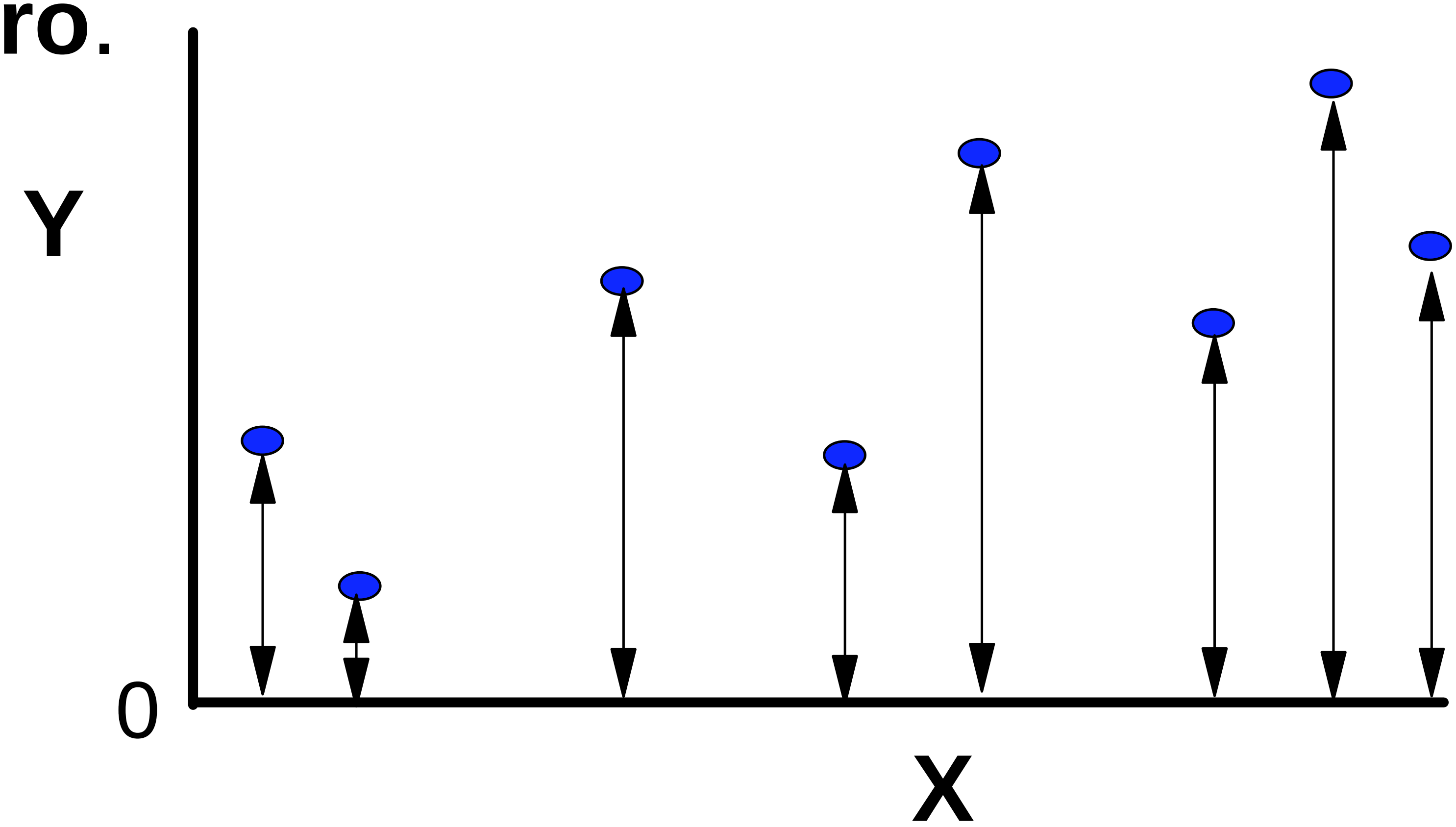
Reg calculations (*continued*)

- **As a result, all simple linear regressions are adjusted for the mean of X and Y and pass through the point ($\bar{X}$, $\bar{Y}$).**



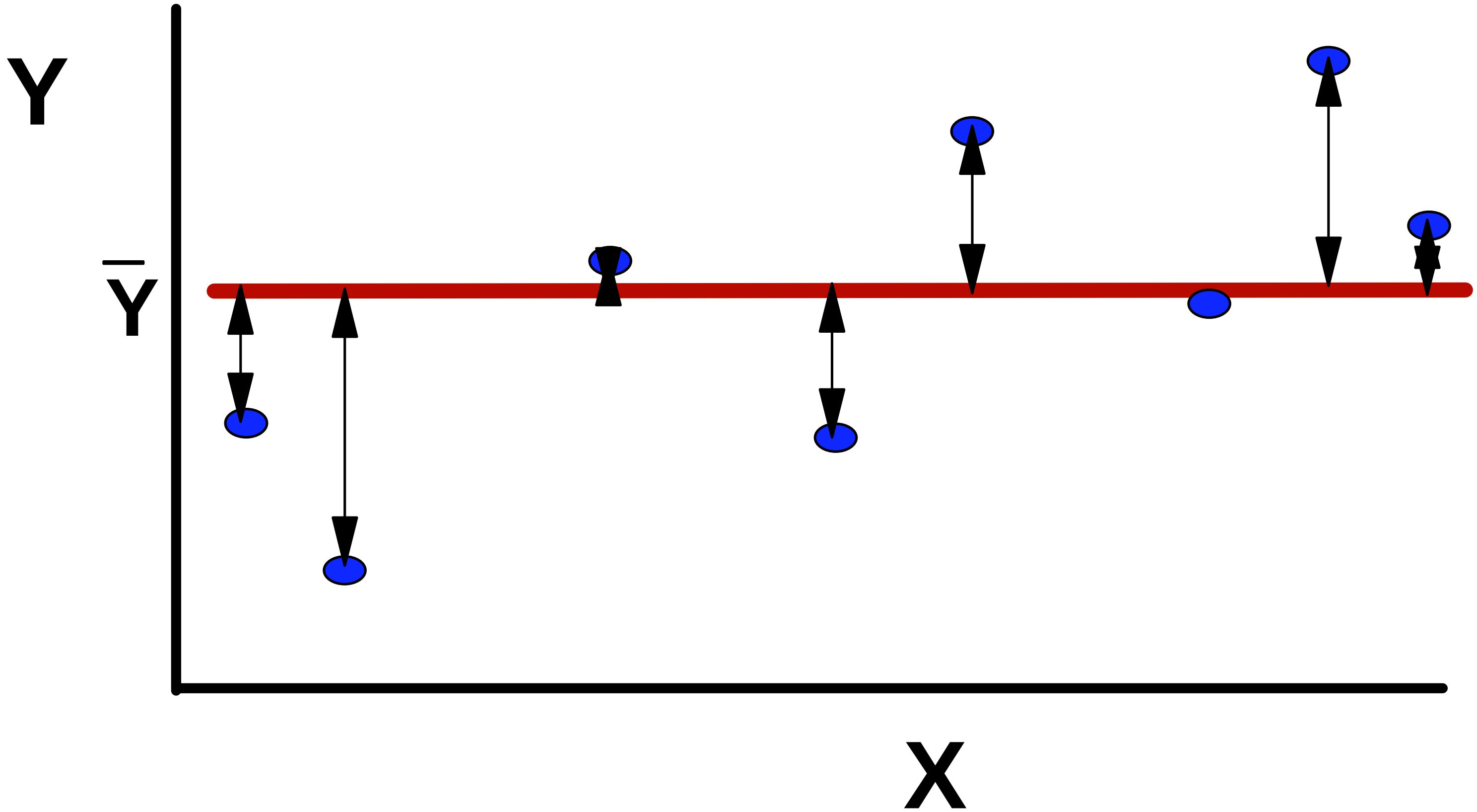
Reg calculations (*continued*)

- **The original sums and sums of squares of Y are distances and squared distances from zero.**



Reg calculations (*continued*)

- The corrected sums are zero (half negative and half positive) and sums of squares of Y are squared distances from the mean of Y.

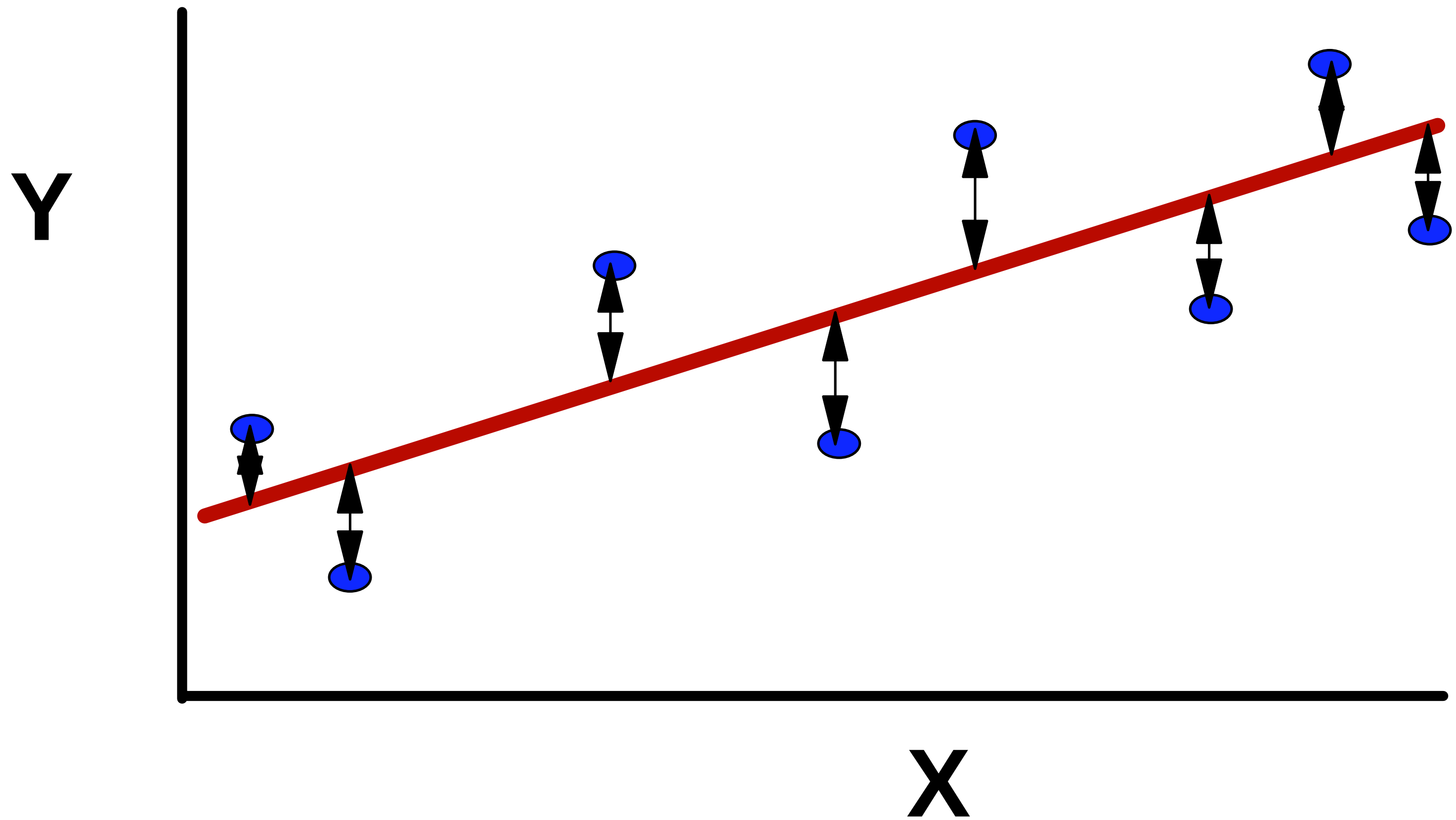


Reg calculations *(continued)*

- **Once corrected sums of squares and crossproducts are obtained (S_{XX}, S_{YY}, S_{XY}), the calculations are:**
- **Slope = $b_1 = S_{XY} / S_{XX}$**
- **Intercept = $b_0 = \bar{Y} - b_1 \bar{X}$**
- **We have fitted the sample equation**
  - **$Y_i = b_0 + b_1 X_i + e_i$**
  - $$\hat{Y}_i = b_0 + b_1 X_i$$

Variance Estimates

- **Once the regression line is fitted, Variance calculations are based on the deviations from the regression.**



Variance Estimates *(continued)*

- **From the regression model**
  - $Y_i = b_0 + b_1 X_i + e_i$
- **We derive the formula for the deviations**
  - $e_i = Y_i - (b_0 + b_1 X_i) =$

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

Variance Estimates (*continued*)

- **As with other calculations of variance, we calculate a sum of squares (corrected for the mean). This is simplified by the fact that the deviations, or residuals, already have a mean of zero.**

$$\text{SSResiduals} = \sum_{i=1}^{n} e_i^2 = \text{SSError}$$

Variance Estimates *(continued)*

- **The degrees of freedom (d.f.) for the variance calculation is n-2, since two parameters are estimated prior to the variance (b_0 and b_1).**
- **The variance estimate is called the MSE (Mean square error). It is the SSError divided by the d.f..**
- **MSE = SSE / (n-2)**

Variance Estimates *(continued)*

- **The variances for the two parameter estimates and the predicted values are all different, but all are based on the MSE, and all have n-2 d.f. (t-tests) or n-2 d.f. for the denominator (F tests).**
- **Variance of the slope = MSE/S_{XX}**
- **Variance of the intercept =**
  - $MSE[(1/n)+(-\bar{X})^2/S_{XX}]$
- **Variance of a predicted value at X_i =**
  - $MSE[(1/n)+(X_i-\bar{X})^2/S_{XX}]$

Variance Estimates *(continued)*

- **Any of these variance can be used for a t-test of an estimate against an hypothesized value for a parameter.**
- **Another common expression of regression results is an ANOVA table.**
  - **Given the SSError (sum of squared deviations from the regression)**
  - **And the initial total sum of squares (S_{YY}), the sum of squares of Y adjusted for the mean**
  - **we can construct an ANOVA table**

ANOVA table

Simple Linear Regression ANOVA table

	d.f.	Sum of Squares	Mean Square	F
Regression	1	SSRegression	MSReg	$\frac{MSReg}{MSError}$
Error	n-2	SSError	MSError	
Total	n-1	S_{YY} = SSTotal		

ANOVA table *(continued)*

- In the ANOVA table
- The SSRegression and SSError sum the SSTotal, so given the total (S_{YY}) and one of the two terms, we can get the other.
- The easiest to calculate is usually the SSRegression since we usually already have the necessary intermediate values.
  - SSRegression = $(S_{XY})^2 / S_{XX}$.

ANOVA table *(continued)*

- The SSRegression is a measure of the "improvement" in the fit due to the regression line. The deviations start at S_{YY} and are reduced to SSError. The difference is the improvement, and is equal to the SSRegression.
- This gives another statistic called the R^2. What portion of the total SS (S_{YY}) is accounted for by the regression.
- R^2 = SSRegression / SSTotal

ANOVA table *(continued)*

- **The degrees of freedom are,**
- **n-1 for the total, one lost for the correction for the mean (which also fits the intercept)**
- **n-2 for the error, since two parameters are estimated to get the regression line.**
- **1 d.f. for the regression, which is the d.f. for the slope.**

ANOVA table *(continued)*

- **The F test is constructed by calculating the MSRegression / MSError.**
  - **This has 1 and (n-2) d.f.**
  - **This is EXACTLY the same test as the t-test of the slope against zero.**
  - **To test the slope against an hypothesized value (say zero) with n-2 d.f., calculate**

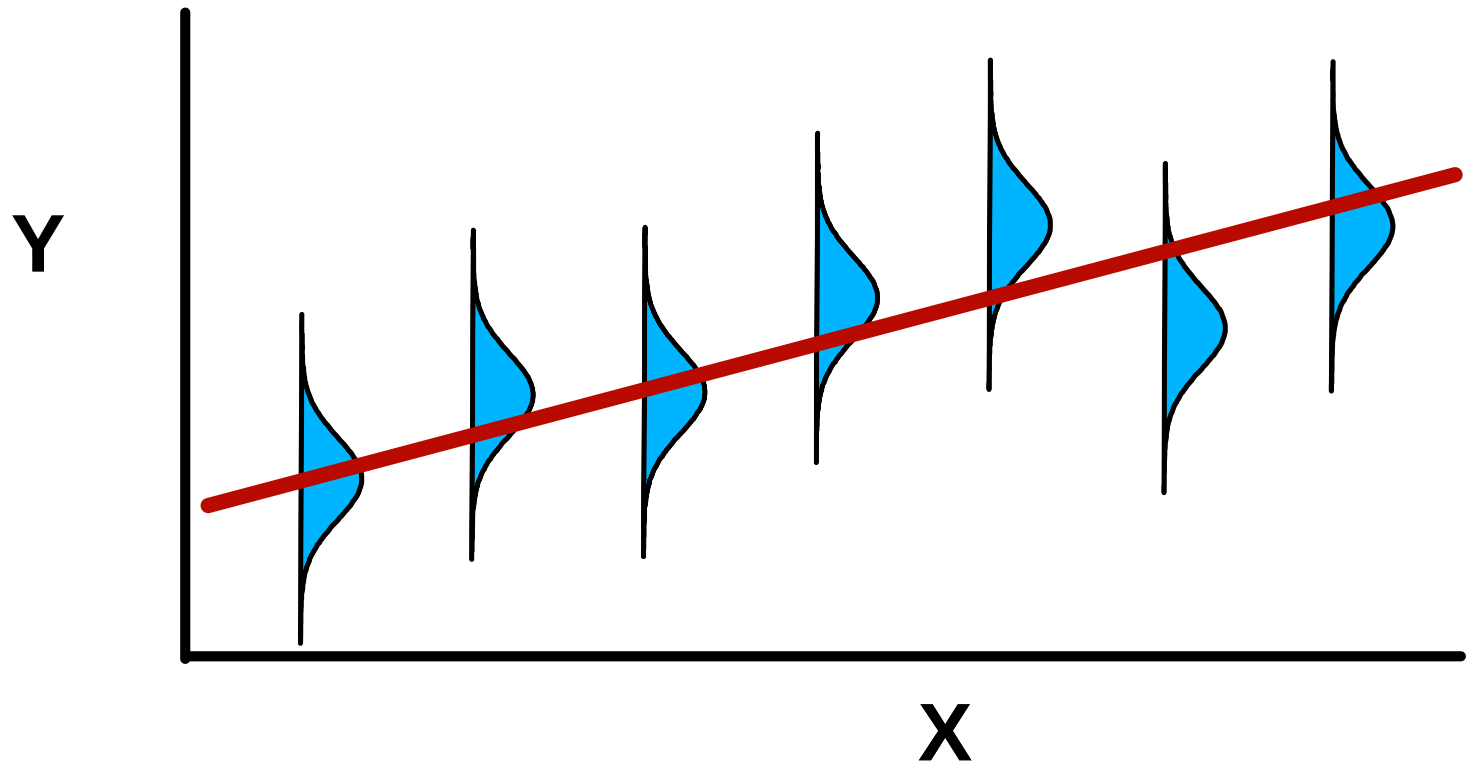
$$t = \frac{b_1 - b_{1Hypothesized}}{S_{b_1}} = \frac{b_1 - 0}{\sqrt{MSE / S_{XX}}}$$

Assumptions for the Regression

- **We will recognize 4 assumptions**
- **1) Normality - We take the deviations from regression and pool them all together into one estimate of variance. Some of the tests we use require the assumption of normality, so these deviations should be normally distributed.**

Assumptions for the Regression *(continued)*

- **For each value of X there is a population of values for the variable Y (normally distributed).**



Assumptions for the Regression *(continued)*

- **2) Homogeneity of variance - When we pool these deviations (variances) we also assume that the variances are the same at each value of X_i. In some cases this is not true, particularly when the variance increases as X increases.**
- **3) X is measured without error! Since variances are measured vertically only, all variance is in Y, no provisions are made for variance in X.**

Assumptions for the Regression *(continued)*

- **Independence. This enters in several places. First, the observations should be independent of each other (i.e. the value of e_i should be independent of e_j).**
- **Also, in the equation for the line**
  - $Y_i = b_0 + b_1X_i + e_i$
- **We assume that the term e_i is independent of the rest of the model. We will talk more of this when we get to multiple regression.**

Assumptions for the Regression *(continued)*

- **So the four assumptions are:**
  - **Normality**
  - **Homogeneity of variance**
  - **Independence**
  - **X measured without error**
- **These are explicit assumptions, and we will often test these assumptions.**

Assumptions for the Regression *(continued)*

- **There are some other assumptions that I consider implicit. We will not state these, but in some cases they can be tested. For example,**
  - **There is order in the Universe**
  - **The underlying fundamental relationship that I just fitted a straight line to really is a straight line. This one can be examined statistically.**

Characteristics of a Regression Line

- **The line will pass through the point $\bar{X}, \bar{Y}$ (also the point 0, b_0)**
- **The sum of deviations will be zero ($\Sigma e_i = 0$)**
- **The sum of squared deviations (measured vertically, $\Sigma e_i^2 = \Sigma(Y_i - b_0 + b_1 X_i)^2$) of the points from the regression line will be a minimum.**
- **Values on the line can be described by the equation** $\hat{Y}_i = b_0 + b_1 X_i$

Characteristics of a Regression

- **The line has some desirable properties (if the assumptions are met)**
  - **$E(b_0) = \beta_0$**
  - **$E(b_1) = \beta_1$**
  - **$E(\bar{Y}_X) = \mu_{X.Y}$**
  - **Therefore, the parameter estimates and predicted values are unbiased estimates.**
- **Note that linear regression is considered statistically robust. That is, the analysis tends to give good results if the assumptions are not violated to a great extent.**

About crossproducts and correlation

- **Crossproducts are used in a number of related calculations.**
- **a crossproduct = Y_iX_i**
- **Sum of crossproducts = $\Sigma Y_iX_i = S_{XY}$**
- **Covariance = S_{XY} / (n-1)**
- **Slope = S_{XY} / S_{XX}**
- **SSRegression = S^2_{XY} / S_{XX}**
- **Correlation = $S_{XY} / \sqrt{S_{XX}S_{YY}}$**
- **$R^2 = r^2 = S^2_{XY} / S_{XX}S_{YY}$ = SSRegression/SSTotal**

Summary

- See Simple linear regression notes from EXST7005 for additional information, including the derivation of the equations for the slope and intercept. You are not responsible for these derivations.
- Know the terminology, characteristics and properties of a regression line, the assumptions, and the components to the ANOVA table.

Summary *(continued)*

- You will not be fitting regressions by hand, but I will expect you to understand where the values on SAS output come from and what they mean.
- Particular emphasis will be placed on working with, and interpreting, numerical regression analyses.