PubH 7405:
REGRESSION ANALYSIS

Review #2:

Simple Correlation & Regression

COURSE INFORMATION

- **Course Information are at address www.biostat.umn.edu/~chap/pubh7405**
- **On each class web page, there is a brief version of the lecture for the day – the part with "formulas"; you can review, preview, or both – and as often as you like.**
- Follow **Reading & Homework** assignments **at the end** of the page when & if applicable.

OFFICE HOURS

- **Instructor's scheduled office hours: 1:15 to 2:15 Monday & Wednesday, in A441 Mayo Building**
- **Other times are available by appointment**
- **When really needed, can just drop in and interrupt me; could call before coming – making sure that I'm in.**
- **I'm at a research facility on Fridays.**

Variables

- A variable represents a characteristic or a class of measurement. It takes on different values on different subjects/persons. Examples include weight, height, race, sex, SBP, etc. The observed values, also called "observations," form items of a data set.
- Depending on the scale of measurement, we have different types of data.

There are "observed variables" (Height, Weight, etc… each takes on different values on different subjects/person) and there are "calculated variables" (Sample Mean, Sample Proportion, etc… each is a "statistic" and each takes on different values on different samples). The Standard Deviation of a calculated variable is called the Standard Error of that variable/statistic.

A “variable” – sample mean, sample standard deviation, etc… included – is like a “function”; when you apply it to a target element in its domain, the result is a “number”. For example, “height” is a variable and “the height of Mrs. X” is 135 lbs; it’s a number.

TYPES OF DATA

- There are binary or dichotomous outcomes, e.g. Sex/gender (male/female), Morbidity (sick/well)
- There are categorical or polytomous outcomes, eg. Race (white/black/Hispanics/Asian)
- There are continuous outcomes, e.g. blood pressure, cholesterol level); of course, you can dichotomized or categorized a continuous outcome to make it binary or categorical – but some information are lost in the process.

In most problem involving statistical inference, we investigate **one variable at a time**. However, in many important investigations, we may have two measurements made on each subject, and the research objective is concerned not with each of them but with the **relationship** between them.

AN EXAMPLE: IN SEARCH OF AN HONEST EMPLOYEE

- Shoplifting is a big problem, it costs up to 2 billions dollars a year in America
- Who done it? Customers?
- Yes, but customer shoplifting ranks second to employee theft which involves between 2% and 3% of all employees.

SOLUTION?

- One approach to curtailing employee theft is screen job applicants so as not to hire those with “high potential” to theft.
- How to do it? How about using polygraph test (lie detector)?
- But **who** want to apply? you need to treat your future employee with dignity!

AN ALTERNATIVE

- May be a less visible pencil-and-paper test as part of the application.
- **Need:** to device some kind of a questionnaire; but its “score” should be “highly correlated” to to the result by the polygraph test.

ANOTHER: RESEARCH IN AN AMUSEMENT PARK?

- Yes, they do it for business planning: designing questionnaires, selecting samples, conducting interviews, and analyzing data that provide information about visitors' attitudes, perceptions, and preferences.
- **Information** about visitors themselves, where they come from and why they came.
- Results would be variety of plans, strategies, and decisions on how to draw visitors to the park & make them to spend more.

SOME OTHER INTERESTING RELATIONSHIPS

- Height and Weight
- Age and Blood Pressure
- Daily Fat Intake and cholesterol Level
- Daily Salt Intake and Blood Pressure
- Weight Gain during pregnancy and Birth weight
- Time to engraftment and time to infection in BMT.
- White Blood Count and a leukemia patient's Survival Time from diagnosis.

EASY WAY OUT?

- We could dichotomize both variables and use the Odds Ratio; for example, Daily Salt Intake (Above/Below average) versus High Blood Pressure (yes/No).
- But by doing so, we would lose the details and the "power" (it always take more data to deal with dichotomous variables!)
- Instead, you learned how to deal with the relationship between continuous variables.

SUB-TYPES OF ANALYSES

- We have have measurements made on each subject, one is the response variable Y, the other **predictor** X. There are two types of analyses:
- **Correlation:** is concerned with the association between them, measuring the strength of the relationship. For example, Is a woman's Age and her SBP related? **How strong** is the relationship?
- **Regression:** To **predict** response from predictor. For example, Is a woman's Age predictive of her SBP? Or Is a woman's Weight Gain during pregnancy predictive of her newborn's Birth Weight?

AN EXAMPLE

Trace metals in drinking water affect the flavor and may pose a health hazard. The following Table shows concentration of Zinc (in mg/l) for both surface (X) and bottom (Y) water at 6 river location. Can we predict bottom water concentration (which is harder to measure) from surface water concentration (which is easier to measure) so that in a continuous monitoring system we can only measure from the surface water? Regression may be needed here.

POLLUTION DATA

Concentration of Zinc (in mg/l) measured at six (6) river locations, both from surface water and bottom water.

Location	Bottom	Surface
1	0.430	0.415
2	0.266	0.238
3	0.567	0.390
4	0.531	0.410
5	0.707	0.605
6	0.716	0.609

NUTRITION AND “IMR”

The following Table gives “Net Food Supply” (X, in number of calories per person per day) and the “**Infant Mortality Rate**” (Y=IMR, number of infant deaths per 1000 live births). Data are listed for 22 selected countries (each country is an unit of observation); data were ontained before World War I (current IMRs are much lower; for USA: current figure is about 11). Are X and Y related? Maybe it’s just a problem of Correlation.

INFANT MORTALITY DATA

Country	x	y	Country	x	y
Argentina	2730	98.8	Iceland	3160	42.4
Australia	3300	39.1	India	1970	161.6
Austria	2990	87.4	Iceland	3390	69.6
Belgium	3000	83.1	Italy	2510	102.7
Burma	1080	202.1	Japan	2180	60.6
Canada	3070	67.4	New Zealand	3260	32.2
Chile	2240	240.8	Netherlands	3010	37.4
Cuba	2610	116.8	Sweden	3210	43.3
Egypt	2450	162.9	england	3100	55.3
France	2880	66.1	USA	3150	53.2
Germany	2960	63.3	Uruguay	2380	94.1

N = 22 Countries

X = Calories Per Person Per Day

Y = Infant Mortality Rate (IMR, Deaths per 1000 Live Births)

Birth weight data

n = 12

X = Birth weight (oz)

Y = Growth in weight between 70th and 100th days of life, as % of birth weight.

It's both: Correlation & Regression

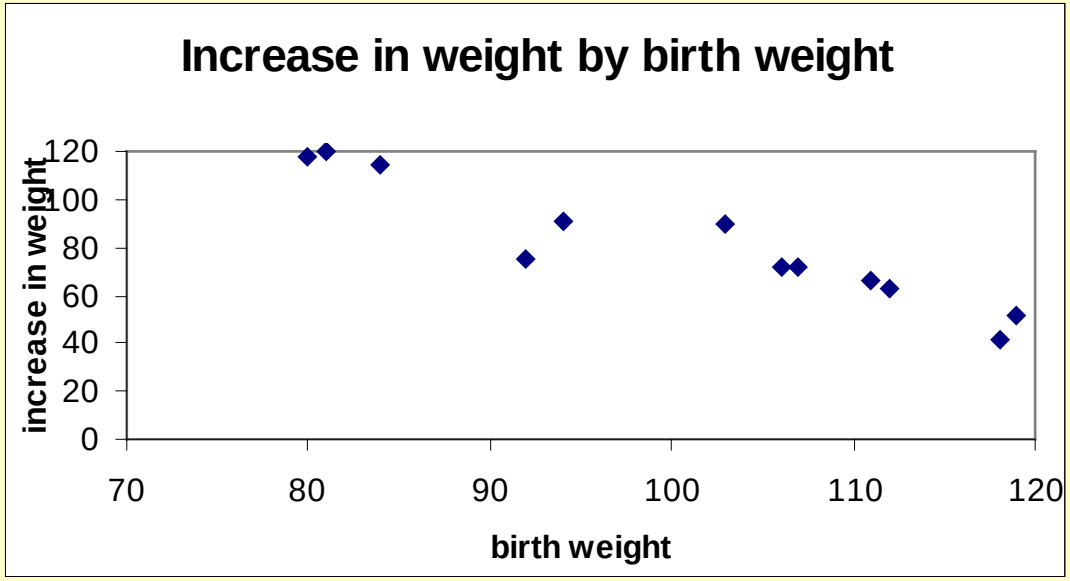
x (oz)	y (%)
112	63
111	66
107	72
119	52
92	75
80	118
81	120
84	114
118	42
106	72
103	90
94	91

MORE APPLICATIONS OF REGRESSION

- Sales of a product could be predicted from amount of advertising expenditures.
- The performance of an employee could be predicted from a battery of tests.
- The size of vocabulary of a child could be predicted from the age of the child and levels of education of parents.
- The length of hospital stay could be predicted from the severity of the operation.

Scatter Diagram

If we let each pair of numbers (*x*,*y*) be represented by a dot in a diagram with the *x*'s on the horizontal axis, we have the figure shown below:

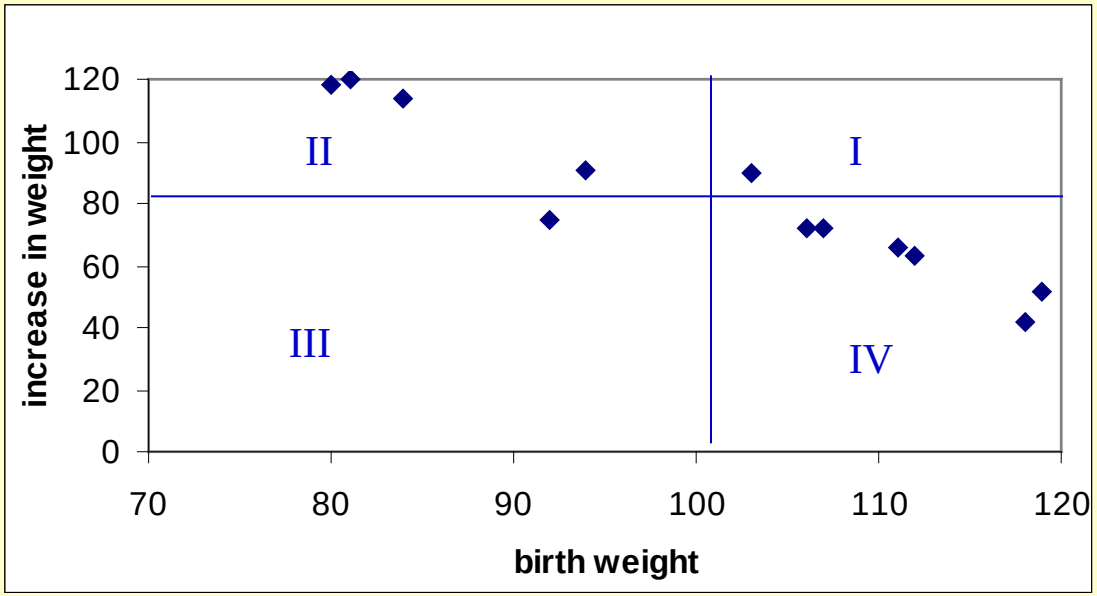


<u>EXAMPLE</u>: For the first child x = 112 oz. & y = 63% (3rd dot from right)

About SCATTER DIAGRAM

- **The dots in a Scatter Diagram do not fall perfectly on a straight line, very typical of a "statistical relationship"-not "deterministic relationship".**
- **The positions of the dots provide information about "direction" as well as the strength of the association.**
- **Positive** association: dots go lower left to upper right
- **Negative** association: dots go upper left to lower right
- **Strong** association: dots are clustered closer to line.
- **Weaker** association: less clustered, form a circle.

ANALYSIS of Scatter Diagram



In this figure, we will draw a vertical line and a horizontal line intersecting at the point $(\overline{x}, \overline{y})$. Together these two lines divide the page into **four quarters**, labeled as I,II,III and IV.

SCATTER DIAGRAM

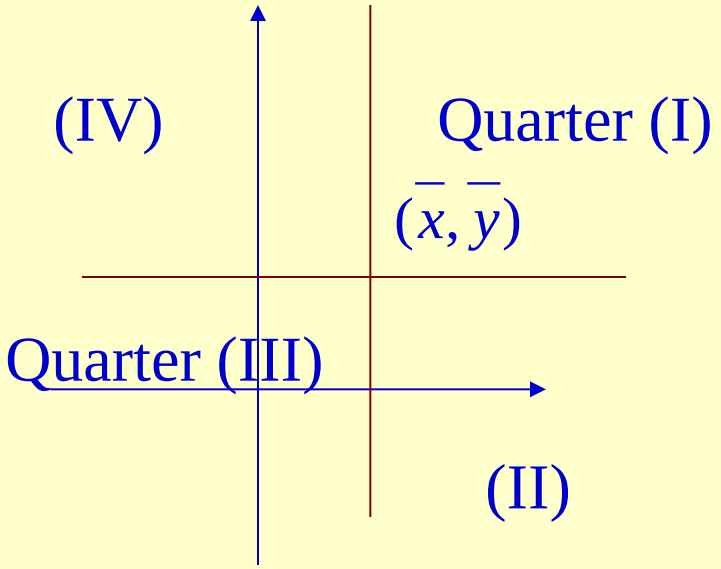
- In quarters I and III,

$$(x-\bar{x})(y-\bar{y})>0$$

- For positive association,

$$\sum(x-\bar{x})(y-\bar{y})>0$$

- In addition, for stronger relationship most of the dots, being closely clustered around the line, are in these two quarters; the above sum is large.



SCATTER DIAGRAM

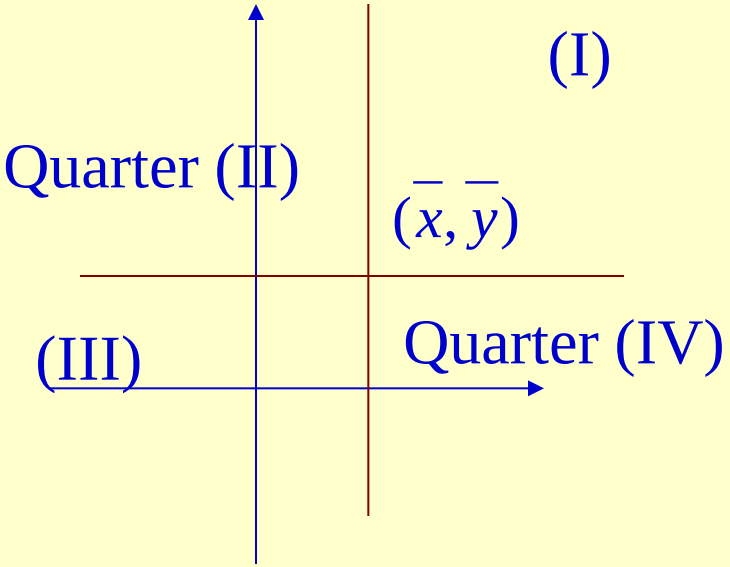
- In quarters II and IV,

$$(x-\bar{x})(y-\bar{y})<0$$

- For negative association,

$$\sum(x-\bar{x})(y-\bar{y})<0$$

- In addition, for stronger relationship most of the dots, being closely clustered around the line, are in these two quarters; the sum is a large negative number.



SUMMARY

- The sum $\sum(x-\bar{x})(y-\bar{y})$ summarizes the "evidence" of the relationship under investigation; It is zero or near zero for weak associations and is large, negative or positive, for stronger associations.
- However, it is "unbounded" making it hard to use because we cannot tell if we have a strong association (how large is "large"?).
- We need to "standardize" it.

COEFFICIENT OF CORRELATION

- With a standardization, we obtain:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2][\sum (y - \bar{y})^2]}}$$

so that $-1 \leq r \leq 1$. The statistic r is called the Correlation Coefficient measuring the strength of the relationship; and here is a "short-cut" formula

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{[\sum x^2 - \frac{(\sum x)^2}{n}][\sum y^2 - \frac{(\sum y)^2}{n}]}}$$

OUR LIMITATION

- We focus only on “linear relationship”, where the dots in the scatter diagram are clustered around a straight line (so, we’ll emphasize on its “slope”).
- There are more complicated “patterns” of association. For example, the dots may cluster around a “curve”, such as a parabola.
- It may seem too restrictive to focus only on linear relationships; fortunately, many real-life applications fit this pattern.

A SMALL EXAMPLE

	x	y	x^2	y^2	xy
	1	3	1	9	3
	2	5	4	25	10
	6	7	36	49	42
Totals	9	15	41	83	55

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\{\sum x^2 - \frac{(\sum x)^2}{n}\}\{\sum y^2 - \frac{(\sum y)^2}{n}\}}} = \frac{55 - \frac{(9)(15)}{3}}{\sqrt{\{41 - \frac{9^2}{3}\}\{83 - \frac{15^2}{3}\}}} = .945$$

	x (oz)	y (%)	x-sq	y-sq	xy
	112	63	12544	3969	7056
	111	66	12321	4356	7326
	107	72	11449	5184	7704
	119	52	14161	2704	6188
	92	75	8464	5625	6900
	80	118	6400	13924	9440
	81	120	6561	14400	9720
	84	114	7056	12996	9576
	118	42	13924	1764	4956
	106	72	11236	5184	7632
	103	90	10609	8100	9270
	94	91	8836	8281	8554
Totals	1207	975	123561	86487	94322

ANALYSIS OF BIRTH WEIGHT DATA

Coefficient of Correlation:

GROWTH Versus BIRTH WEIGHT

Using these five total, we obtain

$$r = \frac{94{,}322 - \frac{(1{,}207)(975)}{12}}{\sqrt{\left[123{,}561 - \frac{1207^2}{12}\right]\left[86{,}487 - \frac{975^2}{12}\right]}}$$

$$= -.946$$

Indicating a **very strong** negative association.

INTERPRETATION

- Values near +1 indicate a strong positive association
- Values near -1 indicate a strong negative association
- Values around 0 indicate a rather weak association.

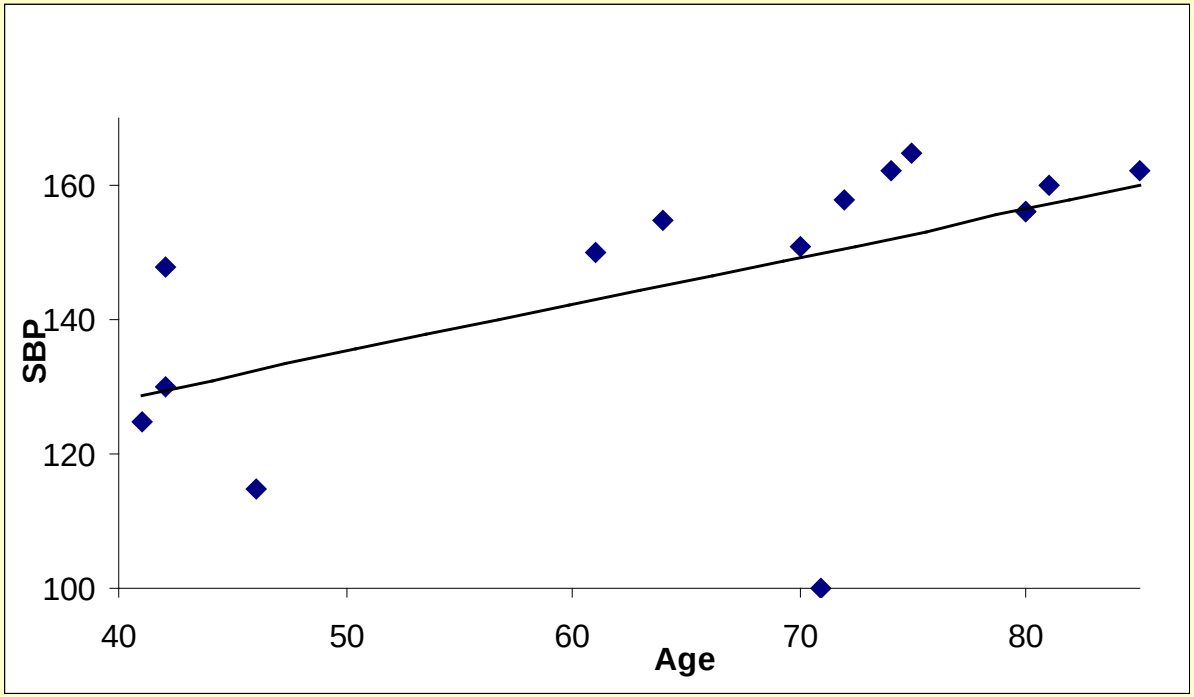
Caution:

- A correlation of 0 does not mean no association, it means no linear association, we assume a straight line relationship to start with. **You may have a correlation near zero and yet a strong relationship**; but **not a linear strong relationship**.
- Check the scatter diagram!

Another example: Age and SBP

n = 15, X = AGE (Years), Y = Systolic Blood Pressure (mm of Hg)

Age (x)	SBP (y)
42	130
46	115
42	148
71	100
80	156
74	162
70	151
80	156
85	162
72	158
64	155
81	160
41	125
61	150
75	165



SBP versus AGE

Analysis:

SBP versus AGE

Age (x)	SBP (y)	x-sq	y-sq	xy
42	130	1764	16900	5460
46	115	2116	13225	5290
42	148	1764	21904	6216
71	100	5041	10000	7100
80	156	6400	24336	12480
74	162	5476	26244	11988
70	151	4900	22801	10570
80	156	6400	24336	12480
85	162	7225	26244	13770
72	158	5184	24964	11376
64	155	4096	24025	9920
81	160	6561	25600	12960
41	125	1681	15625	5125
61	150	3721	22500	9150
75	165	5625	27225	12375
984	2193	67954	325929	146260

Totals

Correlation Coefficient: SBP versus Age

Using these five total, we obtain

$$r = \frac{146{,}260 - \dfrac{(984)(2193)}{15}}{\sqrt{\left[67{,}954 - \dfrac{984^2}{15}\right]\left[325{,}929 - \dfrac{2193^2}{15}\right]}}$$

$$= .564$$

indicating a moderate positive association

COEFFICIENT OF DETERMINATION

- The square of the Coefficient of Correlation r, called the **"Coefficient of Determination"** r^2, when expressed as percentage, represents the proportion of the degree of variation (as measured by the Variance) among the values of one variable which is accounted by its relationship with the other variable.
- Example: 32.04% (square of .564) of variation in SBP among women are due to their different ages.
- This provides a more powerful interpretation for correlation analysis.

HOW STRONG IS A CORRELATION?

- The Coefficient of Determination r^2, when expressed as percentage, represents the proportion of the degree of variation among the values of one variable which is accounted by its relationship with the other variable.
- When $r^2 > 50\%$, one variable is responsible for more than half of the variation in the other; the relationship is obviously strong.
- A correlation with **r >.7** is therefore conventionally considered as a strong.

TESTING FOR INDEPENDENCE

- The Coefficient of Correlation r measures the strength of the relationship between two variables, say the Mother's Weight and her Newborn's Birth Weight. But r is only a Statistic; it is an Estimate of an unknown Population Coefficient of Correlation ρ (rho), the same way the sample $\bar{x}$ is used as an estimate of the Population mean μ.
- The basic question is concerned: H_0: $\rho = 0$; only when H_0 is true,the two variables are not correlated.

Statistics Versus Parameters

- Parameter: A numerical characteristic of a population; parameters are fixed but unknown. Example: population coefficient of correlation ρ
- Statistic: A summarized figure from sample data (used to estimate parameters). Statistics are known but vary from sample to sample. Example: (sample) coefficient of correlation r

Try to separate a “statistic” from a “parameter”. When r = 0, it only imply that values of the two factors, as measured from that sample, are not related. But you can’t generalize that yet (what you found might happen by chance, if you do it again you might not see it again); Only when $\rho = 0$, we can conclude that the factors are not related **- population-wise** .

TESTING FOR INDEPENDENCE

- The Coefficient of Correlation r measures the strength of the relationship between two variables; but as statistic it involves "random variation" in its sampling distribution. We are interested in knowing if we can conclude that: $\rho \neq 0$, that the two variables under investigation are really correlated - not just by chance.
- It is a two-sided Test of the Null Hypothesis of No Association, H_0: $\rho = 0$, against H_A: $\rho \neq 0$ of Real Association (you can do it as one-sided too).

TESTING FOR INDEPENDENCE

- The Test Statistic is:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

- It is the same t-test as used in the comparison of two Population Means; the Degree of Freedom is: df = n-2 (same way to form your rejection decision and to calculate p-value) .

We refer to this as a “t-test” but not as “one-sample t-test”, nor “two-sample t-test” (those later two terms are for the comparison of means). You can call it as the “t-test for independence”.

EXAMPLE #1

- For the Birth-Weight problem, we have: **n=12 and r = -.946** leading to:

$$t = (-.946)\sqrt{\frac{12-2}{1-(-.946)^2}}$$

$$t = -9.23$$

- At α=.05 and df = 10, the tabulated coefficient is 2.228 (2.5% tail) indicating that the Null Hypothesis should be rejected (t=-9.23<-2.228); (two-sided) p-value < .001.

EXAMPLE #2

- For the "SBP vs. Age" problem, we have **n=15 and r = .566** leading to:

$$t = (.566)\sqrt{\frac{15-2}{1-(.566)^2}}$$

$$t = 2.475$$

- At α=.05 and df = 13, the tabulated coefficient is 2.16 (2.5% tail) indicating that the Null Hypothesis should be rejected (t=2.475>2.16); (two-sided) p-value = .028.

EXAMPLE #3

	x	y	x^2	y^2	xy
	1	3	1	9	3
	2	5	4	25	10
	6	7	36	49	42
Totals	9	15	41	83	55

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\{\sum x^2 - \frac{(\sum x)^2}{n}\}\{\sum y^2 - \frac{(\sum y)^2}{n}\}}} = \frac{55 - \frac{(9)(15)}{3}}{\sqrt{\{41 - \frac{9^2}{3}\}\{83 - \frac{15^2}{3}\}}} = .945$$

EXAMPLE #3

x	y	x^2	y^2	xy
1	3	1	9	3
2	5	4	25	10
6	7	36	49	42
9	15	41	83	55

- Here: n = 3, r = .945

$$t = (.945)\sqrt{\frac{3-2}{1-(.945)^2}}$$

$$t = 2.889$$

- At α=.05 and df = 1, the tabulated coefficient is 12.706 (2.5% tail) indicating that the Null Hypothesis should not be rejected (even though r=.945$\neq$0) (t=2.889<12.706); n=3 is the smallest size that we can apply the procedure.

CORRELATION & REGRESSION

- We have have measurements made on each subject, one is the response variable Y, the other predictor X. There are two types of analyses:
- **Correlation:** is concerned with the association between them, measuring the strength of the relationship & test for the Null Hypothesis H_0: $\rho=0$; For example, Is a woman's Age & her SBP related?
- **Regression:** To predict response from predictor. For example, Is a woman's age predictive of her SBP? Or Is a woman's Weight Gain during pregnancy predictive of her newborn's Birth Weight? How?

DIFFERENT ROLES OF VAIRABLES

- X is the “predictor”; also called “explanatory variable” or “independent variable”.
- In a causal relationship, or when they come in sequentially, X comes first - and we place it on the **horizontal axis** of the scatter diagram.
- Y is the “response”; also called “dependent variable” or “outcome variable”.
- In a causal relationship, or when they come in sequentially, Y comes later - and we place it on the **vertical axis** of the scatter diagram.

AN IMPORTANT NOTE

- In Correlation Analysis, the roles of “X” and “Y” are exchangeable; you should note the formula for the coefficient of correlation “r” is symmetric with respect to X and Y (that we get the same result regardless of which one is X).
- In Regression Analysis, each has a well-defined role; we’ll predict “response Y” from (a new) value of “predictor X”

EXAMPLE:

n = 12

X = Birth weight (oz)

Y = Growth in weight between 70th and 100th days of life, as % of birth weight.

Birth weight data:

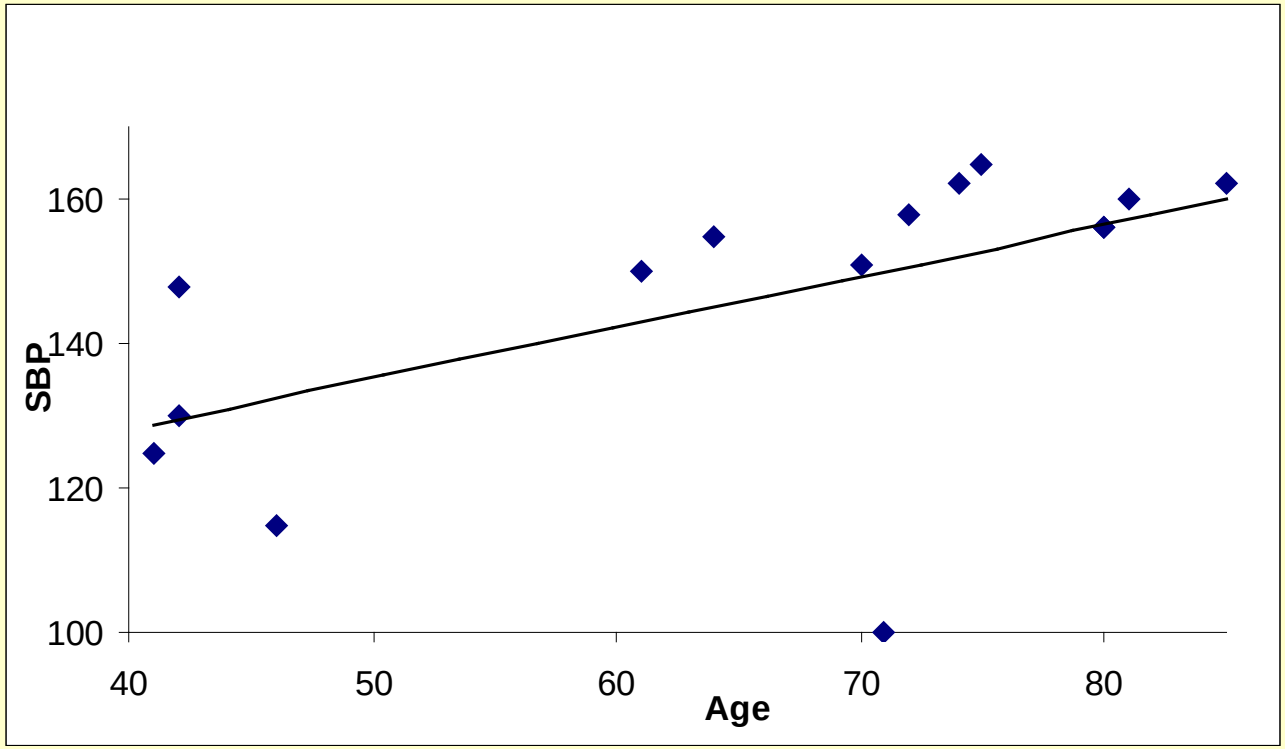
x (oz)	y (%)
112	63
111	66
107	72
119	52
92	75
80	118
81	120
84	114
118	42
106	72
103	90
94	91

Here, the “Birth Weight” is the Predictor, “Growth” the Response

Age and SBP

n = 15, X = AGE (Years), Y = Systolic Blood Pressure) (in mm of Hg)

Age (x)	SBP (y)
42	130
46	115
42	148
71	100
80	156
74	162
70	151
80	156
85	162
72	158
64	155
81	160
41	125
61	150
75	165



SBP versus AGE

(AGE is Predictor, SBP is Response)

RATIONALE FOR PREDICTION

- When the Coefficient of Correlation r is large, so is the Coefficient of Determination.
- If the Coefficient of Determination r^2 is large, say $r^2 = 80\%$, almost all variation among responses are due to different values of its predictor;
- That means we can predict almost precisely the value of the response if we know the value of the predictor. For example, the question could be : what would be a boy's birth weight if his mother gained 37 lbs during her pregnancy?

HISTORICAL ORIGIN

Regression analysis was first developed by Sir Francis Galton in the later part of the 19th century. Galton had studied **the relation between heights of parents and children** and noted that the heights of children of both tall and short parents appeared to "revert" or "regress" to the mean of the group. He considered this tendency be a "regression to mediocrity". The term "regression" persists to this day to describe statistical relations between variables – even nothing "regresses"!.

A functional relation between two variables is expressed by a mathematical formula. For example, if X denotes the independent variable and Y the dependent variable, a a functional relation could be of the form Y = f(X). Given a particular value of X, the function f(.) would give the corresponding value of Y

DETERMINISTIC & STATISTICAL RELATIONSHIPS

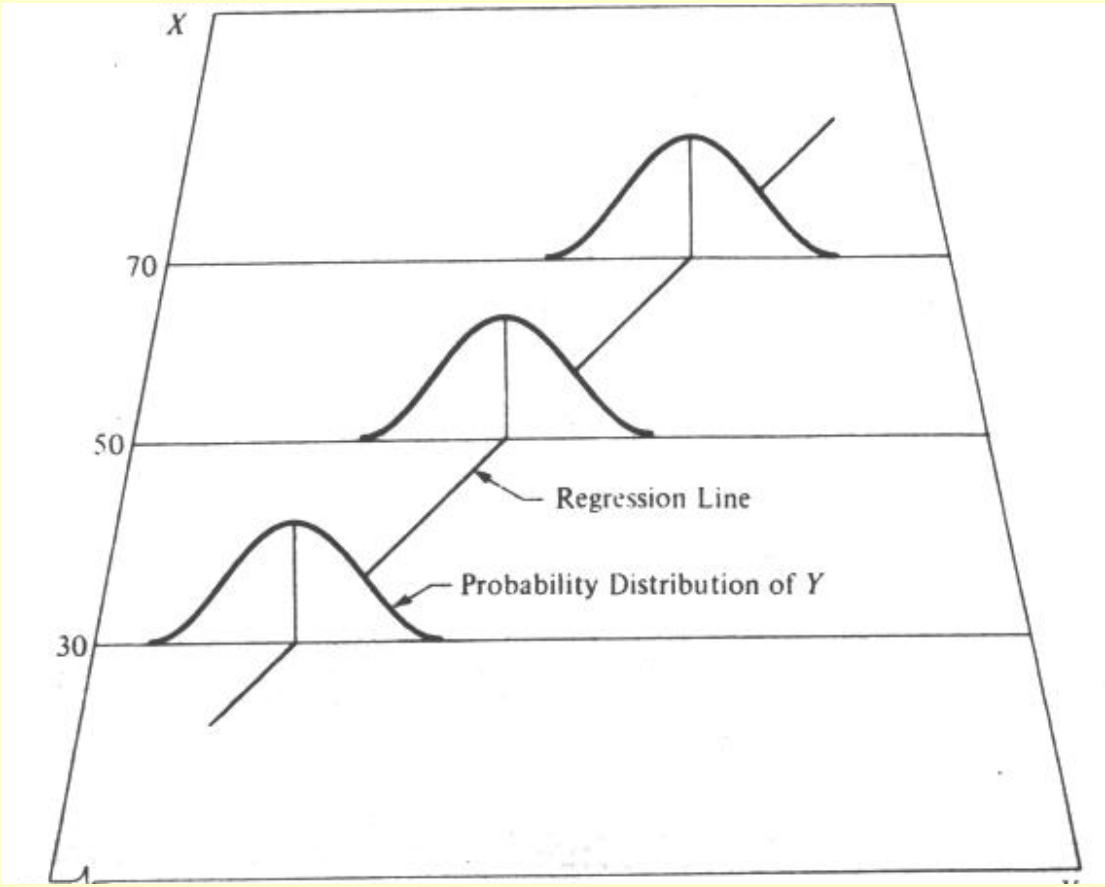
- In a deterministic relation, the value of X determines the value of Y precisely. For example, if the admission costs $100 and each day of stay costs $200; then staying for X=3 days will cost the patient Y=100+(3)(200)= $700.
- A statistical relation is not a perfect one; the observations (i.e. the “dots”) do not fall perfectly on a straight line or a curve – as seen from a scatter diagram.
- Our targets are statistical relations

REGRESSION MODEL

- Let Y be the Response Variable, X the Predictor (also called Explanatory or Independent Variable). For a particular value x of the Predictor X, the values of the Response or Dependent Variable Y is assumed to be “normally distributed”.
- For example, among the mothers who gained 37 lbs during their pregnancies and gave birth to baby boys, the boys’ birth weights may not be all the same but form certain normal distribution.

REGRESSION MODEL

- Let Y be the Response Variable, X the Predictor (also called Explanatory or Independent Variable). The Response or Dependent Variable Y, for the sub-population with X=x, is assumed to be "normally distributed". The Regression Model describes the Mean of that Normal Distribution as a function when X takes value X=x.
- Since we focus only on linear relationship, the above function represents the equation of a straight line- with a Slope and an Intercept- when we graph X on the horizontal axis and Y on the vertical axis.


X
70
50
30
Regression Line
Probability Distribution of Y

EXAMPLE

- For example, among the mothers who gained 37 lbs during their pregnancies and gave birth to baby boys, the boys' birth weights may not be all the same but form certain normal distribution.
- The Mean of that Normal Distribution depends on the weight gain:
  Mean (of BW) = Intercept + (Slope)(37)
- But that's the Mean, an individual BW is the Mean plus certain "deviation from the mean".

REGRESSION MODEL

- Model: $Y = \beta_0 + \beta_1 x + \varepsilon$ where β_0 and β_1 are two new parameters called regression coefficients, the Intercept and the Slope, respectively. The last term, ε, is the "error" representing the random fluctuation of y-values around their mean, $\beta_0 + \beta_1 x$, when $X=x$.
- The presence of the error term is an important characteristic of a statistical relationship; the points on a scatter diagram do not fall perfectly on the line.
- The scatter diagram is an useful diagnostic tool for checking out the Model (e.g. to see if it is linear).

REGRESSION COEFFICIENTS

- The error term ε would tell how spread the dots are around the regression line.
- The regression coefficients, β_0 and β_1, determine the position of the line and are important quantities in the analysis process. In “correlation analysis”, we need to know only the coefficient of correlation r which is proportional to the slop β_1 (we’ll see); but in a “regression analysis”, with new emphasis on prediction , so we need them both, β_0 and β_1.
- As parameters, both β_0 and β_1 are unknown; but they can be “estimated” by statistics from data

THE INTERCEPT

- If the scope of the model include $X = 0$, β_0 gives the Mean of Y when $X = 0$; otherwise, it does not have any particular meaning as a separate term.
- If the scope of the model does not include $X = 0$, we may choose a "transformation" such as:
  (New) $x = x - \overline{x}$
  Under this transformation, α gives the Mean of Y when $X = \overline{x}$, i.e. a "typical" subject (with value $\overline{x}$)

THE SLOPE

- The Slope is a more important parameter:
- (i) If X is binary (=0/1) representing an exposure, β_1 represents the increase in the mean of Y associated with the exposure (or a decrease if β_1 is negative);
- (ii) If X is on a continuous scale, β_1 represents the increase in the mean of Y associated with one unit increase in the value of X, X=x+1 vs. X=x, (or a decrease if β_1 is negative).
- The slope β_1 and the coefficient of correlation r are of the same "sign"; β_1 is positive for a positive association and negative for a negative association.

EXAMPLE

- For example, let X be a mother's weight gain during her pregnancy and Y the birth weight of the newborn. When X=x, the birth weights (BW) of all infants form certain normal distribution.
- The Mean of that Normal Distribution depends on the weight gain:
  Mean (of BW) = Intercept + (Slope)(x)
- The "slope" represents the average increase in birth weight for every pound the mother gained.

ESTIMATION OF PARAMETERS

- By the Model, when X=x, the Mean of Y is $\beta_0 + \beta_1 x$.
- The quantity $(\beta_0 + \beta_1 x)$ is the mean and Y is an observation when X=x; Y can be used as an estimate of that mean (sample of size 1). The error of that estimate is $[Y - (\beta_0 + \beta_1 x)]$ so that $\mathbf{Q = \Sigma [Y - (\beta_0 + \beta_1 x)]^2}$ represents the "total errors" (not distinguishing an under-estimation from an over-estimation); called "the sum of squared errors"
- The method of least squares requires that we find "good estimates" of β_0 and β_1 the values of b_0 and b_1 so as to minimize the "sum of squared deviations" Q.
- (We need Math, "Calculus", to carry out this step)

ESTIMATION OF PARAMETERS

- The "Least Squares" Estimates are:

$$b_1 = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

- Given the estimates "b_0" of the Intercept and "b_1" of the Slope, Estimate of Y (for a "new" value x of X) is $\hat{Y} = b_0 + b_1 x$. You can see that the slope b and the correlation r are proportional, that if one is 0 the other is 0.

SUM OF SQUARED ERRORS

- Since $[Y - (b_0 + b_1x)]$ represents the "error" of our prediction; $SSE = \Sigma [Y - (b_0 + b_1x)]^2$ is referred to as the (observed) "sum of squared errors", very much like the numerator of the sample variance s^2.
- The the Regression Model, the error term ε is assumed to have a Normal Distribution with mean 0 and variance σ^2. The variance σ^2 is estimated by SSE/(n-2); 2 degrees of freedom were lost due to the need to estimate the intercept and slope.

EXAMPLE #1

	x	y	x^2	y^2	xy
	1	3	1	9	3
	2	5	4	25	10
	6	7	36	49	42
Totals	9	15	41	83	55

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\{\sum x^2 - \frac{(\sum x)^2}{n}\}\{\sum y^2 - \frac{(\sum y)^2}{n}\}}} = \frac{55 - \frac{(9)(15)}{3}}{\sqrt{\{41 - \frac{9^2}{3}\}\{83 - \frac{15^2}{3}\}}} = .945$$

EXAMPLE #1

$$b_1 = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

x	y	x^2	y^2	xy
1	3	1	9	3
2	5	4	25	10
6	7	36	49	42
Totals	9	15	41	83	55

the estimates of the Slope and the Intercept are:

$$b_1 = \frac{55 - \frac{(9)(15)}{3}}{41 - \frac{(9)^2}{3}} = .714$$

$$b_0 = \frac{15}{3} - (.714)(\frac{9}{3}) = 2.858$$

For example, for new subject with X=5, it is predicted that its average y-value would be:
2.858 + (.714)(5) = 6.428

EXAMPLE #2:

n = 12

X = Birth weight (oz)

Y = Growth in weight between 70th and 100th days of life, as % of birth weight.

Birth weight data:

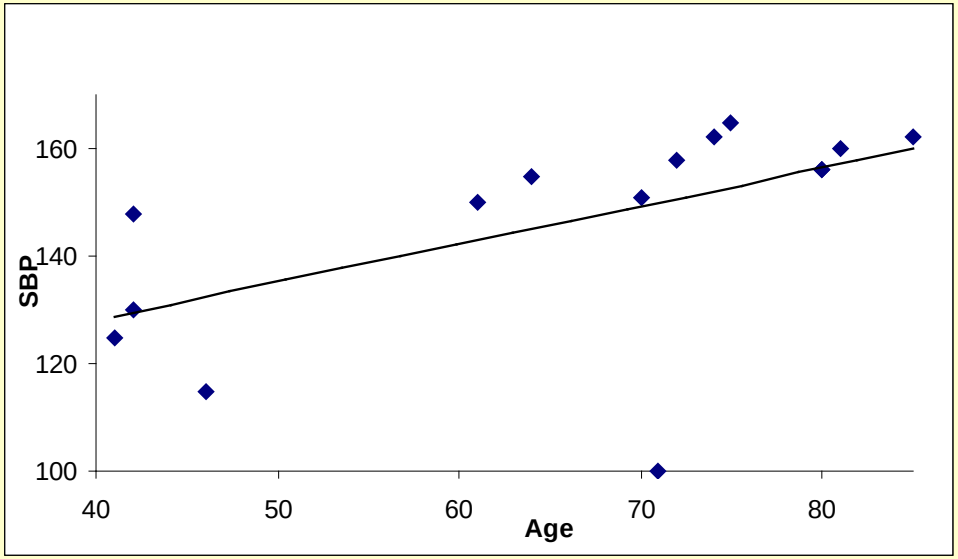
x (oz)	y (%)
112	63
111	66
107	72
119	52
92	75
80	118
81	120
84	114
118	42
106	72
103	90
94	91

Note: if the birth weight is 95 ounces, it is predicted that mean increase between days 70 & 100 would be 256.3 + (-1.74)(95) = 91%

Example #3: Age and SBP

n = 15, X = AGE (Years), Y = Systolic Blood Pressure (in mm of Hg)

Age (x)	SBP (y)
42	130
46	115
42	148
71	100
80	156
74	162
70	151
80	156
85	162
72	158
64	155
81	160
41	125
61	150
75	165



SBP versus AGE

Note: for 60-year-old women, it is predicted that their mean systolic blood pressure would be 99.6 + (.71)(60) = 142.2 mmHg.

How the data set was generated?

Are the results more suitable for some form of data than others? For example, is the method for regression applicable to “correlation data”?

Ideally the “story” should go like this:

Step #1: The investigator chooses n, the number of data points,

Step #2: The investigator chooses the levels of X: $x_1, x_2, \ldots, x_n$,

Step #3: “Nature makes n draws at random with replacement from the magic “error box” whose average is 0; call them $\varepsilon_1, \varepsilon_2, .., \varepsilon_n$.

Step#4: “nature” computes $y_1, y_2, \ldots, y_n$ from the formula/model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; the parameters are known to nature but not to investigator, nor statistician.

Step #5: Investigator get the data values $y_1, y_2, \ldots, y_n$ but none of the ingredients of the model: $\beta_0, \beta_1, \varepsilon 1, \varepsilon 2, .., \varepsilon n$.

Step#6: In the final step, the statistician’s task is to estimate the parameters β_0, β_1 and provide the standard errors for these estimates.

Data for Correlation & Regression Analysis, however, may be obtained from any sources: observational as well as experimental studies. All the results are equally applicable.

Besides estimating values of the “response” Y, at given values of the “predictor” X; efforts in regression analysis are also focus on the slope. These include forming its confidence intervals and/or testing if its true value is zero (if so, X and Y are not correlated). To do these, we need the “Standard Error” of the slope; this formula is given in the next slide, and details will be developed in the next few lectures.

$$Var(b_1) = \frac{\sigma^2}{\sum (x - \bar{x})^2}$$

$$SE(b_1) = \sqrt{\frac{\sigma^2}{\sum (x - \bar{x})^2}}$$

SCOPE OF THE MODEL

In formulating a regression model, we need to restrict the “coverage” of the model to some interval of values of the independent variable X; this is determined either by the design or the availability of data at hand. The shape of the regression function outside this range would be in doubt because the investigation provided no evidence as to the nature of the statistical relation outside this range. In short, one should not do any extrapolation.

DUE AS HOMEWORK

We have data on the conduct of a number of cancer clinical trials from "ClinicalTrials.gov" (File: Minority Enrollment); the aim is to is to investigate potential factors which might affect the enrollment of black patients. There were n=113 trials and the (response) variable under investigation is the percent of black patients ("Black") among those recruited for each trial. To provide possible explanations, we'll investigate 9 possible exploratory (or independent) factors represented by 10 variables: Age (1= under 18, 2 = 18 and above), Gender (1 = Male, 2 = Female, 3 = both), Funder (1 = Government, 2 = Industry, 4 = Combination), Trial Duration (in months), Allocation (1 = Randomized, 2 = Non-randomized), Intervention Model (or Design; 1 = Parallel (multiple arms), 2 = Single group, 3 = Cross-over), Primary Purpose (1 = Therapeutic, 2 = Non-therapeutic), Masking (1 = Open Label, 2 = Double Blind). The final factor, Trial Size, is represented by two variables: Actual enrollment, and Accrual Percentage which expressed accrual as percentage of Planned Accrual.

#2.1 Investigate the role of Trial Duration, Actual Enrollment, and Accrual Percentage using Simple Correlation (calculating Coefficient of correlation & test for independence).

#2.2 Are Actual Enrollment and Accrual Percentage correlated? (Optional Question: Why we would be interested in or concerned about relationship between independent variables?)