## UCLA STAT 13
## Introduction to Statistical Methods


- **Instructor:** **Ivo Dinov,**
  Asst. Prof. In Statistics and Neurology
- **Teaching Assistants:** Tom Daula and Kaiding Zhu,
  UCLA Statistics

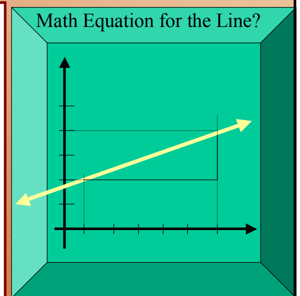
University of California, Los Angeles, Fall 2002
http://www.stat.ucla.edu/~dinov/




## Chapter 12: Lines in 2D
**(Regression and Correlation)**

- Vertical Lines
- Horizontal Lines
- Oblique lines
- Increasing/Decreasing
- Slope of a line
- Intercept
- $Y=\alpha X + \beta$, in general.

Math Equation for the Line?

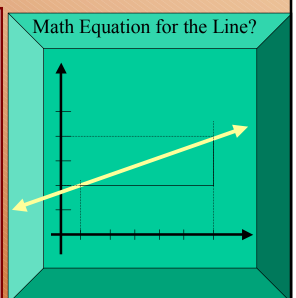




## Chapter 12: Lines in 2D
**(Regression and Correlation)**

- Draw the following lines:
- Y=2X+1
- Y=-3X-5
- Line through $(X_1,Y_1)$ and $(X_2,Y_2)$.
- $(Y-Y_1)/(Y_2-Y_1)=$
  $(X-X_1)/(X_2-X_1)$.

Math Equation for the Line?



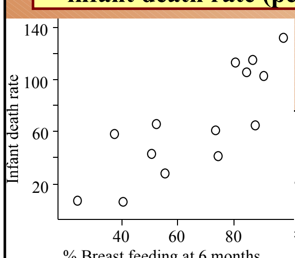


## Approaches for modeling data relationships
**Regression and Correlation**

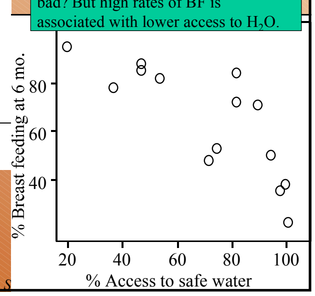
- There are random and nonrandom variables
- Correlation applies if both variables (X/Y) are random (e.g., We saw a previous example, systolic vs. diastolic blood pressure SISVOL/DIAVOL) and are treated symmetrically.
- Regression applies in the case when you want to single out one of the variables (response variable, Y) and use the other variable as predictor (explanatory variable, X), which explains the behavior of the response variable, Y.



## Causal relationship?
## – infant death rate (per 1,000) in 14 countries

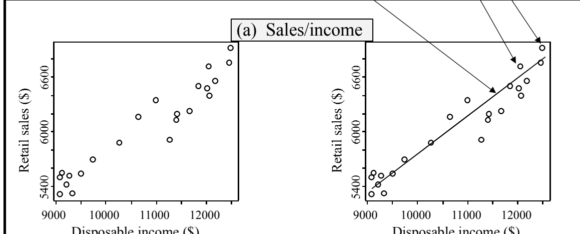



Strong evidence (linear pattern) of death rate increase with increasing level of breastfeeding (BF)? Naïve conclusion breast feeding is bad? But high rates of BF is associated with lower access to $H_2O$.




Predict behavior of Y (response) Based on the values of X (explanatory var.) Strategies for uncovering the reasons (causes) for an observed effect.

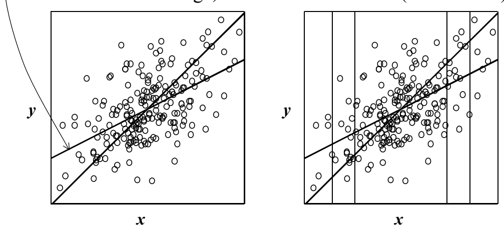
## Regression relationship = trend + residual scatter




- Regression is a way of studying relationships between variables (random/nonrandom) for predicting or explaining behavior of 1 variable (response) in terms of others (explanatory variables or predictors).

## Looking vertically

Flatter line gives better prediction, since it approx. goes through the middle of the Y-range, for each fixed x-value (vertical line)




(a) Which line? (b) Flatter line gives better predictions.

**Figure 3.1.8** Educating the eye to look vertically.





## Correlation Coefficient

Correlation coefficient ($-1 \le R \le 1$): a measure of linear association, or clustering around a line of multivariate data.

Relationship between two variables (X, Y) can be summarized by: $(\mu_X, \sigma_X)$, $(\mu_Y, \sigma_Y)$ and the correlation coefficient, $R$. $R=1$, perfect positive correlation (straight line relationship), $R=0$, no correlation (random cloud scatter), $R=-1$, perfect negative correlation.

Computing $R(X,Y)$: (standardize, multiply, average)

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

$X=\{x_1, x_2, \ldots, x_N\}$
$Y=\{y_1, y_2, \ldots, y_N\}$
$(\mu_X, \sigma_X)$, $(\mu_Y, \sigma_Y)$
sample mean / SD.



## Correlation Coefficient

Example:

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

| Student $i$ | Height $x_i$ | Weight $y_i$ | $x_i-\bar{x}$ | $y_i-\bar{y}$ | $(x_i-\bar{x})^2$ | $(y_i-\bar{y})^2$ | $(x_i-\bar{x})(y_i-\bar{y})$ |
|---|---|---|---|---|---|---|---|
| 1 | 167 | 60 | 6 | 4.67 | 36 | 21.8089 | 28.02 |
| 2 | 170 | 64 | 9 | 8.67 | 81 | 75.1689 | 78.03 |
| 3 | 160 | 57 | -1 | 1.67 | 1 | 2.7889 | -1.67 |
| 4 | 152 | 46 | -9 | -9.33 | 81 | 87.0489 | 83.97 |
| 5 | 157 | 55 | -4 | -0.33 | 16 | 0.1089 | 1.32 |
| 6 | 160 | 50 | -1 | -5.33 | 1 | 28.4089 | 5.33 |
| Total | 966 | 332 | 0 | $\approx 0$ | 216 | 215.3334 | 195.0 |



## Correlation Coefficient

Example:

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

$$\mu_x = \frac{966}{6} = 161\,\text{cm}, \quad \mu_y = \frac{332}{6} = 55\,\text{kg},$$

$$\sigma_x = \sqrt{\frac{216}{5}} = 6.573, \quad \sigma_y = \sqrt{\frac{215.3}{5}} = 6.563,$$

$$Corr(X,Y) = R(X,Y) = 0.904$$



## Correlation Coefficient - Properties

Correlation is invariant w.r.t. linear transformations of X or Y

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right) =$$

$$R(aX+b, cY+d), \quad \text{since}$$

$$\left(\frac{ax_k+b-\mu_{ax+b}}{\sigma_{ax+b}}\right) = \left(\frac{ax_k+b-(a\mu_x+b)}{a\times\sigma_x}\right) =$$

$$\left(\frac{a(x_k-\mu_x)+b-b}{a\times\sigma_x}\right) = \left(\frac{x_k-\mu_x}{\sigma_x}\right)$$

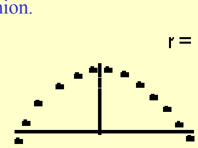


## Correlation Coefficient - Properties

Correlation is Associative

$$R(X,Y) = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right) = R(Y,X)$$

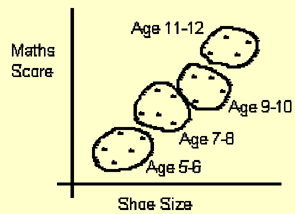
Correlation measures linear association, NOT an association in general!!! So, Corr(X,Y) could be misleading for X & Y related in a non-linear fashion.



## Correlation Coefficient - Properties

$$R(X,Y)=\frac{1}{N}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)=R(Y,X)$$

1. *R* measures the extent of linear association between two continuous variables.
2. Association does not imply causation - both variables may be affected by a third variable – age was a confounding variable.






## Essential Points

6. If the experimenter has control of the levels of *X* used, how should these levels be allocated to the available experimental units?

At random! Example, testing hardness of concrete, Y, based on levels of cement, X, incorporated. Factors effecting Y: amount of $H_2O$, ratio stone-chips to sand, drying conditions, etc. To prevent uncontrolled differences in batches of concrete in confounding our impression of cement effects, we should choose which batch ($H_20$ levels, sand, dry-conditions) gets what amount of cement at random! Then investigate for X-effects in Y observations. If some significance test indicates observed trend is significantly different from a random pattern → we have evidence of causal relationship, which may strengthen even further if the results are replicable.



## Essential Points

7. What theories can you explore using regression methods?

Prediction, explanation/causation, testing a scientific hypothesis/mathematical model:

a. Hooke's spring law: amount of stretch in a spring, Y, is related to the applied weight X by $Y=\alpha+\beta X$, a, b are spring constants.

b. Theory of gravity: force of gravity F between 2 objects is given by $F = \alpha/D^\beta$, where D=distance between objects, a is a constant related to the masses of the objects and $\beta=2$, according to the inverse square law.

c. Economic production function: $Q=\alpha L^\beta K^\gamma$, Q=production, L=quantity of labor, K=capital, $\alpha,\beta,\gamma$ are constants specific to the market studied.



## Essential Points

8. People fit theoretical models to data for three main purposes.

a. To test the model, itself, by checking if the data is reasonably close agreement with the relationship predicted by the model.

b. Assuming the model is correct, to test if theoretically specified values of a parameter are consistent with the data (y=2x+1 vs. y=2.1x-0.9).

c. Assuming the model is correct, to estimate unknown constants in the model so that the relationship is completely specified (y=ax+5, a=?)



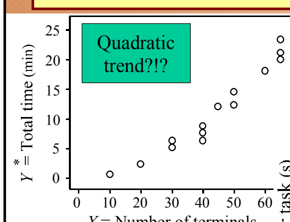
## Trend and Scatter - Computer timing data

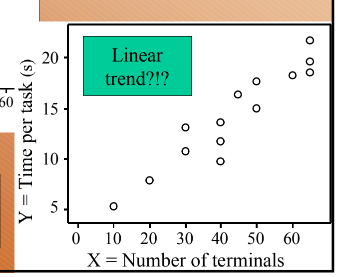
- The major components of a regression relationship are trend and scatter around the trend.
- To investigate a trend – fit a math function to data, or smooth the data.
- Computer timing data: a mainframe computer has X users, each running jobs taking Y min time. The main CPU swaps between all tasks. Y* is the total time to finish all tasks. Both Y and Y* increase with increase of tasks/users, but how?

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| X | = | Number of terminals: | 40 | 50 | 60 | 45 | 40 | 10 | 30 | 20 |
| Y* | = | Total Time (mins): | 6.6 | 14.9 | 18.4 | 12.4 | 7.9 | 0.9 | 5.5 | 2.7 |
| Y | = | Time Per Task (secs): | 9.9 | 17.8 | 18.4 | 16.5 | 11.9 | 5.5 | 11 | 8.1 |
| X | = | Number of terminals: | 50 | 30 | 65 | 40 | 65 | 65 | | |
| Y* | = | Total Time (mins): | 12.6 | 6.7 | 23.6 | 9.2 | 20.2 | 21.4 | | |
| Y | = | Time Per Task (secs): | 15.1 | 13.3 | 21.8 | 13.8 | 18.6 | 19.8 | | |



## Trend and Scatter - Computer timing data

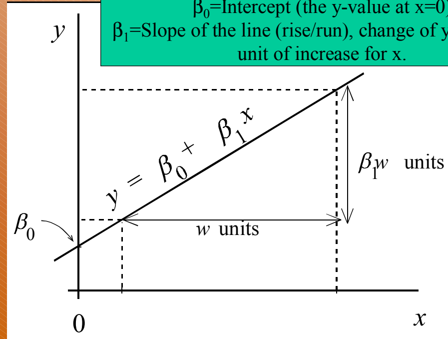



We want to find reasonable models (descriptions) for these data!

## Equation for the straight line – linear/affine function

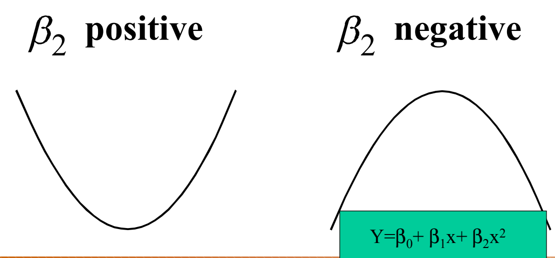
$\beta_0$=Intercept (the y-value at x=0)
$\beta_1$=Slope of the line (rise/run), change of y for every unit of increase for x.






## The quadratic curve

### Quadratic Curve

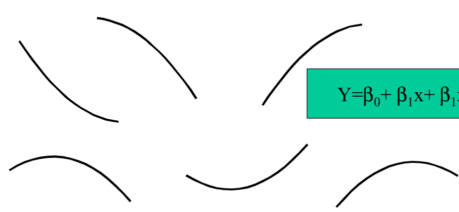



$Y=\beta_0+\beta_1x+\beta_2x^2$



## The quadratic curve

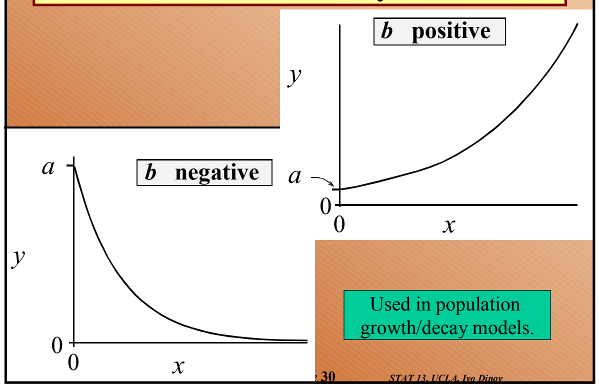
### Segments of the curve



$Y=\beta_0+\beta_1x+\beta_1x^2$



## The exponential curve, $y = a\, e^{bx}$




Used in population growth/decay models.



## Effects of changing x for different functions/curves

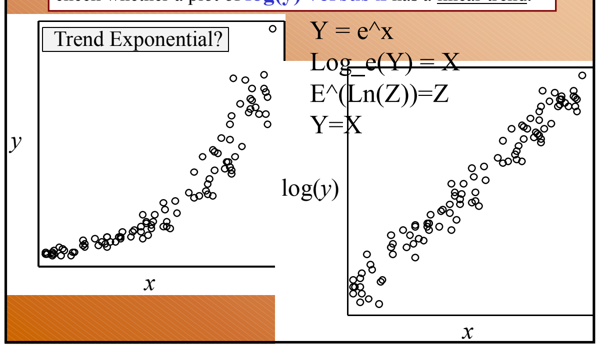
A straight ***line*** changes by a fixed ***amount*** with each unit change in *x*.

An ***exponential*** changes by a fixed ***percentage*** with each unit change in *x*.



## To tell whether a trend is exponential ….

check whether a plot of **log(y) versus x** has a linear trend.

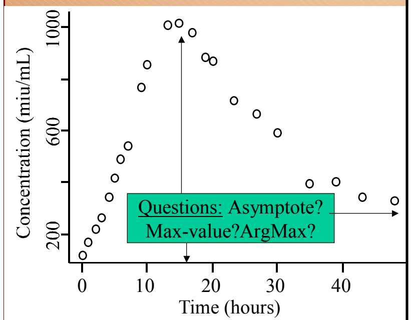



Y = e^x
Log_e(Y) = X
E^(Ln(Z))=Z
Y=X

## Creatine kinase concentration in patient's blood

**You should not let the questions you want to ask be dictated by the tools you know how to use.**

Here Y=creatine kinase concentration in blood for a set of heart attack patients vs. the time, X.




No symmetry so $X^2$ models won't work!



## Comments

1. In statistics what are the two main approaches to summarizing trends in data? (model fitting; smoothing – done by the eye!)
2. In $y = 5x + 2$, what information do the 5 and the 2 convey? (slope, y-intercept)
3. In $y = 7 + 5x$, what change in $y$ is associated with a 1-unit increase in $x$? with a 10-unit increase? (5; 50)

   How about for $y = 7- 5x$. (-5; -50)

5. How can we tell whether a trend in a scatter plot is exponential? (plot ***log***(Y) vs. X, should be linear)



Choosing the "best-fitting" line

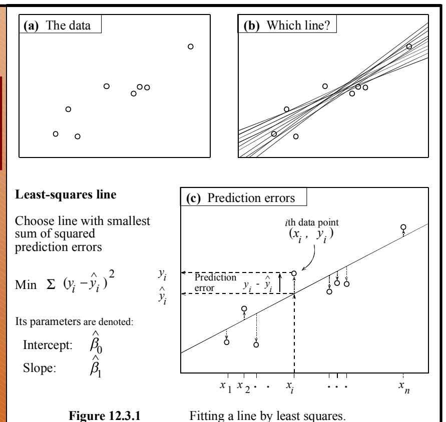
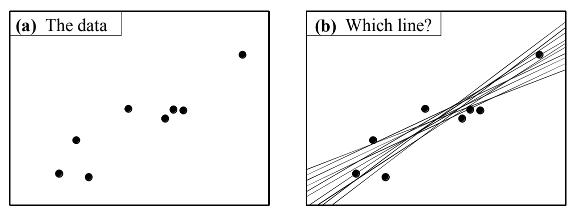



**Figure 12.3.1** Fitting a line by least squares.

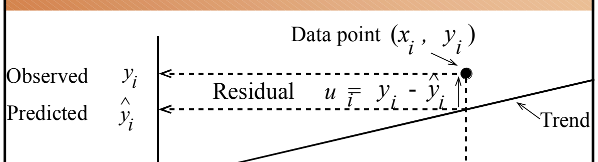


## Fitting a line through the data






## The idea of a residual or prediction error






## Least squares criterion

***Least squares criterion***: Choose the values of the parameters to *minimize the sum of squared prediction errors* (or sum of squared residuals),

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

## The least squares line

**Least-squares line**

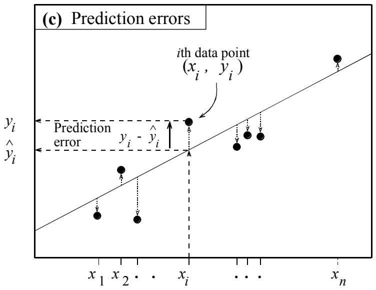
Choose line with smallest sum of squared prediction errors

Min $\Sigma\ (y_i - \hat{y}_i)^2$

Its parameters are denoted:

Intercept: $\hat{\beta}_0$

Slope: $\hat{\beta}_1$




***Least-squares line:*** $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$



## The least squares line

***Least-squares line:*** $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left[(x_i - \bar{x})(y_i - \bar{y})\right]}{\sum_{i=1}^{n} (x_i - \bar{x})^2}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



## Computer timings data – linear fit

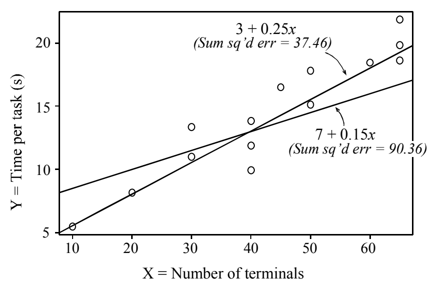



**Figure 12.3.2** Two lines on the computer-timings data.

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.



## Computer timings data

**TABLE 12.3.1 Prediction Errors**

| | | 3 + 0.25x | | 7 + 0.15x | |
|---|---|---|---|---|---|
| $x$ | $y$ | $\hat{y}$ | $y-\hat{y}$ | $\hat{y}$ | $y-\hat{y}$ |
| 40 | 9.90 | 13.00 | -3.10 | 13.00 | -3.10 |
| 50 | 17.80 | 15.50 | 2.30 | 14.50 | 3.30 |
| 60 | 18.40 | 18.00 | 0.40 | 16.00 | 2.40 |
| 45 | 16.50 | 14.25 | 2.25 | 13.75 | 2.75 |
| 40 | 11.90 | 13.00 | -1.10 | 13.00 | -1.10 |
| 10 | 5.50 | 5.50 | 0.00 | 8.50 | -3.00 |
| 30 | 11.00 | 10.50 | 0.50 | 11.50 | -0.50 |
| 20 | 8.10 | 8.00 | 0.10 | 10.00 | -1.90 |
| 50 | 15.10 | 15.50 | -0.40 | 14.50 | 0.60 |
| 30 | 13.30 | 10.50 | 2.80 | 11.50 | 1.80 |
| 65 | 21.80 | 19.25 | 2.55 | 16.75 | 5.05 |
| 40 | 13.80 | 13.00 | 0.80 | 13.00 | 0.80 |
| 65 | 18.60 | 19.25 | -0.65 | 16.75 | 1.85 |
| 65 | 19.80 | 19.25 | 0.55 | 16.75 | 3.05 |
| Sum of squared errors | | | 37.46 | | 90.36 |



## Adding the least squares line

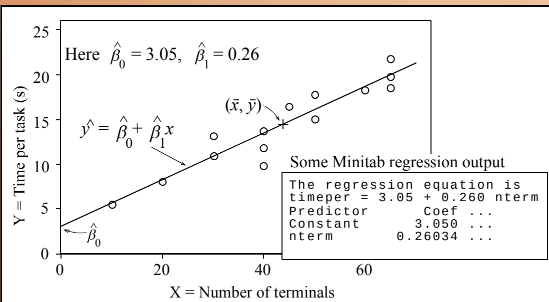



**Figure 12.3.3** Computer-timings data with least-squares line.

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.



## Review, Fri., Oct. 19, 2001

1. The least-squares line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ passes through the points ($x$ = 0, $\hat{y}$ = ?) and ($x = \bar{x}$, $\hat{y}$ = ?). Supply the missing values.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left[(x_i - \bar{x})(y_i - \bar{y})\right]}{\sum_{i=1}^{n} (x_i - \bar{x})^2}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

## Hands – on worksheet !

1. X={-1, 2, 3, 4}, Y={0, -1, 1, 2},

| X | Y | $x-\bar{x}$ | $y-\bar{y}$ | $(x-\bar{x})^2$ | $(y-\bar{y})^2$ | $(x-\bar{x})\times(y-\bar{y})$ | |
|---|---|---|---|---|---|---|---|
| -1 | 0 | | | | | | |
| 2 | -1 | | | | | | |
| 3 | 1 | | | | | | |
| 4 | 2 | | | | | | |

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$



## Hands – on worksheet !

1. X={-1, 2, 3, 4}, Y={0, -1, 1, 2}, $\bar{x} = 2$, $\bar{y} = 0.5$

| X | Y | $x-\bar{x}$ | $y-\bar{y}$ | $(x-\bar{x})^2$ | $(y-\bar{y})^2$ | $(x-\bar{x})\times(y-\bar{y})$ | |
|---|---|---|---|---|---|---|---|
| -1 | 0 | -3 | -0.5 | 9 | 0.25 | 1.5 | |
| 2 | -1 | 0 | -1.5 | 0 | 2.25 | 0 | |
| 3 | 1 | 1 | 0.5 | 1 | 0.25 | 0.5 | |
| 4 | 2 | 2 | 1.5 | 4 | 2.25 | 3 | |
| 2 | 0.5 | | | 14 | 5 | 5 | |

$\beta_1$=5/14
$\beta_0$=y^-β1*x^
$\beta_0$= 0.5-10/14



## Course Material Review

1. ===========Part I=================
2. Data collection, surveys.
3. Experimental vs. observational studies
4. Numerical Summaries (5-#-summary)
5. Binomial distribution (prob's, mean, variance)
6. Probabilities & proportions, independence of events and conditional probabilities
7. Normal Distribution and normal approximation



## Course Material Review – cont.

1. ===============Part II=================
2. Central Limit Theorem – sampling distribution of $\overline{X}$
3. Confidence intervals and parameter estimation
4. Hypothesis testing
5. Paired vs. Independent samples
6. Analysis Of Variance (1-way-ANOVA, one categorical var.)
7. Correlation and regression
8. Best-linear-fit, least squares method