# UCLA STAT 13

## Introduction to Statistical Methods for the Life and Health Sciences


- **Instructor**: **Ivo Dinov**,
  Asst. Prof. In Statistics and Neurology
- **Teaching Assistants**: Sovia Lau, Jason Cheng
  UCLA Statistics

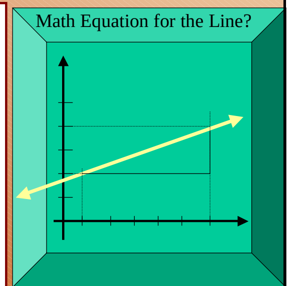
University of California, Los Angeles, Fall 2003


*http://www.stat.ucla.edu/~dinov/courses_students.html*



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

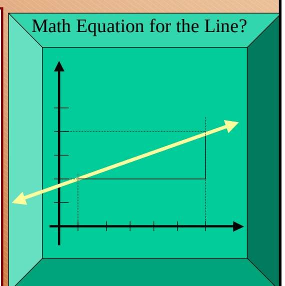
- Vertical Lines
- Horizontal Lines
- Oblique lines
- Increasing/Decreasing
- Slope of a line
- Intercept
- $Y=\alpha X + \beta$, in general.






## 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 $(X_2,Y_2)$.
- $(Y-Y_1)/(Y_2-Y_1)= (X-X_1)/(X_2-X_1)$.



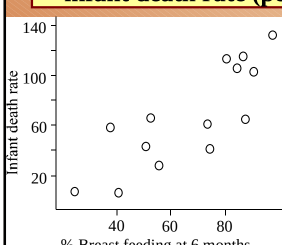



## Approaches for modeling data relationships
**Regression and Correlation**

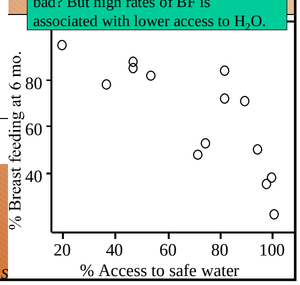
- There are random and nonrandom variables
- Correlation applies if both variables (X/Y) are random (e.g., We saw a previous example, systolic vs. diastolic blood pressure SISVOL/DIAVOL) and are treated symmetrically.
- Regression applies in the case when you want to single out one of the variables (response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

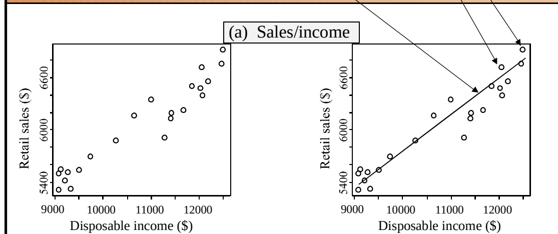



Strong evidence (linear pattern) of death rate increase with increasing level of breastfeeding (BF)? Naïve conclusion breast feeding is bad? But high rates of BF is associated with lower access to $H_2O$.




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

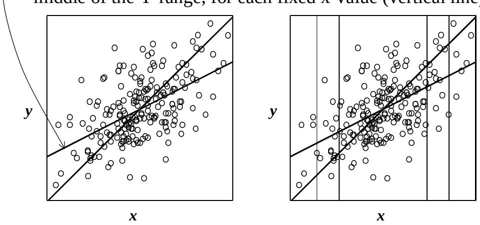
## Regression relationship = trend + residual scatter




- 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:

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

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

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

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



## Correlation Coefficient - Properties

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

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

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

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

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

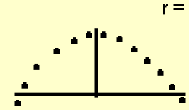


## Correlation Coefficient - Properties

Correlation is Associative

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

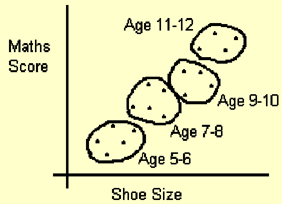
Correlation measures linear association, NOT an association in general!!! So, Corr(X,Y) could be misleading for X & Y related in a non-linear fashion.



## 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.





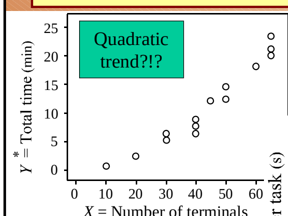

## Trend and Scatter - Computer timing data

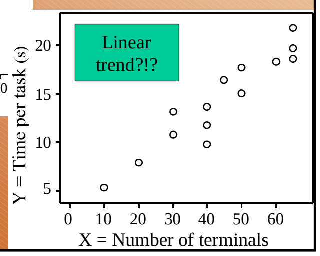
- The major components of a regression relationship are trend and scatter around the trend.
- To investigate a trend – fit a math function to data, or smooth the data.
- Computer timing data: a mainframe computer has X users, each running jobs taking Y min time. The main CPU swaps between all tasks.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

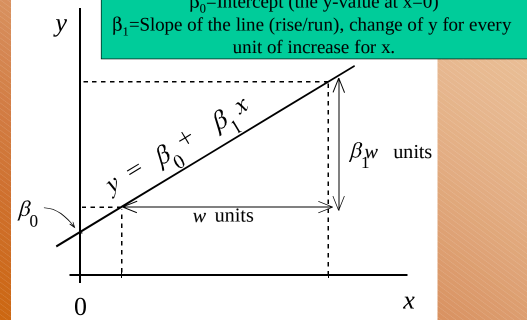



We want to find reasonable models (descriptions) for these data!

## 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.

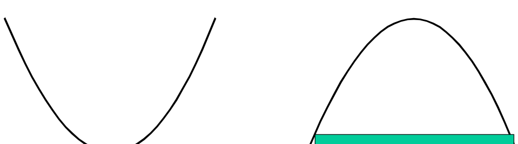





## The quadratic curve

### Quadratic Curve

$\beta_2$ **positive** $\beta_2$ **negative**

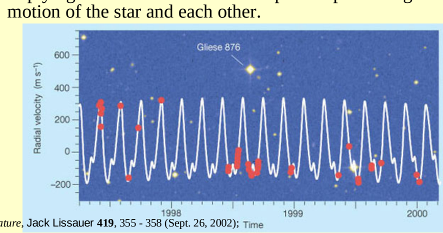


$Y=\beta_0+\beta_1 x+\beta_2 x^2$



## Other Non-linear model curves (trigonometric, piece-wise polynomial)

- Data from the Keck telescope in Hawaii (red points) show the variation over time of the radial velocity of the star *Gliese 876*. The white curve is the best fit to the data points, implying that there are two unseen planets perturbing the motion of the star and each other.




*Nature*, Jack Lissauer **419**, 355 - 358 (Sept. 26, 2002);

## Choosing the "best-fitting" line

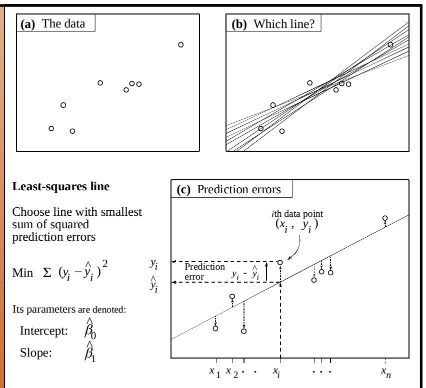
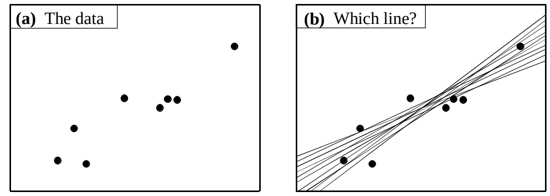



**Figure 12.3.1** Fitting a line by least squares.

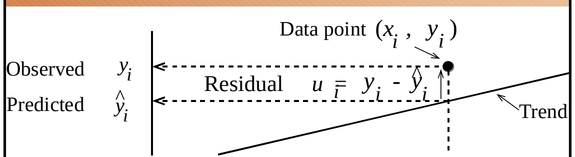
From Chance Encounters by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

## Fitting a line through the data






## The idea of a residual or prediction error





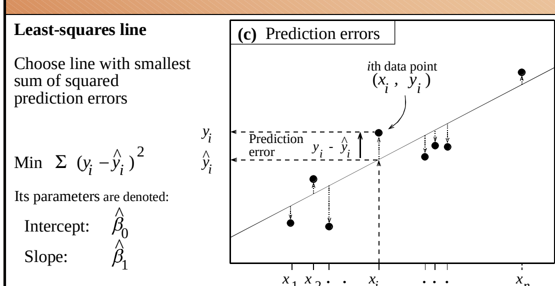

## Least squares criterion

***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:*** $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$



## 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

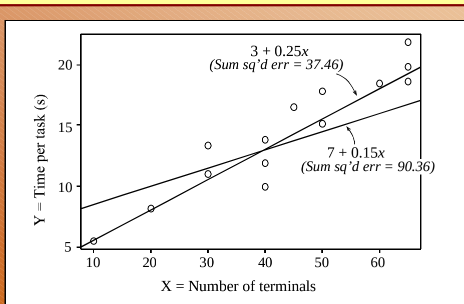



**Figure 12.3.2** Two lines on the computer-timings data.





## 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

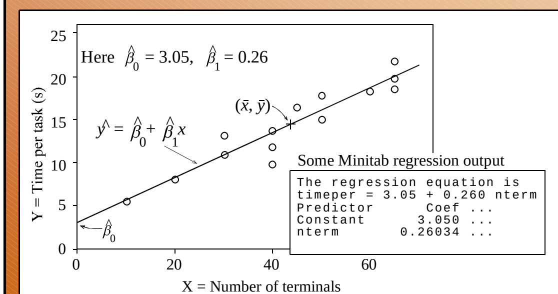



**Figure 12.3.3** Computer-timings data with least-squares line.





## 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 !

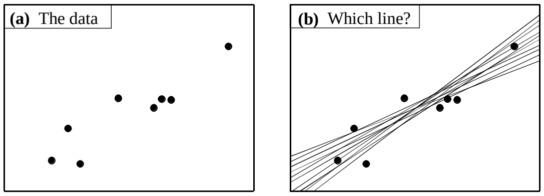
1. X={-1, 2, 3, 4}, Y={0, -1, 1, 2}, $\bar{x} = 2$, $\bar{y} = 0.5$

| 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

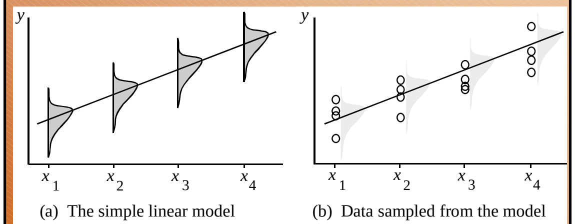
## Fitting a line through the data

Show the Regression-Line Simulation Applet
RegressionApplet.html






## The simple linear model




When $X = x$, $Y \sim \text{Normal}(\mu_Y, \sigma)$ where $\mu_Y = \beta_0 + \beta_1 x$, **OR**

when $X = x$, $\mathbf{Y = \beta_0 + \beta_1\, x + U}$, where $U \sim \text{Normal}(0,\sigma)$

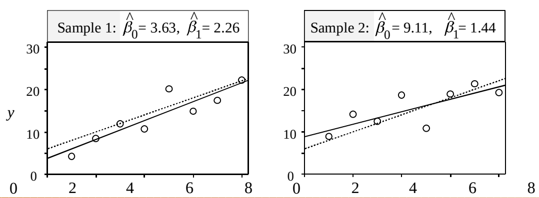
Random error



**Data generated from** $Y = 6 + 2x$ + error (U)

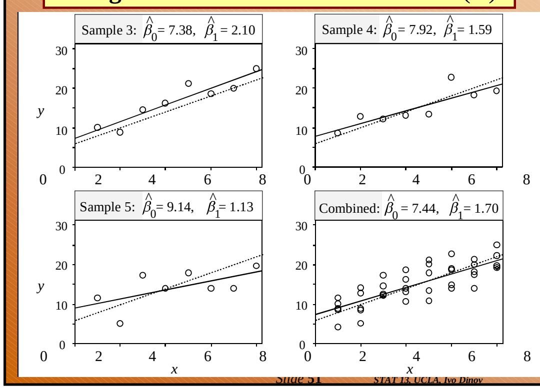
Dotted line ··········· is true line and solid line ——— is the data-estimated LS line.

Note differences between true $\beta_0=6$, $\beta_1=2$ and their estimates $\beta_0$^ & $\beta_1$^.



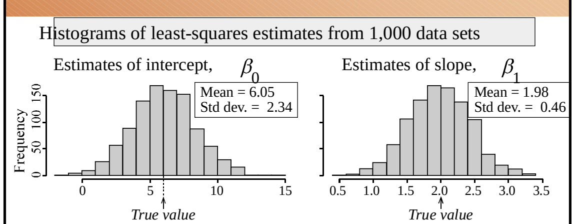



## Data generated from $Y = 6 + 2x$ + error(U)






## Data generated from $Y = 6 + 2x$ + error(U)




Data generated from the model $Y = 6 + 2\,x + U$
where $U \sim \text{Normal}(\mu = 0, \sigma = 3)$.



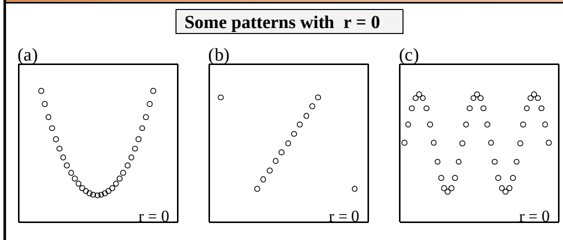
## Recall the correlation coefficient…

Another form for the correlation coefficient is:

$$R(X;Y) = Corr(X;Y) = \frac{\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right]}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\times\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}} =$$

$$\frac{\sum_{i=1}^{n}\left[y_i x_i + \bar{x}\bar{y} - \bar{y}x_i - \bar{x}y_i\right]}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}} = \frac{\sum_{i=1}^{n}\left(y_i x_i\right) - n\times\bar{x}\times\bar{y}}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\times\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}}$$

## Misuse of the correlation coefficient




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

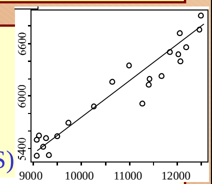


## Linear Regression

- Regression relationship = trend + residual scatter

$$\hat{y} = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x} + \underbrace{Err}$$

- Trend=best linear fit Line (LS)




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

- Scatter = residual (prediction) error Err=Obs-Pred

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = (y_1-\hat{y}_1)^2 + (y_2-\hat{y}_2)^2 + \ldots + (y_n-\hat{y}_n)^2$$



## Another Notation for the Slope of the LS line

1. Note that there is a slight difference in the formula for the slope of the Least Squares Best-Linear Fit line:

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

$$\hat{\beta}_1 = Corr(X;Y)\times\frac{SD(Y)}{SD(X)}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$



## Another Notation for the Slope of the LS line

$$\hat{\beta}_1^{New} = Corr(X;Y)\times\frac{SD(Y)}{SD(X)} =$$

$$\frac{\sum_{i=1}^{n}[(x_i-\bar{x})(y_i-\bar{y})]}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\times\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}} \times \frac{\sqrt{\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)\frac{1}{N-1}}}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\frac{1}{N-1}}} =$$

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



## 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 $\bar{X}$
3. Confidence intervals and parameter estimation
4. Hypothesis testing
5. Paired vs. Independent samples
6. Chi-Square ($\chi^2$) Goodness-of-fit Test
7. Analysis Of Variance (1-way-ANOVA, one categorical var.)
8. Correlation and regression
9. Best-linear-fit, least squares method