

# Applied Regression Analysis

## Week 3

1. Homework week 2: highlights
2. Correlation coefficient I
3. Correlation coefficient II
4. Coefficient of determination,  $r^2$
5. The ANOVA table for straight line regression I
6. The ANOVA table for straight line regression II
7. The ANOVA table for straight line regression III
8. Homework

Stanley Lemeshow, Professor of Biostatistics  
*College of Public Health, The Ohio State University*




THE OHIO STATE UNIVERSITY

**DEF: The correlation coefficient provides a measure of how two random variables are associated in a sample.**

- It is also a measure of the strength of the straight-line relationship between  $x$  and  $y$ .

$$\begin{aligned}
 r &= \frac{\widehat{\text{cov}}(x, y)}{\sqrt{\widehat{\text{var}}(x) \widehat{\text{var}}(y)}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \\
 &= \frac{\sum_{i=1}^n x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\left\{ \left[ \sum_{i=1}^n x_i^2 - \frac{(\sum x_i)^2}{n} \right] \left[ \sum_{i=1}^n y_i^2 - \frac{(\sum y_i)^2}{n} \right] \right\}^{1/2}}
 \end{aligned}$$

$s_x s_y$  

note: since  $\hat{\beta}_1 = \frac{\widehat{\text{cov}}(x, y)}{\widehat{\text{var}}(x)} = \frac{s_{xy}}{s_x^2}$

and  $r = \frac{s_{xy}}{s_x s_y}$

we have that  $r = \frac{s_x}{s_y} \hat{\beta}_1$

Example - For the AGE, SBP data,

$$r = \frac{199,576 - \frac{(1354)(4276)}{30}}{\left\{ \left[ 67894 - \frac{1354^2}{30} \right] \left[ 624260 - \frac{4276^2}{30} \right] \right\}^{1/2}} = 0.66$$

or, more simply, since  $\hat{\beta}_1 = 0.97$ ,  $r = \frac{15.29}{22.58} (0.97) = 0.66$

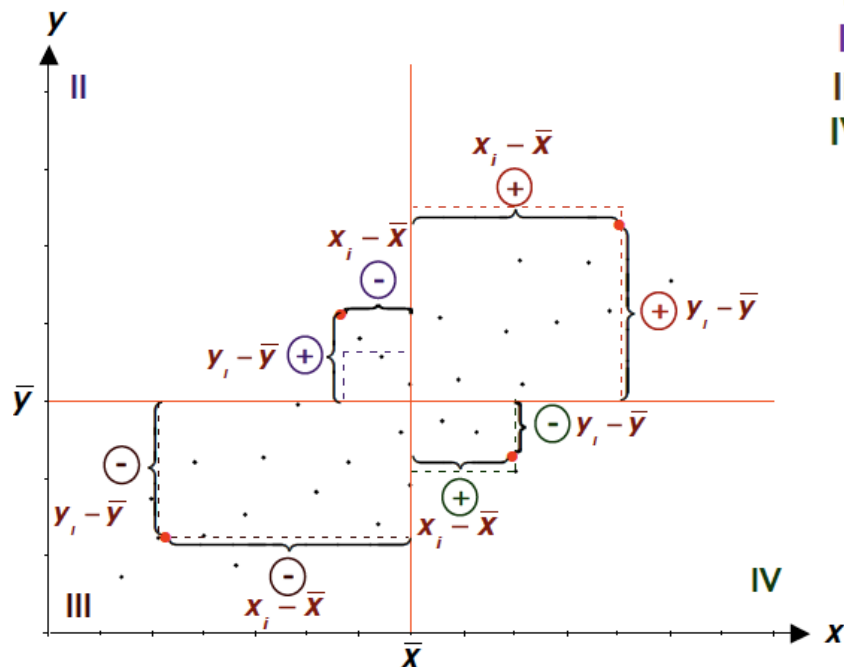
now  $-1 \leq r \leq +1$ , and  $r$  is dimensionless

- i.e., it is independent of the units of measurement of  $x$  or  $y$ .

finally,  $r$  always has the same sign as  $\hat{\beta}_1$

Actually,  $r$  is the standardized covariance.

Let us motivate what is meant by the covariance between  $x$  and  $y$ .

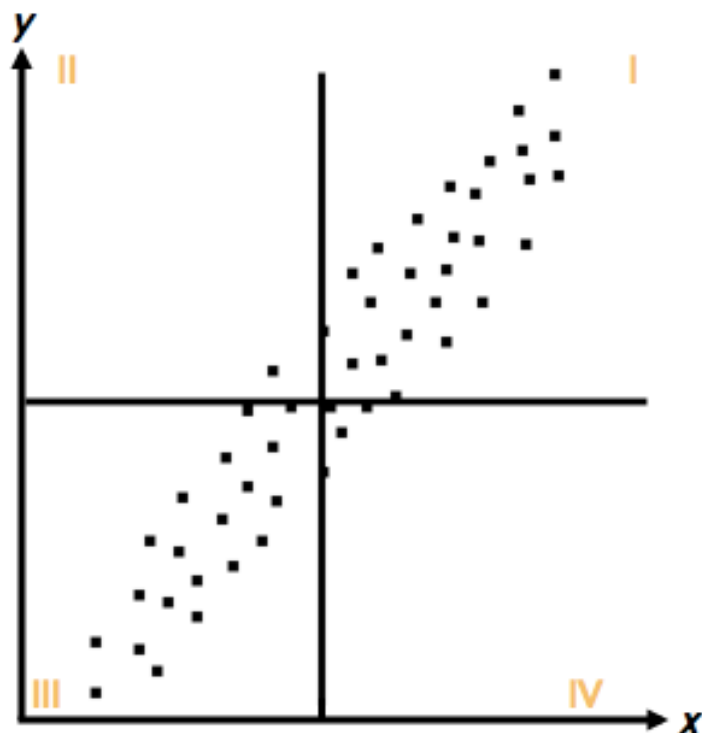


Quadrant	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
I	+	+	+
II	-	+	-
III	-	-	+
IV	+	-	-

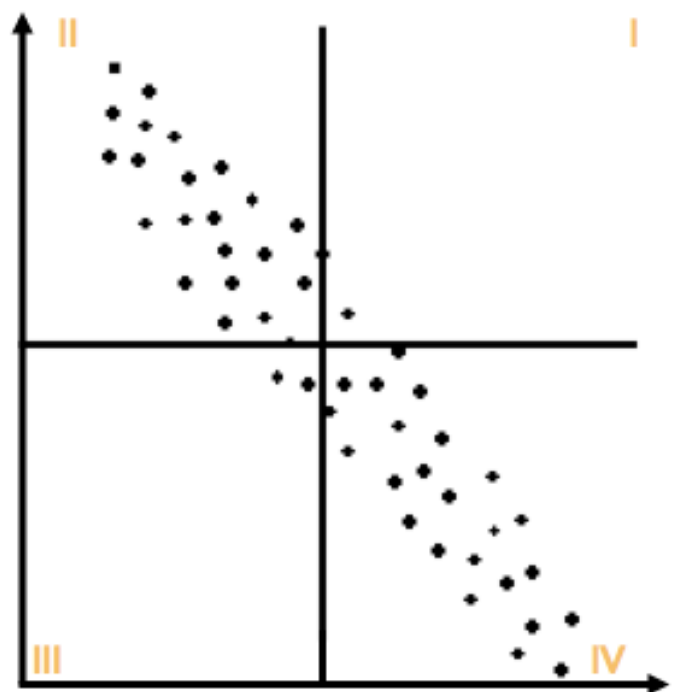
note: covariance =  $\sigma_{xy} = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{N}$

sample covariance:  $s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

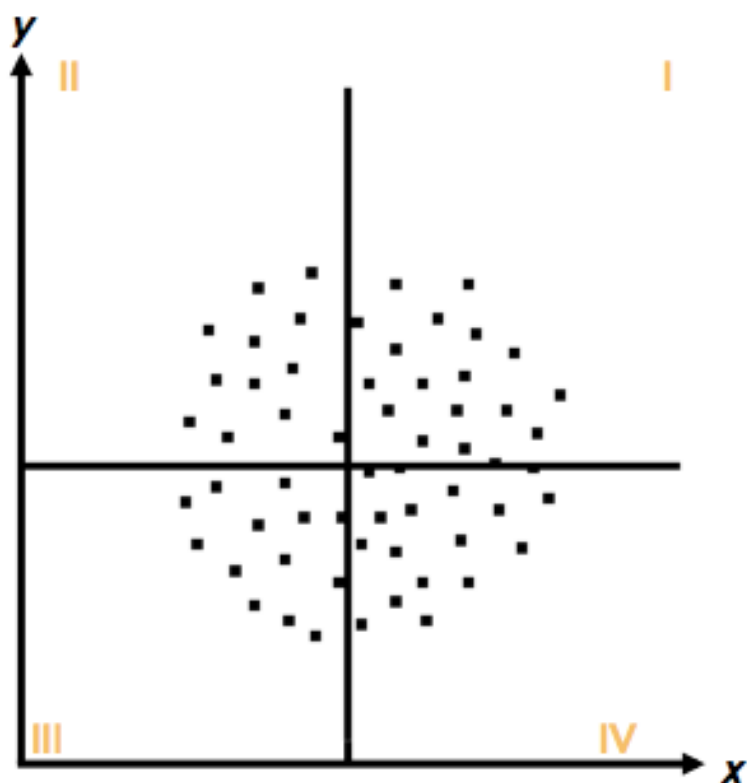
Now, if points look like:



$s_{xy} > 0$  since most points are in QI and QIII  
 $\therefore r > 0, \hat{\beta}_1 > 0$



$s_{xy} < 0$  since most points are in QII and QIV  
 $\therefore r < 0, \hat{\beta}_1 < 0$

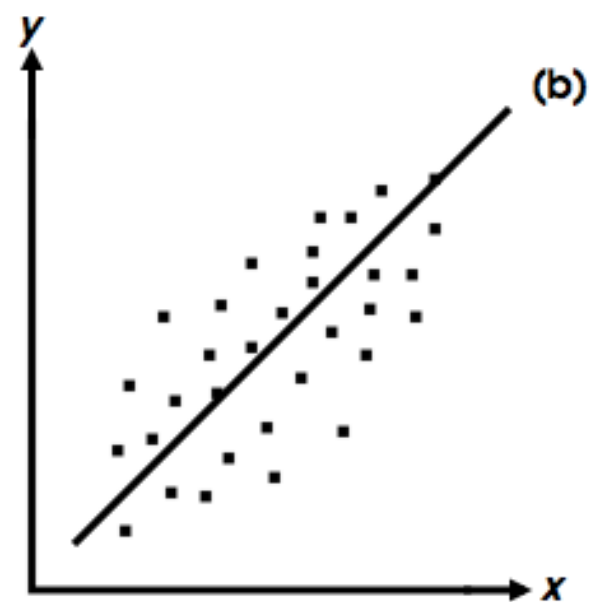
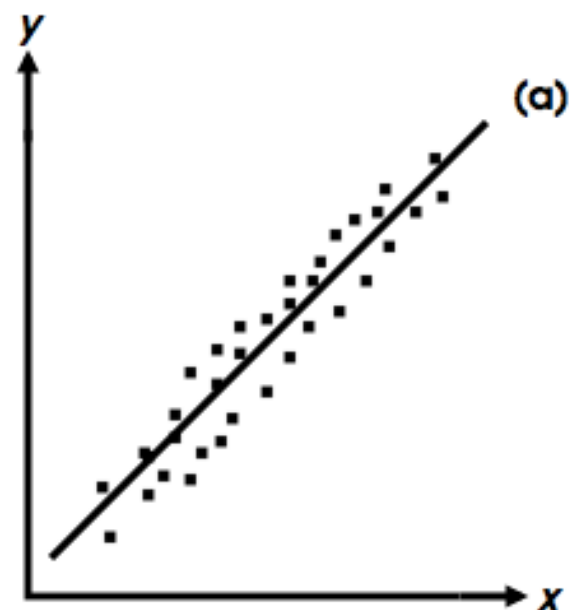


$S_{xy} = 0$  since + points are offset by - points

$$\therefore r = 0, \hat{\beta}_1 = 0$$

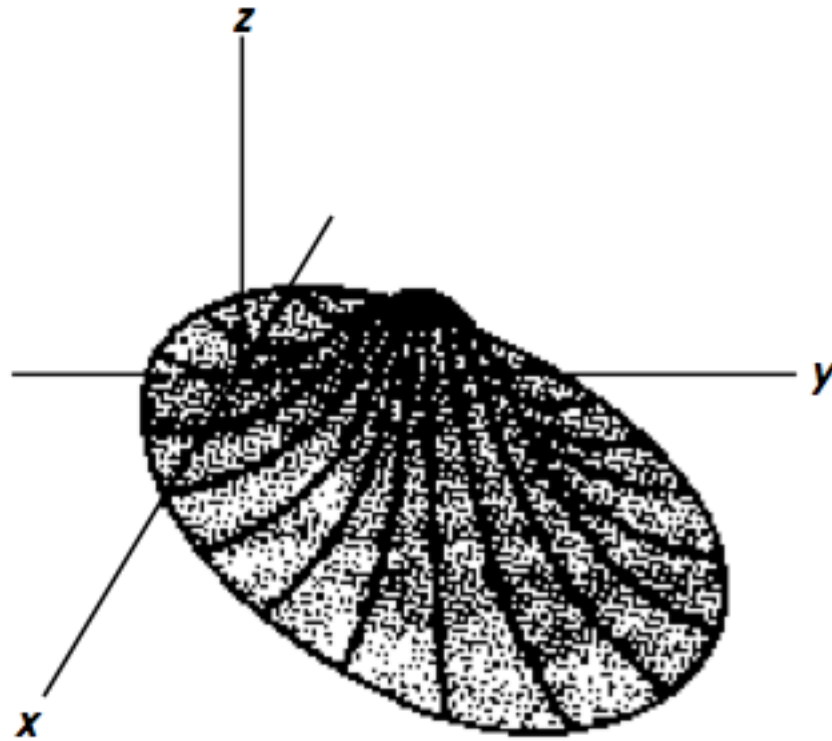
$r$  in (a) is much greater than  $r$  in (b) since there are fewer points in QII and IV in (a)

This is true even if though the slopes are identical

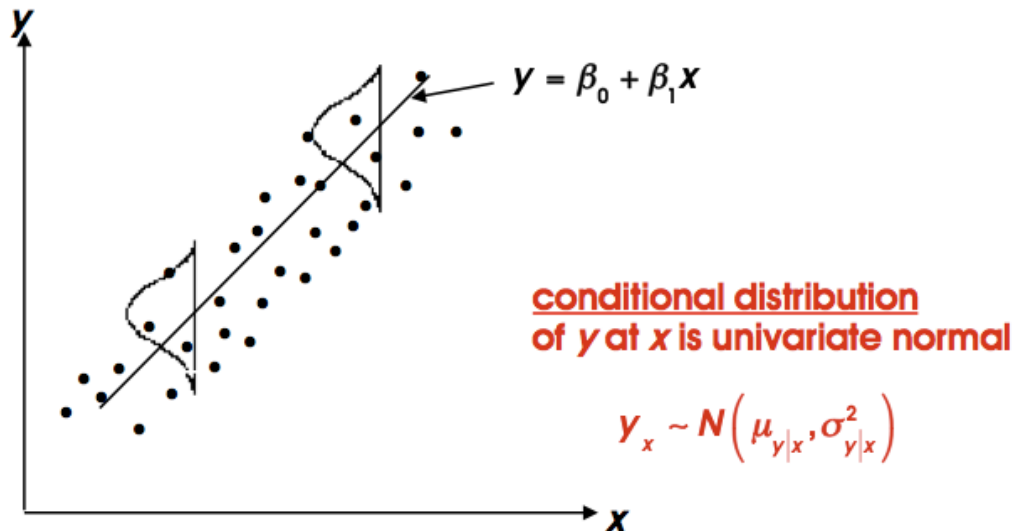


Let us assume that  $x$  and  $y$  are random variables having the bivariate normal distribution.

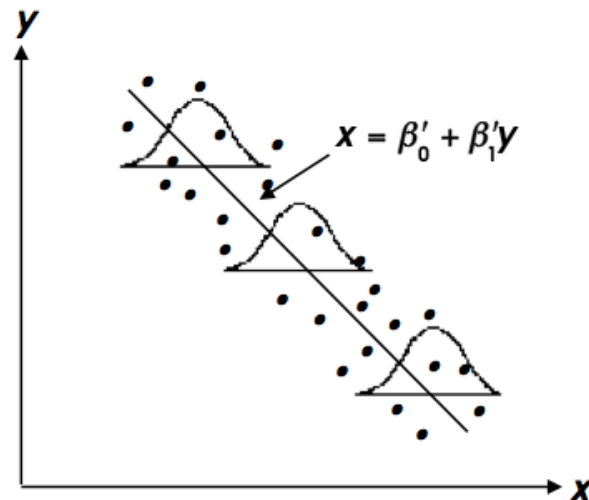
- This distribution has the following appearance:



One property of this distribution is that the distribution of  $y$  for any fixed  $x$  is normal



and the distribution of  $x$  for any fixed  $y$  is normal



We also assume  
homoskedasticity

note:  $\beta_0 \neq \beta'_0$   
 $\beta_1 \neq \beta'_1$



It follows from statistical theory that

$$\mu_{y|x} = \mu_y + \rho_{xy} \frac{\sigma_y}{\sigma_x} (x - \mu_x)$$

and

$$\sigma_{y|x}^2 = \sigma_y^2 (1 - \rho_{xy}^2)$$

Let  $\beta_1 = \rho_{xy} \left( \frac{\sigma_y}{\sigma_x} \right)$  and  $\beta_0 = \mu_y - \beta_1 \mu_x$

Then 
$$\mu_{y|x} = \underbrace{\beta_0 + \beta_1 \mu_x}_{\mu_y} + \underbrace{\beta_1}_{\rho_{xy} \left( \frac{\sigma_y}{\sigma_x} \right)} (x - \mu_x) = \beta_0 + \beta_1 x$$

hence we have the familiar straight-line model

Also,  $\mu_{y|x}$  can be estimated by

$$\hat{\mu}_{y|x} = \bar{y} + r \left( \frac{s_y}{s_x} \right) (x - \bar{x})$$

and, since  $\hat{\beta}_1 = r \left( \frac{s_y}{s_x} \right) = \frac{s_{xy}}{s_x s_y} \left( \frac{s_y}{s_x} \right) = \frac{s_{xy}}{s_x^2}$

$$\hat{\mu}_{y|x} = \bar{y} + \hat{\beta}_1 (x - \bar{x}) \leftarrow \text{usual least squares line}$$

Thus, the least squares formulae for  $\hat{\beta}_0$  and  $\hat{\beta}_1$  can be developed by assuming that  $x$  and  $y$  are random variables having the bivariate normal distribution and by substituting the usual estimates of  $\mu_x$ ,  $\mu_y$ ,  $\sigma_x$ ,  $\sigma_y$ , and  $\rho_{xy}$  into the expression for  $\mu_{y|x}$ .

Also,  $\hat{\sigma}_{y|x}^2 = s_{y|x}^2 = \frac{n-1}{n-2} (s_y^2 - \hat{\beta}_1^2 s_x^2) = \frac{n-1}{n-2} s_y^2 (1 - r^2)$

$$\approx s_y^2 (1 - r^2)$$

Now, as we've seen,

$$\sigma_{y|x}^2 = \sigma_y^2 (1 - \rho_{xy}^2)$$

where  $\sigma_y^2$  = unconditional variance of  $y$

i.e., it's the variance of  $y$  when we know nothing about  $x$

On the other hand,

$$\sigma_{y|x}^2 = \text{conditional variance of } y$$

i.e., it's the variance of  $y$  when we know the corresponding value of  $x$

Hence, the reduction in the variance of  $y$  due to knowledge of  $x$  is:

$$\sigma_y^2 - \sigma_{y|x}^2 = \rho_{xy}^2 \sigma_y^2$$

and

$$\rho_{xy}^2 = \frac{\sigma_y^2 - \sigma_{y|x}^2}{\sigma_y^2}$$

Hence, the squared correlation coefficient is the proportion of the variance of  $y$  "explained by" knowledge of  $x$ .

When  $\rho_{xy} = 0$  this means that  $\sigma_{y|x}^2 = \sigma_y^2$

i.e., none of the variance in  $y$  is explained by the regression of  $y$  on  $x$ .

When  $\rho_{xy} = 1$  this means that  $\sigma_{y|x}^2 = 0$

i.e., all of the variance in  $y$  is explained by the regression of  $y$  on  $x$ .

i.e., the relationship between  $y$  and  $x$  is perfectly linear.

Hence, defining  $SSY = \sum_{i=1}^n (y_i - \bar{y})^2$

and

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

if the model fits,  $SSE \ll SSY$ . Then

$$r^2 = \frac{SSY - SSE}{SSY}$$

is a quantitative measure of the improvement in the fit obtained by using  $x$  and

$$0 \leq r^2 \leq 1$$

## Note:

One should not be led to a false sense of security by considering the magnitude of  $r$  rather than of  $r^2$  when assessing the strength of the linear association between  $x$  and  $y$ .

e.g. when  $r = .5$ ,  $r^2 = .25$

$$r = .7, \quad r^2 = .49$$

$$r = .3, \quad r^2 = .09$$

## Example

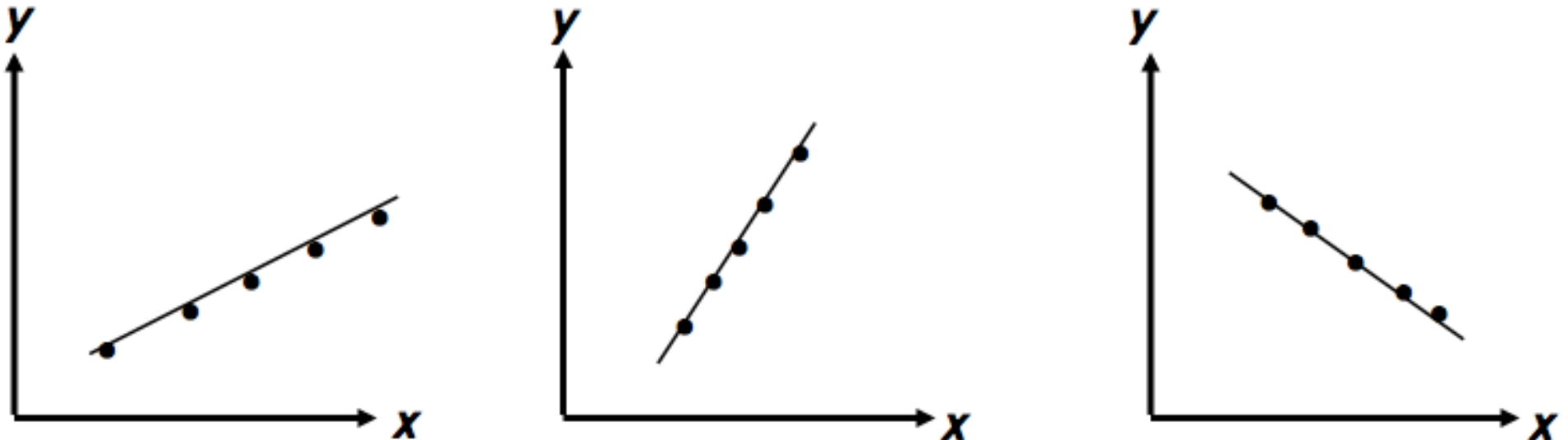
For the age-SBP data

$$r = .66$$

$$r^2 = .44$$

Also note:

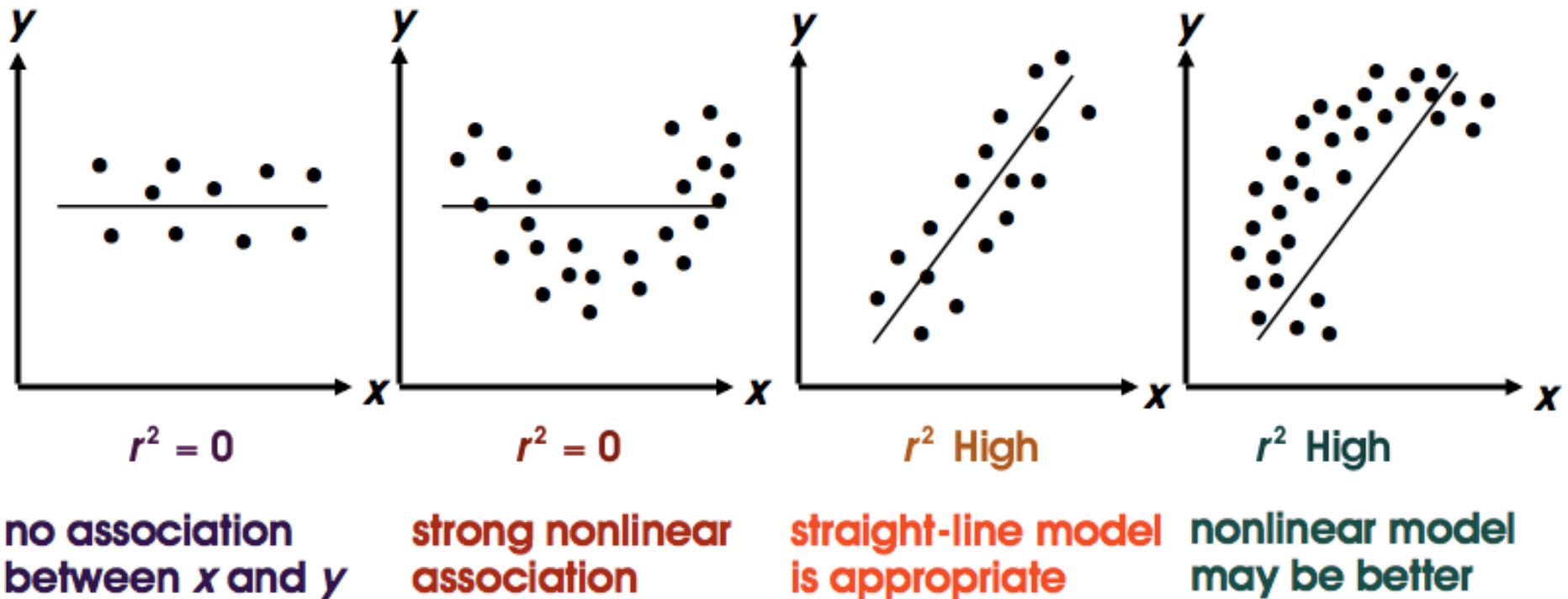
- $r^2$  is not a measure of the magnitude of the slope of the regression line



In all instances  $r^2 = 1$  but the slopes differ widely.

- $r^2$  is a measure of the clustering of points about the regression line.

$r^2$  is not a measure of the appropriateness of the straight-line model





To test  $H_0 : \rho_{xy} = 0$

vs.  $H_a : \rho_{xy} \neq 0$

(or one-sided)

we may simply test

$$H_0 : \beta_1 = 0$$

vs.  $H_a : \beta_1 \neq 0$

as we learned before

or, we may use

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{and} \quad t \sim t(n-2)$$

Note:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \equiv \frac{\hat{\beta}_1 s_x}{s_{y|x} / \sqrt{n-1}}$$

that we learned before

e.g., in the age-SBP problem,

$$r = .66$$

and

$$t = \frac{.66\sqrt{30-2}}{\sqrt{1-(.66)^2}} = 4.62$$

which is the same value as that obtained in the test for the slope

To test

$$H_0 : \rho_{xy} = \rho_0$$

$$\text{vs. } H_a : \rho_{xy} \neq \rho_0 \quad \text{where } \rho_0 \neq 0$$

we cannot use  $t$  as described previously.

Also,  $H_0 : \rho_{xy} = \rho_0 (\neq 0)$  is not equivalent to  $H_0 : \beta_1 = \beta_1^{(0)}$ .

We must consider the distribution of  $r$

- $r \sim$  symmetric only when  $\rho_{xy} = 0$
- $r$  is not symmetric when  $\rho_{xy} \neq 0$ 
  - In that case the sampling distribution of  $r$  is skewed.

Hence, we cannot use  $t$  (that has a normally distributed estimator in the numerator and an estimate of its standard deviation in the denominator).

Fortunately, through an appropriate transformation,  $r$  can be changed to a statistic that is approximately normal.

## FISHER' S Z TRANSFORMATION

$$z = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right)$$

note:  $z = \text{inv hyp tan } (r)$

has been shown to be approximately normal with

$$E(z) = \frac{1}{2} \ln \left( \frac{1 + \rho_{xy}}{1 - \rho_{xy}} \right) = Z \quad \text{and} \quad \text{var}(z) = \frac{1}{n-3}$$

i.e.,

$$z \sim N \left( Z, \frac{1}{n-3} \right)$$

The inverse Fisher transformation is

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

note:  $r = \text{hyp tan } (z)$

Example: for the AGE-SBP data

$$\text{Test: } H_0 : \rho_{xy} = .85$$

$$\text{vs: } H_a : \rho_{xy} \neq .85$$

$$r = 0.66, \quad n = 30$$

$$z = \frac{1}{2} \ln \left( \frac{1 + .66}{1 - .66} \right) = .7928$$

$$Z = \frac{1}{2} \ln \left( \frac{1 + .85}{1 - .85} \right) = 1.2561$$

Then

$$z' = \frac{z - Z}{\frac{1}{\sqrt{n-3}}} = \frac{.7928 - 1.2561}{\frac{1}{\sqrt{27}}} = -2.41$$

and reject  $H_0$  if  $z' > 1.96$

or if  $z' < -1.96 \quad \therefore \text{reject } H_0.$

Similarly, we can get a  $(1 - \alpha)\%$  C.I. estimate for  $\rho_{xy}$

$$z - 1.96\sqrt{\frac{1}{n-3}} \leq Z \leq z + 1.96\sqrt{\frac{1}{n-3}}$$

$$.7928 - 1.96\sqrt{\frac{1}{27}} \leq Z \leq .7928 + 1.96\sqrt{\frac{1}{27}}$$

$$.4156 \leq Z \leq 1.1700$$

Using the inverse transformation

$$r_L = \frac{e^{2(.4156)} - 1}{e^{2(.4156)} + 1} = \frac{1.296}{3.296} = .393$$

$$r_U = \frac{e^{2(1.1700)} - 1}{e^{2(1.1700)} + 1} = \frac{9.38}{11.38} = .824$$

$$.393 \leq \rho_{xy} \leq .824$$

```
. z_r sbp age (STB-32: sg51)
(sample correlations, n=30)
      sbp      age
sbp  1.0000
age  0.6576  1.0000

(lower\upper 95% confidence limits)
      sbp      age
sbp  1.0000  0.8229
age  0.3896  1.0000
```

Let

$$(y_i - \bar{y}) = \text{"total" deviation} = a$$

$$(y_i - \hat{y}_i) = \text{"unexplained" deviation} = b$$

$$(\hat{y}_i - \bar{y}) = \text{"explained" deviation} = c$$

Note:

$$(y_i - \bar{y}) = (y_i - \hat{y}_i + \hat{y}_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

$$a = b + c$$

now, squaring both sides of the equation

$$(y_i - \bar{y})^2 = (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

and summing over all  $n$  points

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 0$$

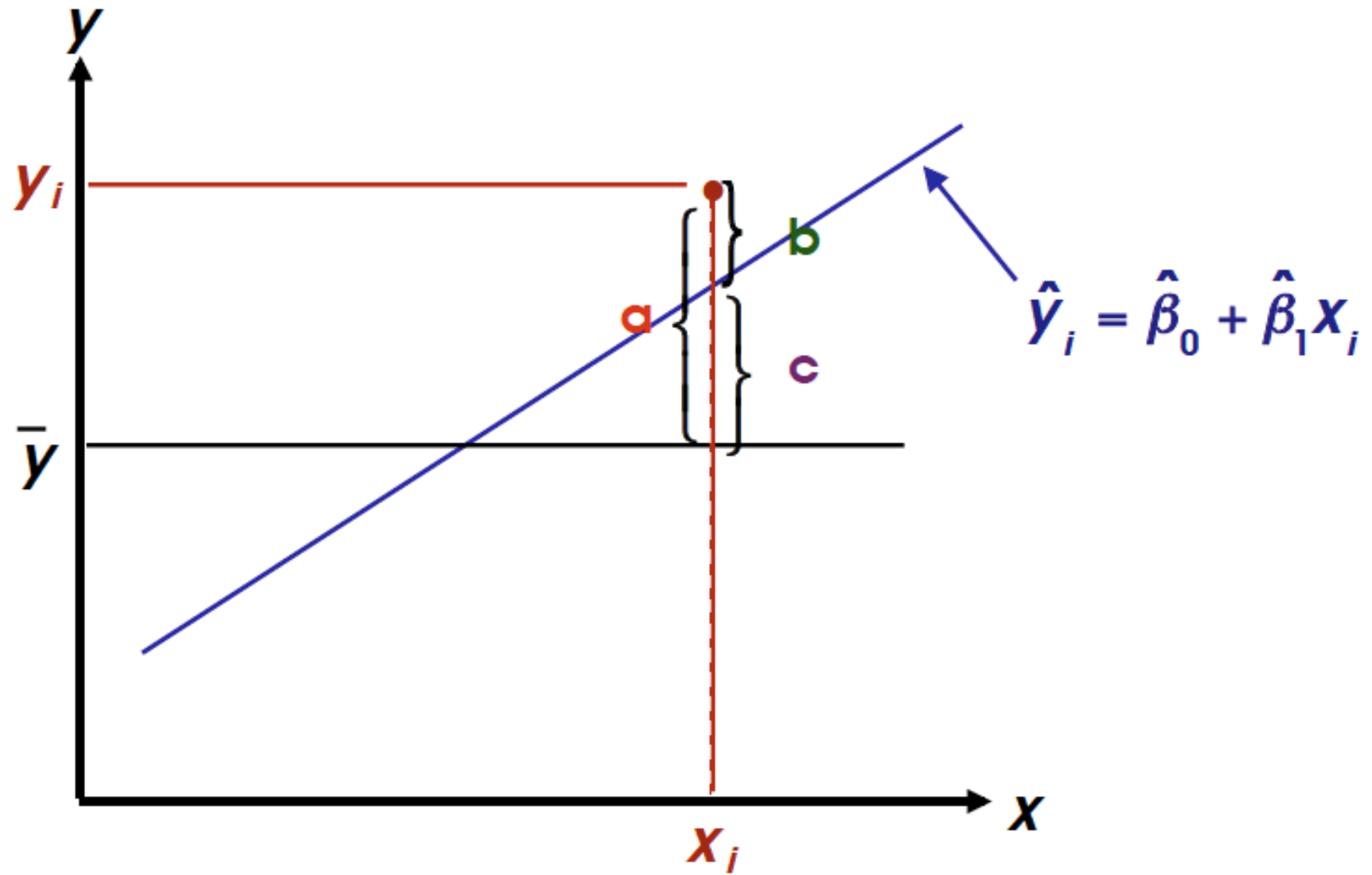
$\sum$  cross products = 0

This is called the:

“fundamental equation of regression analysis”



graphically,



**Recall:**

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \text{SSY} \quad \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \text{SSE} \quad \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \text{SSY} - \text{SSE}$$

**Now, for the age-SBP data this can all be summarized in an ANOVA table as follows:**

Source	df	SS	MS	EMS	F
Regression (x)	1	SSY-SSE = 6394.02	6394.02	$\sigma^2 + c\beta_1^2$	21.33
Residual	28	SSE = 8393.44	299.77	$\sigma^2$	
Total	29	SSY = 14787.46			

**Note:**  $r^2 = \frac{\text{SSY-SSE}}{\text{SSY}} = \frac{6394.02}{14787.46} = 0.43$

$$\frac{\text{SSE}}{\text{df}} = \text{MSE} = s_{y|x}^2 \quad \text{we computed earlier}$$

$s_{y|x}^2$  is an estimate of  $\sigma^2$  if the true regression model is linear.

SSY – SSE provides an estimate of  $\sigma^2$  only if the variable  $x$  does not help to predict the dependent variable,  $y$  (i.e.,  $\beta_1 = 0$ )

$MS_1$  and  $MS_2$  are independent and, if  $H_0 : \beta_1 = 0$  is true, then

$$F = \frac{MS_1}{MS_2} \sim F(1, n - 2)$$

This is also a test of  $H_0 : \rho_{xy} = 0$

Fortunately this  $F$  test is equivalent to the 2-sided  $t$ -test discussed previously.

recall:

$$F(1, \nu) = t^2(\nu)$$

so

$$F_{.95}(1, \nu) = \left( t_{.975}(\nu) \right)^2$$

e.g.,

in the age-SBP example

$$t = 4.62 \quad t^2 = 21.33 = F$$

$$\text{also, } t_{.975}(28) = 2.05 \quad \text{and } (2.05)^2 = 4.20 = F_{.95}(1, 28)$$

Hence, the two sided critical region:

$$\begin{aligned} &\text{reject } H_0 \text{ if } t > t_{.975}(\nu) \\ &\quad \text{or if } t < t_{.025}(\nu) \end{aligned}$$

is equivalent to

$$\text{reject } H_0 \text{ if } F > F_{.95}(1, \nu)$$

# Spreadsheets For Computing ANOVA Table In Regression

ID	SBP (y)	AGE (x)	$x - \bar{x}$	$(x - \bar{x})^2$	$y - \bar{y}$	$(y - \bar{y})^2$	$(x - \bar{x}) \times (y - \bar{y})$	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$	$(\hat{y} - \bar{y}) \times (\hat{y} - \bar{y})$
1	144	39	-6.13	37.62	1.47	2.15	-9.00	136.58	7.42	55.08	-5.95	35.46	-44.19
2	220	47	1.87	3.48	77.47	6001.08	144.60	144.35	75.65	5723.58	1.81	3.28	137.11
3	138	45	-0.13	0.02	-4.53	20.55	0.60	142.40	-4.40	19.39	-0.13	0.02	0.57
4	145	47	1.87	3.48	2.47	6.08	4.60	144.35	0.65	0.43	1.81	3.28	1.19
5	162	65	19.87	394.68	19.47	378.95	386.74	161.82	0.18	0.03	19.29	372.03	3.45
:	:	:	:	:	:	:	:	:	:	:	:	:	:
25	160	44	-1.13	1.28	17.47	305.08	-19.80	141.43	18.57	344.73	-1.10	1.21	-20.43
26	158	53	7.87	61.88	15.47	239.22	121.67	150.17	7.83	61.30	7.64	58.33	59.80
27	144	63	17.87	319.22	1.47	2.15	26.20	159.88	-15.88	252.16	17.35	300.89	-275.45
28	130	29	-16.13	260.28	-12.53	157.08	202.20	126.87	3.13	9.80	-15.66	245.34	-49.03
29	125	25	-20.13	405.35	-17.53	307.42	353.00	122.99	2.01	4.05	-19.55	382.08	-39.36
30	175	69	23.87	569.62	32.47	1054.08	774.87	165.70	9.30	86.40	23.17	536.92	215.38
SUM	4276	1354	0.00	6783.47	0.00	14787.47	6585.87		0.00	8393.44	0.00	6394.02	0.00
MEAN	142.53	45.13											
VAR	509.91	233.91											

N= 30  
 SLOPE= 0.97  
 INTERCEPT= 98.71  
 CORR= 0.66

SOURCE	DF	SS	MS	F	p
REGRESSION	1	6394.02	6394.02	21.33	0.00
RESIDUAL	28	8393.44	299.77		
TOTAL	29	14787.47			