UCLA STAT XL 10
Introduction to Statistical Reasoning

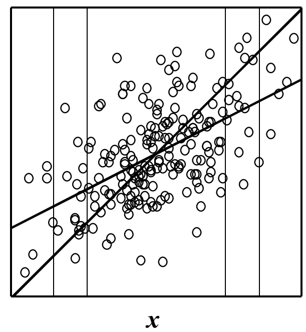
- **Instructor**: **Ivo Dinov,**

Asst. Prof. In Statistics and Neurology

University of California, Los Angeles, Spring 2002

http://www.stat.ucla.edu/~dinov/

The Basic Problems – When, Why, How?

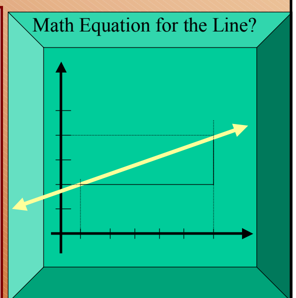


1. Identify the Trend of the Data
2. Predict Future Observations.

Chapters 7-10: 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?

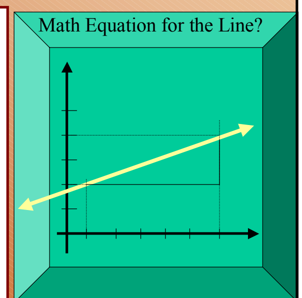


Chapters 7-10: Lines in 2D
(Regression and Correlation)

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

  $(\mathbf{X}-X_1)/(X_2-X_1)$.

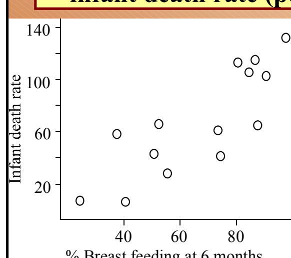
Math Equation for the Line?



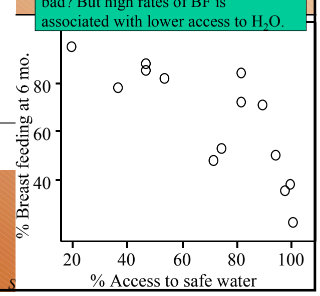
Approaches for modeling data relationships
Regression and Correlation

- 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

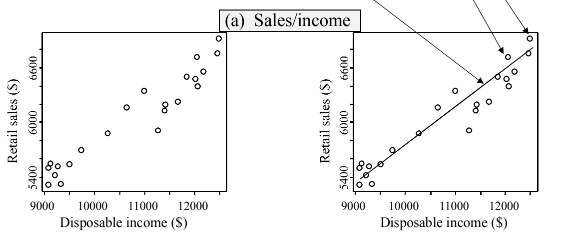


Strong evidence (linear pattern) of death rate increase with increasing level of breastfeeding (BF)? Naïve conclusion breast feeding is bad? But high rates of BF is associated with lower access to H_2O.



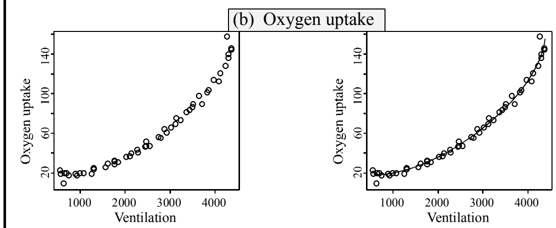
Predict behavior of Y (response) Based on the values of X (explanatory var.) Strategies for uncovering the reasons (causes) for an observed effect.

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).

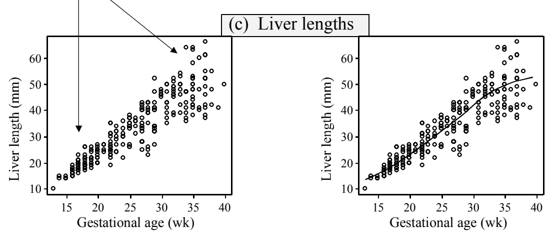
Trend (does not have to be linear) + scatter (could be of any type/distribution)



From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 1999.

Trend + scatter (fetus liver length in *mm*)

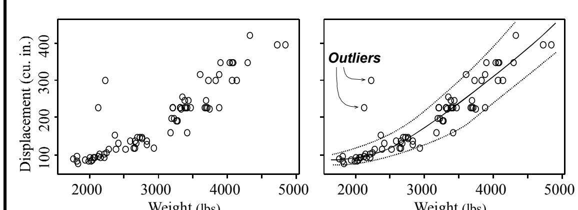
Change of scatter with age



From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 1999.

Trend + scatter

Dotted curves (confidence intervals) represent the extend of the scatter.



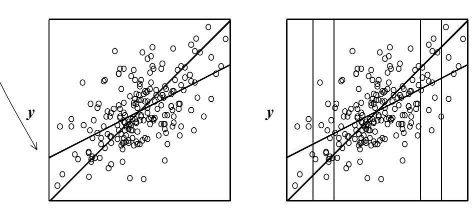
(a) Scatter plot (b) With trend plus scatter

Displacement versus weight for 74 models of automobile.

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

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)

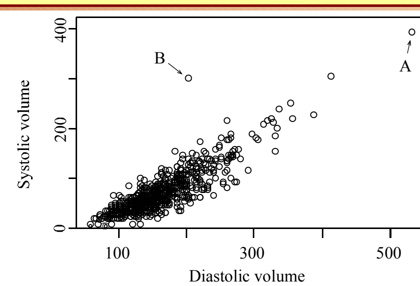


(a) Which line? (b) Flatter line gives better predictions.

Educating the eye to look vertically.

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

Outliers – odd, atypical, observations (errors, B, or real data, A)



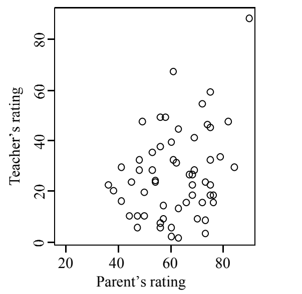
Scatter plot from the heart attack data.

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

A weak relationship

58 abused children are rated (by non-abusive parents and teachers) on a psychological disturbance measure.

How do we quantify weak vs. strong relationship?



Parent's rating versus teacher's rating for abused children.

A note of caution!

In observational data, strong relationships are ***not*** necessarily causal. It is virtually **impossible** to conclude a cause-and-effect relationship between variables using observational data!

Essential Points

1. What essential difference is there between the correlation and regression approaches to a relationship between two variables? (In correlation independent variables; regression response var depends on explanatory variable.)
2. What are the most common reasons why people fit regression models to data? (predict Y or unravel reasons/causes of behavior.)
3. Can you conclude that changes in *X* caused the changes in *Y* seen in a scatter plot if you have data from an observational study? (No, there could be lurking variables, hidden effects/predictors, also associated with the predictor X, itself, e.g., time is often a lurking variable, or may be that changes in Y cause changes in X, instead of the other way around).

Essential Points

5. When can you reliably conclude that changes in *X* cause the changes in *Y*? (Only when controlled randomized experiments are used – levels of X are randomly distributed to available experimental units, or experimental conditions need to be identical for different levels of X, this includes time.

Correlation Coefficient

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

Relationship between two variables (X, Y) can be summarized by: (μ_X, σ_X), (μ_Y, σ_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)

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

$X=\{x_1, x_2, \ldots, x_N\}$
$Y=\{y_1, y_2, \ldots, y_N\}$
(μ_X, σ_X), (μ_Y, σ_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	≈ 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

$$R(X,Y)=\frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{xk-\mu x}{\sigma x}\right)\left(\frac{yk-\mu y}{\sigma y}\right)=$$

$R(aX+b,cY+d)$, since

$$\left(\frac{axk+b-\mu ax+b}{\sigma ax+b}\right)=\left(\frac{axk+b-(a\mu x+b)}{a\times\sigma x}\right)=$$

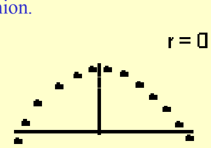
$$\left(\frac{a(xk-\mu)+b-b}{a\times\sigma x}\right)=\left(\frac{xk-\mu x}{\sigma x}\right)$$

Correlation Coefficient - Properties

Correlation is Associative

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

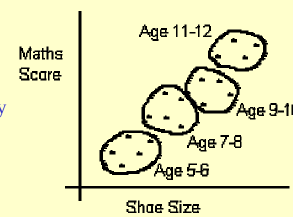
Correlation measures linear association, NOT an association in general!!! So, Corr(X,Y) could be misleading for X & Y related in a non-linear fashion.



Correlation Coefficient - Properties

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

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 <u>which batch</u> (H_20 levels, sand, dry-conditions) gets what <u>amount of cement</u> 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 <u>inverse square law</u>.

c. Economic production function: $Q=\alpha L^{\beta}K^{\gamma}$, Q=production, L=quantity of labor, K=capital, α,β,γ 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).

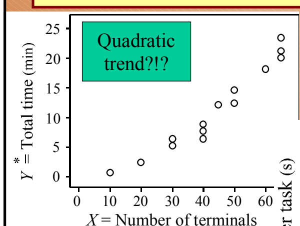
   c. Assuming the model is correct, to estimate unknown constants in the model so that the relationship is completely specified (y=ax+5, a=?)

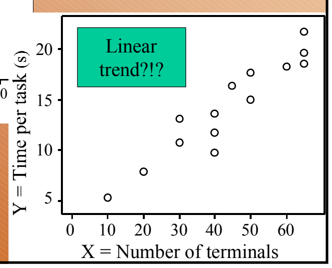
Trend and Scatter - Computer timing data

- 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		

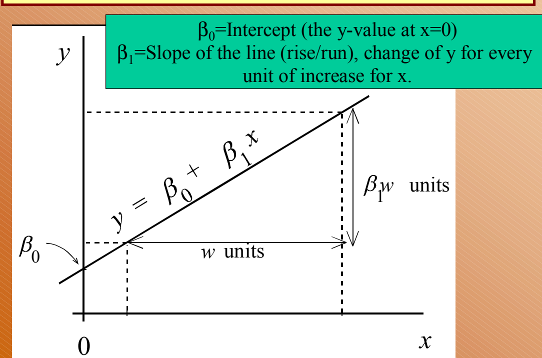
Trend and Scatter - Computer timing data





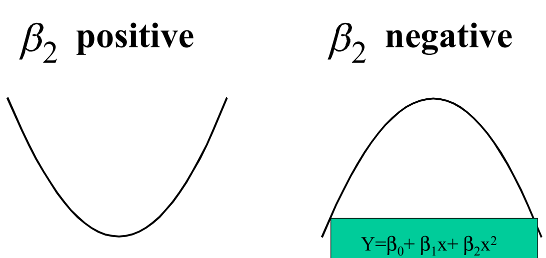
We want to find reasonable models (descriptions) for these data!

Equation for the straight line – linear/affine function



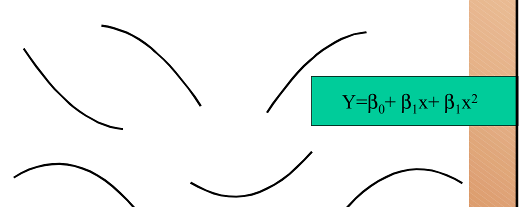
The quadratic curve

Quadratic Curve

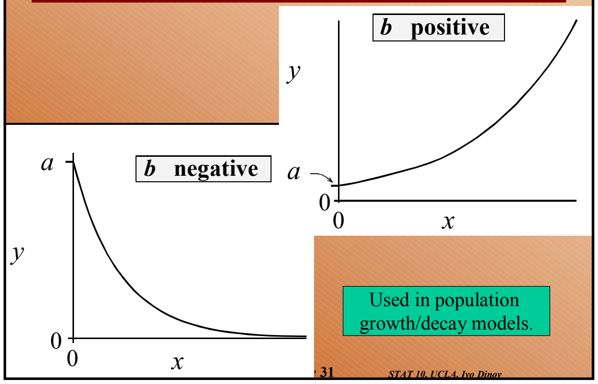


The quadratic curve

Segments of the curve



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



Used in population growth/decay models.

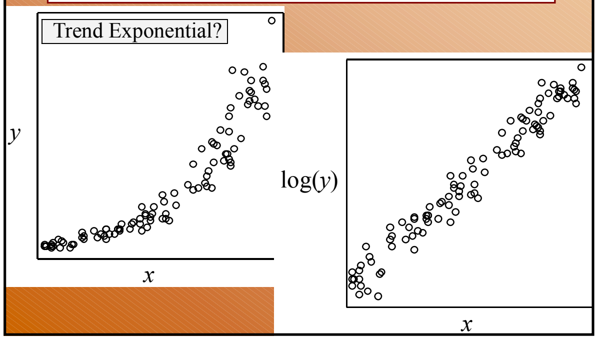
Effects of changing *x* for different functions/curves

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.

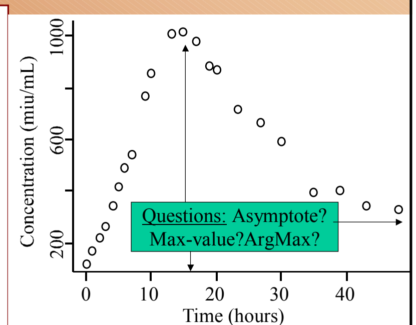


Creatine kinase concentration in patient's blood

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!



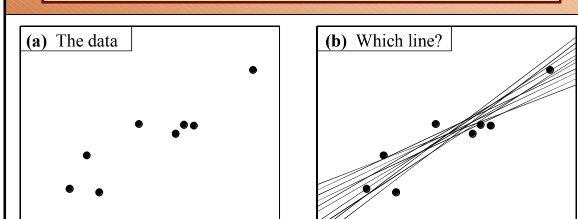
Comments

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)

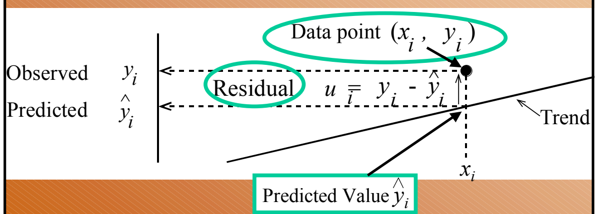
Fitting a line through the data

Choosing the "best-fitting" line



The idea of a residual or prediction error

Show the Regression-Line Simulation Applet:
RegressionApplet.html



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),

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

For each point $P_1(x_1,y_1)$, $P_2(x_2,y_2)$, … , $P_n(x_n,y_n)$.

The least squares line

Least-squares line

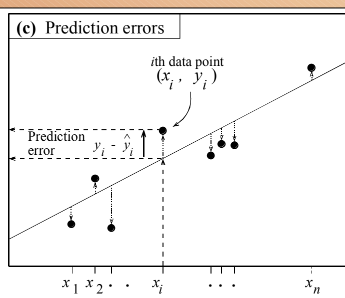
Choose line with smallest sum of squared prediction errors

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

Its parameters are denoted:

Intercept: $\hat{\beta}_0$

Slope: $\hat{\beta}_1$



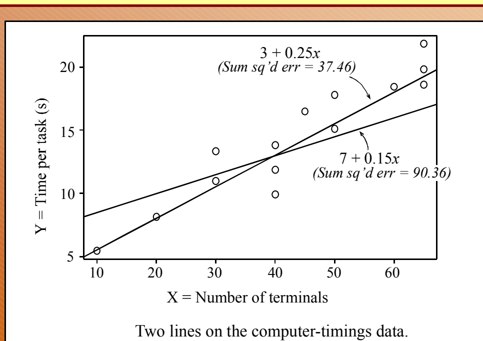
Least-squares line: $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$

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



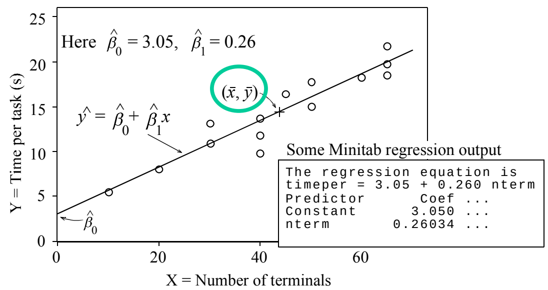
Two lines on the computer-timings data.

Computer timings data

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



Computer-timings data with least-squares line.

Review

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.

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

Review

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						

Review

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	

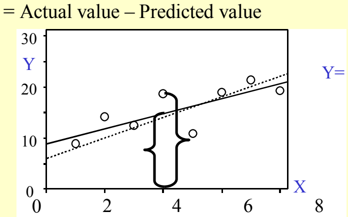
β_1=5/14
β_0=y^-β1*x^
β_0= 0.5-10/14

Review

1. What are the quantities that specify a particular line?
2. Explain the idea of a prediction error in the context of fitting a line to a scatter plot. To what visual feature on the plot does a prediction error correspond?
3. What property is satisfied by the line that fits the data best in the least-squares sense?
4. 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.

RMS Error for regression

- Error = Actual value – Predicted value



Y= $\beta_0 + \beta_1$ X

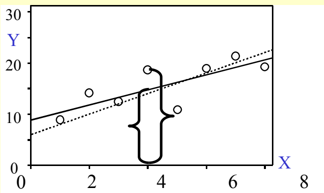
- The RMS Error for the regression line Y= $\beta_0 + \beta_1$ X is

$$\sqrt{\frac{(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2+(y_4-\hat{y}_4)^2+(y_5-\hat{y}_5)^2}{5-1}}$$

where $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k, \quad 1 \le k \le 5$

Compute the RMS Error for this regression line

- Error = Actual value – Predicted value



X	Y
1	9
2	15
3	12
4	19
5	11
6	20
7	22
8	18

- The RMS Error for the regression line $Y = \beta_0 + \beta_1 X$ is

$$\sqrt{\frac{(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2+(y_4-\hat{y}_4)^2+(y_5-\hat{y}_5)^2}{5-1}}$$

where $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k, \quad 1 \le k \le 5$

Slide 49 STAT 10, UCLA, Ivo Dinov

Compute the RMS Error for this regression line

- Error = Actual value – Predicted value
- The RMS Error for the regression line $Y = \beta_0 + \beta_1 X$ is

$$\sqrt{\frac{(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2+(y_4-\hat{y}_4)^2+(y_5-\hat{y}_5)^2}{5-1}}$$

where $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k, \quad 1 \le k \le 5$

- First compute the LS linear fit (estimate β_0^ + β_1^)
- Then Compute the individual errors
- Finally compute the cumulative RMS measure.

X	Y
1	9
2	15
3	12
4	19
5	11
6	20
7	22
8	18

Slide 50 STAT 10, UCLA, Ivo Dinov

Compute the RMS Error for this regression line

- First compute the LS linear fit (estimate β_0^ +β_1^),μ_X=4.5,μ_Y=15.75

X	Y	X-μ_X	Y-μ_Y	(X-μ_X)^2	(Y-μ_Y)^2	(X-μ_X)*(Y-μ_Y)
1	9					
2	15					
3	12					
4	19					
5	11					
6	20					
7	22					
8	18					
Total:						

- Compute

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

Slide 51 STAT 10, UCLA, Ivo Dinov

Compute the RMS Error for this regression line

- Then Compute the individual errors

$(y_k - \hat{y}_k)^2$, where $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k, \quad 1 \le k \le 8$

- Finally compute the cumulative RMS measure.

$$\sqrt{\frac{(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2+(y_4-\hat{y}_4)^2+(y_5-\hat{y}_5)^2}{5-1}}$$

where $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k, \quad 1 \le k \le 5$

X	Y
1	9
2	15
3	12
4	19
5	11
6	20
7	22
8	18

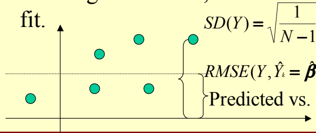
Slide 52 STAT 10, UCLA, Ivo Dinov

Compute the RMS Error for this regression line

- The RMS Error for the regression line $Y = \beta_0 + \beta_1 X$ says how far away from the (model/predicting) regression line is each observation.
- Observe that the SD(Y) is also a RMS Error measure of another specific line – horizontal line through the average of the Y values. This line may also be taken for a regression line, but often it's not the best linear fit.

$$SD(Y) = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(Y_k-\bar{Y})^2} \quad \text{vs.}$$

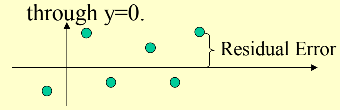
$$RMSE(Y,\hat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 X) = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(Y_k-\hat{Y}_k)^2}$$



Slide 53 STAT 10, UCLA, Ivo Dinov

Plotting the Residuals

- The Residuals=Observed –Predicted for the regression line $Y = \beta_0 + \beta_1 X$ (just like the error).
- Residuals average to zero, mathematically, and the regression line for the residuals is a horizontal line through y=0.



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

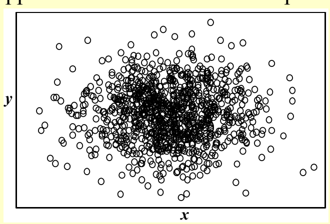
when $X = x$, $\mathbf{Y = \beta_0 + \beta_1 x + U}$, where $U \sim \text{Normal}(0, \sigma)$

Random error

Slide 54 STAT 10, UCLA, Ivo Dinov

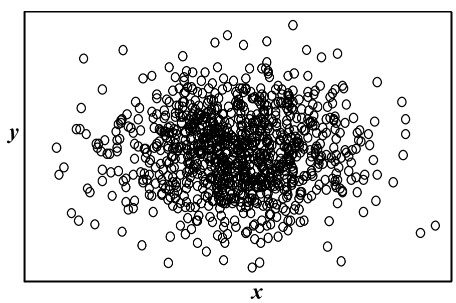
Plotting the Residuals – patterns?

- The Residuals=Observed –Predicted for the regression line $Y = \beta_0 + \beta_1 X + U$ should show no clear trend or pattern, for our linear model to be a good and useful approximation to the unknown process.



Is there always an X Y relationship? Linear Relationship?

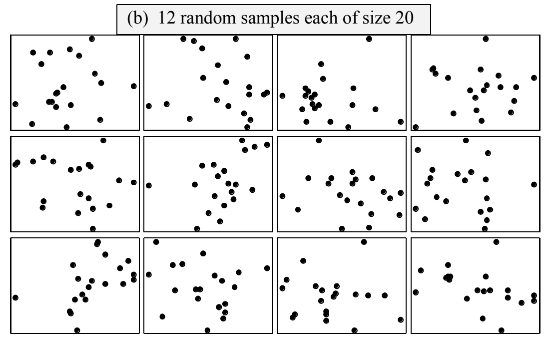
(a) 1000 data points with no relationship between *X* and *Y*



From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 1999.

Random samples from these 1000 data points

(b) 12 random samples each of size 20



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

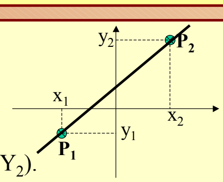
Review

1. Describe a fundamental difference between the way regression treats data and the way correlation treats data.
2. What is the correlation coefficient intended to measure?
3. For what shape(s) of trend in a scatter plot does it make sense to calculate a correlation coefficient?
4. What is the meaning of a correlation coefficient of $r = +1$? $r = -1$? $r = 0$?

Summary

Lines in the Plane

- Draw the following lines:
- Y = 3.4X + 13
- Y = –3X — 5.7
- Line through (X_1,Y_1) and (X_2,Y_2).



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

Concepts

- Relationships between quantitative variables should be explored using **scatter plots**.
  - Usually the *Y* variable is continuous
    (or behaves like one in that there are few repeated values)
  - and the *X* variable is discrete or continuous.
- **Regression** singles out one variable (***Y***) as the response and uses the explanatory variable (***X***) to explain or predict its behavior.
- **Correlation** treats both variables symmetrically as random.

Concepts cont.

In practical problems, regression models may be fitted for any of the following reasons:

- To understand a **causal relationship** better. Ex?
- To find relationships which may be **causal**. Ex?
- To make **predictions**. Ex?
  - But be cautious about predicting outside the range of the data
- To **test theories**. Ex?
- To **estimate parameters** in a theoretical model.

Concepts cont.

- In observational data, strong relationships are not necessarily causal.
- We can only have reliable evidence of causation from controlled, randomized, designed experiments.
- Be aware of the possibility of **lurking** variables which may effect both *X* and *Y*.

Concepts cont.

- The two main approaches to summarizing trends in data are using *smoothing* and *fitting models* (e.g., regression lines).
- The ***least-squares criterion*** for fitting a mathematical curve is to choose the values of the parameters (e.g. β_0 and β_1) to minimize the sum of squared prediction errors, $\sum (y_i - \hat{y}_i)^2$.

Linear Relationship

- We fit the linear relationship $\hat{y} = \beta_0 + \beta_1 x$.
- The slope β_1 is the change in $\hat{y}$ associated with a one-unit increase in *x*.

Least-squares estimates

- The least-squares estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to minimize $\sum (y_i - \hat{y}_i)^2$.
- The **least-squares regression line** is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Residuals and outliers

- These assumptions should be checked using residual plots. The *i*-th *residual* (or *prediction error*) is

$$y_i - \hat{y}_i = \text{observed - predicted.}$$

- An ***outlier*** is a data point with an unexpectedly large residual (positive or negative).

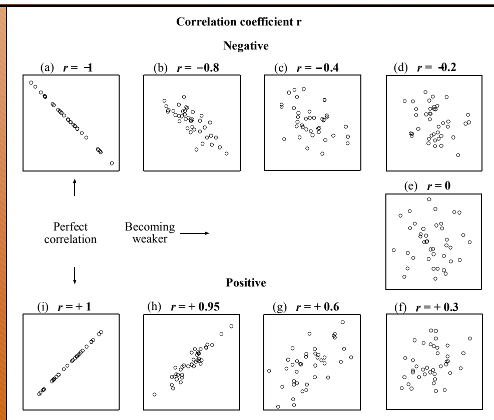
Correlation coefficient

The correlation coefficient *r* is a measure of linear association with $-1 \le r \le 1$.

- If **r = 1**, then X and Y have a perfect positive linear relationship.
- If **r = –1**, then X and Y have a perfect negative linear relationship.
- If **r = 0**, then there is no linear relationship between X and Y.
- Correlation does not necessarily imply causation.

Correlation coefficient – interpret the following. Give examples!

- Correlation is invariant w.r.t. linear transformations of X or Y.
- Correlation is Associative.
- Correlation measures linear association, NOT an association in general!!!



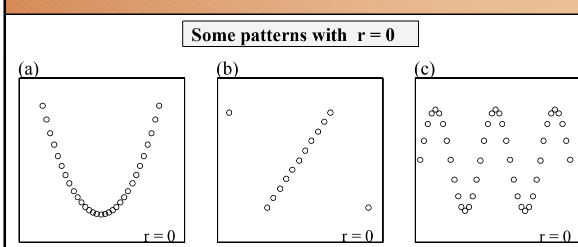
From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Recall the correlation coefficient…

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

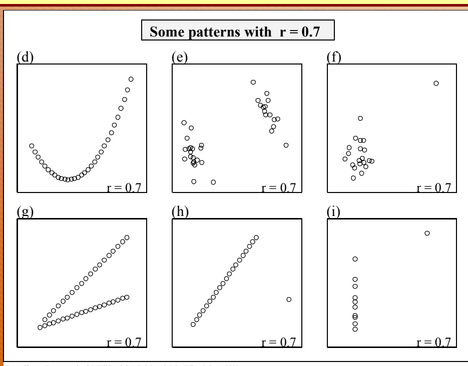
$$\frac{\sum_{i=1}^{n}\left[y_i x_i + \bar{x}\bar{y} - \bar{y}x_i - \bar{x}y_i\right]}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\times\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}} = \frac{\sum_{i=1}^{n}(y_i x_i) \; - \; n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\times\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}}$$

Misuse of the correlation coefficient



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

Misuse of the correlation coefficient



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

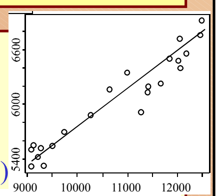
Correlation does not necessarily imply causation.

Linear Regression

- Regression relationship = trend + residual scatter

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

- 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**

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

Textbook vs. Lecture Notation …

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

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

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

Textbook vs. Lecture Notation …

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

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

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

Redo the problem from last time using:

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						

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

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

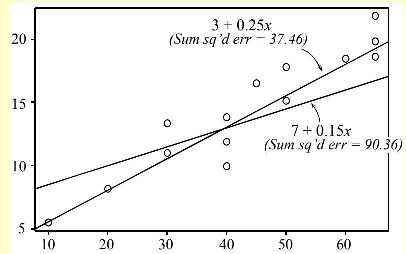
Redo the problem from last time using:

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	

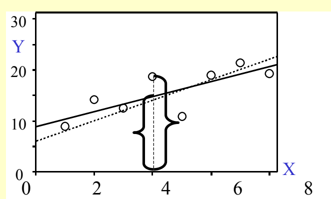
Properties of Linear Regression

- Linear Fit that minimizes the sum-square error of $\hat{y} = \beta_0 + \beta_1 x$ obs. vs. predicted values: $\sum (y_i - \hat{y}_i)^2$.



Properties of Linear Regression

- The points ($x = 0$, $y=\hat{\beta}_0$) and ($x = \bar{x}$, $\hat{y} = \bar{y}$) lie on the LS line.
- RMS error – indicates how far are typical points from the regression line (up/down)



$$\sqrt{\frac{(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2+(y_4-\hat{y}_4)^2+(y_5-\hat{y}_5)^2}{5-1}}$$

where $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k, \quad 1 \le k \le 5$