LESSON 10 & 11
SIMPLE LINEAR REGRESSION

SAMIE L.S. LY

1. Simple Linear Regression Model
2. Least Square Method
3. Coefficient of Determination
4. Model Assumptions
5. Testing for Significance
6. Covariance & Coefficient of Correlation
7. Using the Estimated Regression –Equation for Estimation and Prediction

COMM215
First the Foundation, then Innovation

SIMPLE LINEAR REGRESSION MODEL

REGRESSION MODEL

REGRESSION EQUATION

ESTIMATED REGRESSION EQUATION

S.L.S.Ly
C215

THE OBJECTIVE OF SLR

Let’s say, I would like to create the ultimate equation to figure out my Final COMM 215 grade.

What influences my Final COMM 215 grade?
Think of a few examples.

S.L.S.Ly
C215

FINAL COMM 215 Grade (y) =

Study Time (x_1) +

Work Time (x_2) +

# of hours spent on Facebook (x_3) +

Family Time (x_4) +

anything else you can think of ($x_{\ldots}$) …

S.L.S.Ly
C215

THE OBJECTIVE OF SLR

Let's say, I would like to create the ultimate equation to figure out my Final COMM 215 grade.

S.L.S.Ly
C215

Let's say,

I study 5 hours a week,

work part-time for 25 hours a week,

I spend 3 hours on facebook,

And I have 2 kids.

Here is 1 scenario.

THE OBJECTIVE OF SLR

Let's say, I would like to create the ultimate equation to figure out my Final COMM 215 grade.

S.L.S.Ly
C215

What if it was someone else?
With a different profile?

Do I have to make my calculations all over again?

If I set up a regression line, I just need to plug in values and get an estimation of my Final COMM 215 grade.

Voila!

Then you might ask… how?

THE OBJECTIVE OF SLR

Let's say, I would like to create the ultimate equation to figure out my Final COMM 215 grade.

S.L.S.Ly C215

How do I create this regression line?

Answer: By gathering data from history.
I am going to take a sample of individuals who took COMM 215 before and write down their profile.

Hours per week	**Study Time**	**Work Time**	**Family Time**	**Facebook Time**
Bob	10	25	5	0
Sally	12	0	15	15
Eric	3	40	10	0

COMM215
First the Foundation, then Innovation

THE OBJECTIVE OF SLR

Let's say, I would like to create the ultimate equation to figure out my Final COMM 215 grade.

Since we are only considering 1 independent variable, let's take just 1, Study Time as the main indicator of your Final COMM 215 grade.

Hours per week	**Study Time(x)**	**Final Grade(y)**
Bob	10	89
Sally	12	67
Eric	3	45

THE OBJECTIVE OF SLR

Let's say, I would like to create the ultimate equation to figure out my Final COMM 215 grade.

We've generated this equation!
Now, if Michelle asks, if I study 8 hours a week, what would be my estimated Final COMM 215 grade?

SLSLy
C215

$$y = 3.4478x + 38.269$$

SIMPLE LINEAR REGRESSION

1 VARIABLE

FINAL COMM 215 GRADE (y) = STUDY TIME (x)

S.L.S.Ly
C215

MULTIPLE LINEAR REGRESSION

MORE THAN 1 VARIABLE

FINAL COMM 215 GRADE (y) =

STUDY TIME (x_1) + WORK TIME(x_2) + ….

As a way of predicting sales

Managerial decisions are often made based on the relationship between two or more variables.

S.L. S.Ly C215

The statistical process is called

regression analysis used to develop an equation showing

how the variables are related.

The variable

being predicted is

called the dependent variable.

S.L.S.Ly C215

The variable or variables

being used to predict the value of the dependence variable are

called independent variables.

FINAL COMM 215 Grade (y) =

Study Time (x_1) +

Work Time (x_2) +

# of hours spent on Facebook (x_3) +

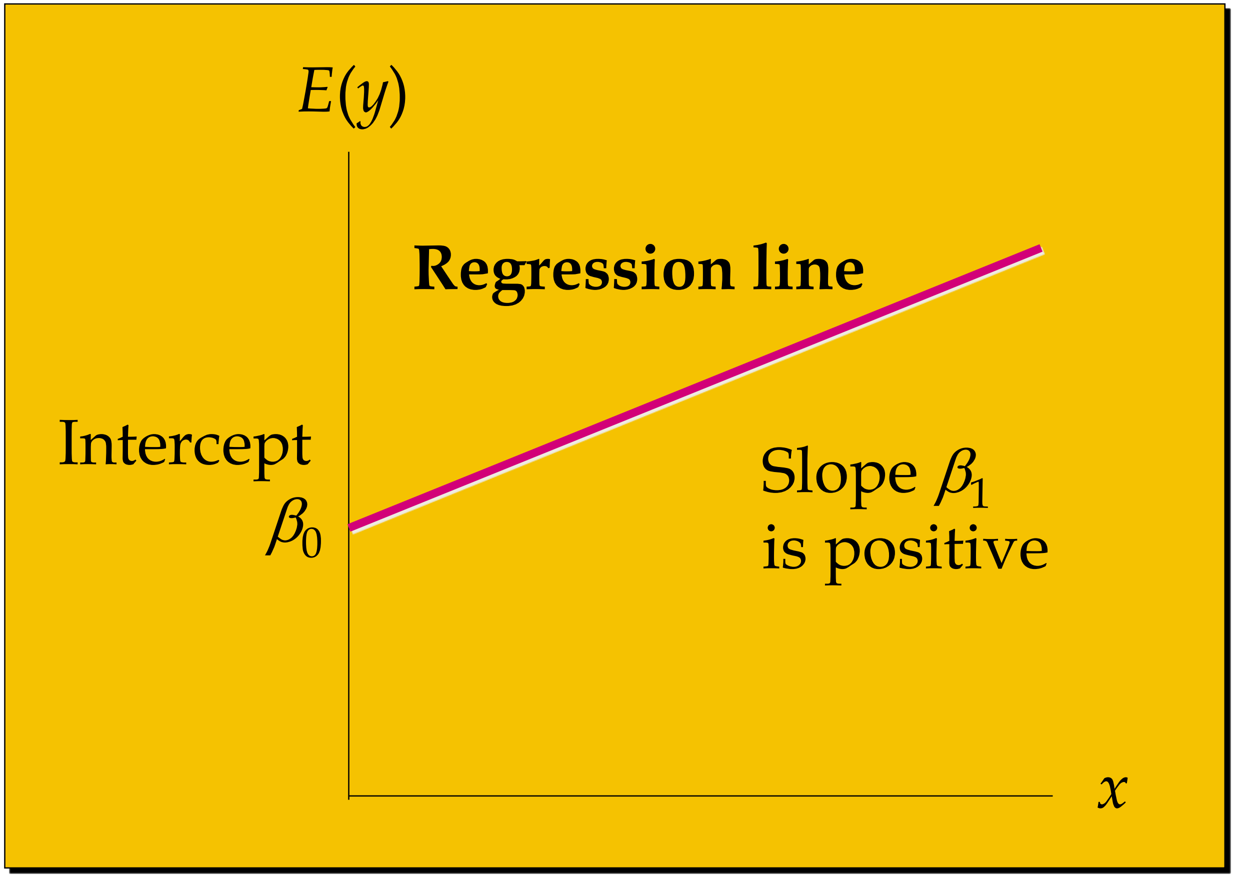
Family Time (x_4) +

anything else you can think of ($x_{...}$) …

S.L.S.Ly
C215

SIMPLE LINEAR REGRESSION EQUATION

- Positive Linear Relationship

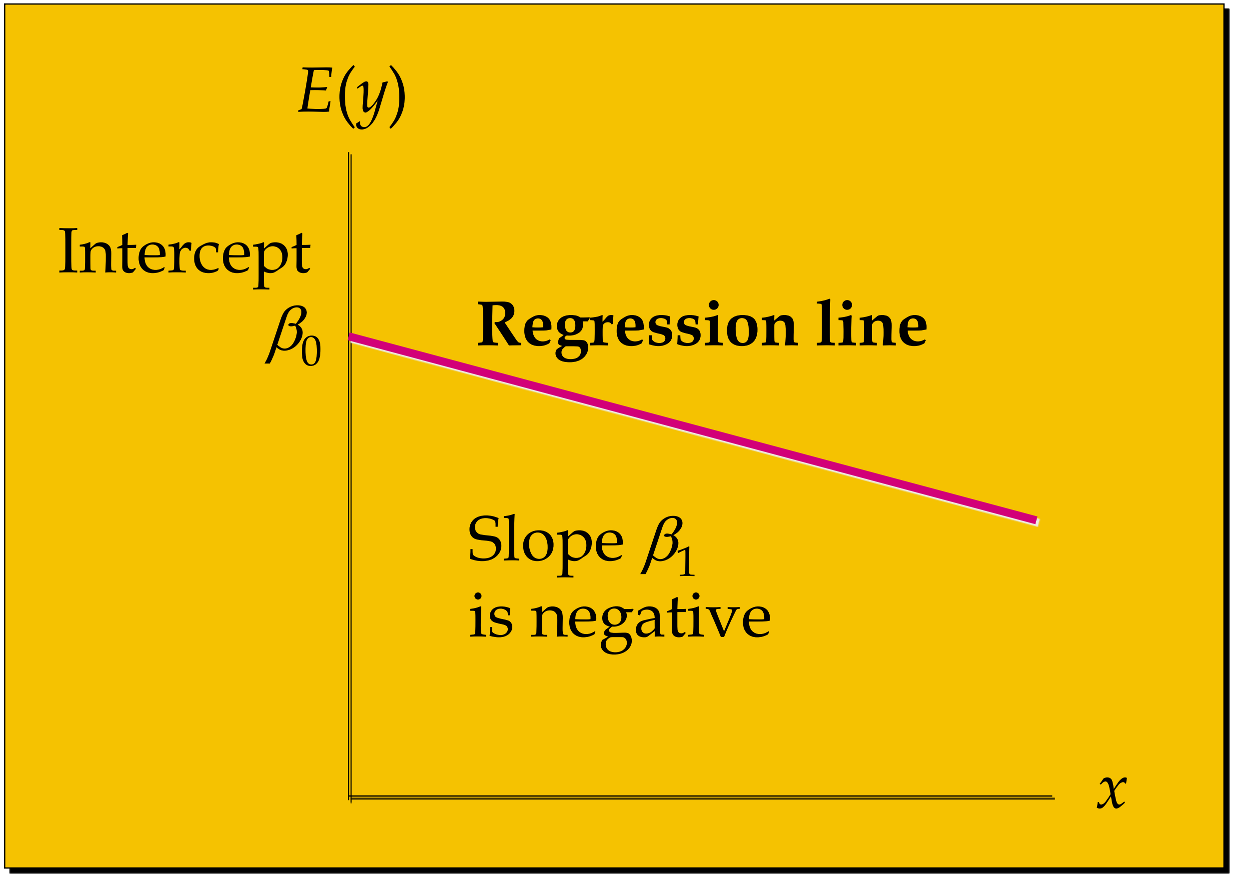


- The relationship between the two variables is approximated by a straight line.

Bowerman, et al. (2017) pp. 534

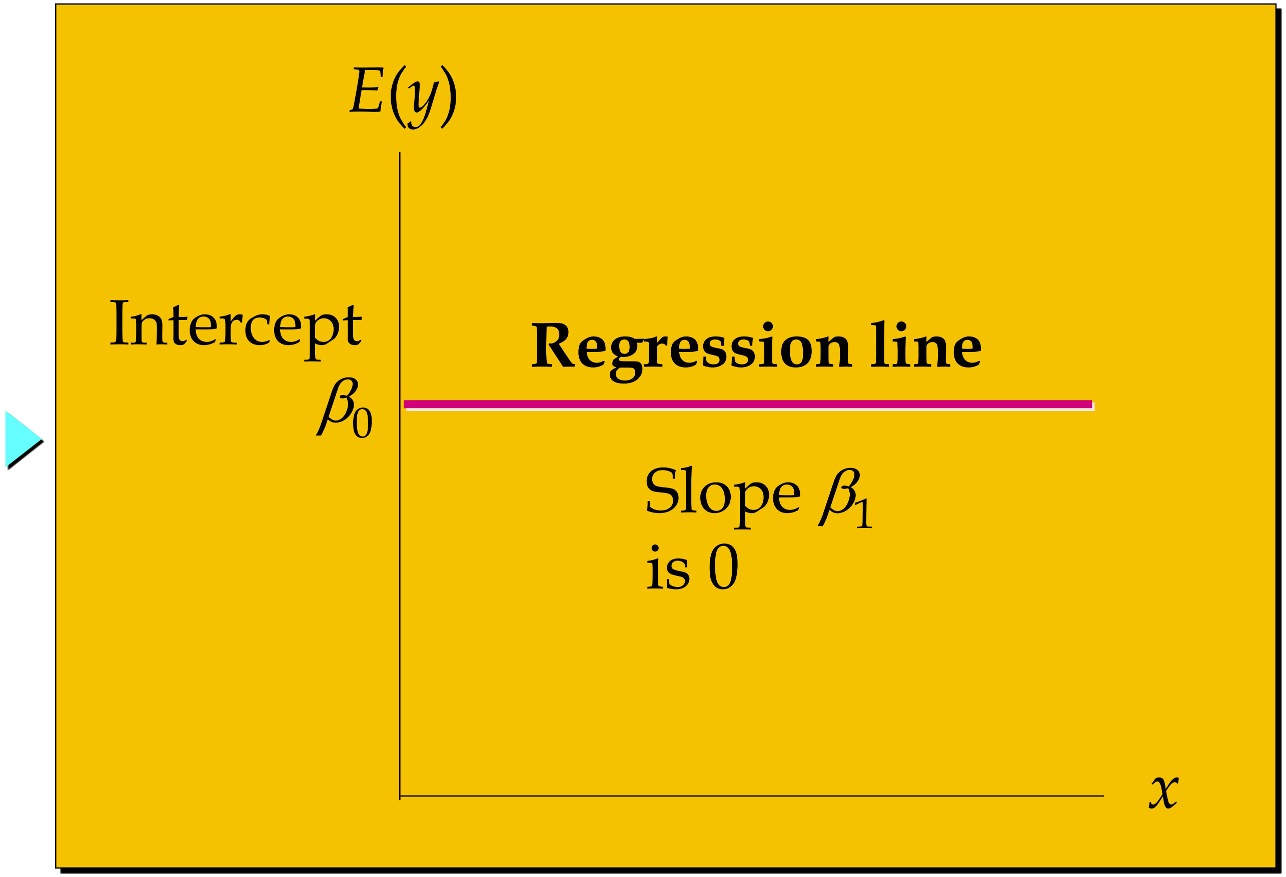
SIMPLE LINEAR REGRESSION EQUATION

- Negative Linear Relationship



SIMPLE LINEAR REGRESSION EQUATION

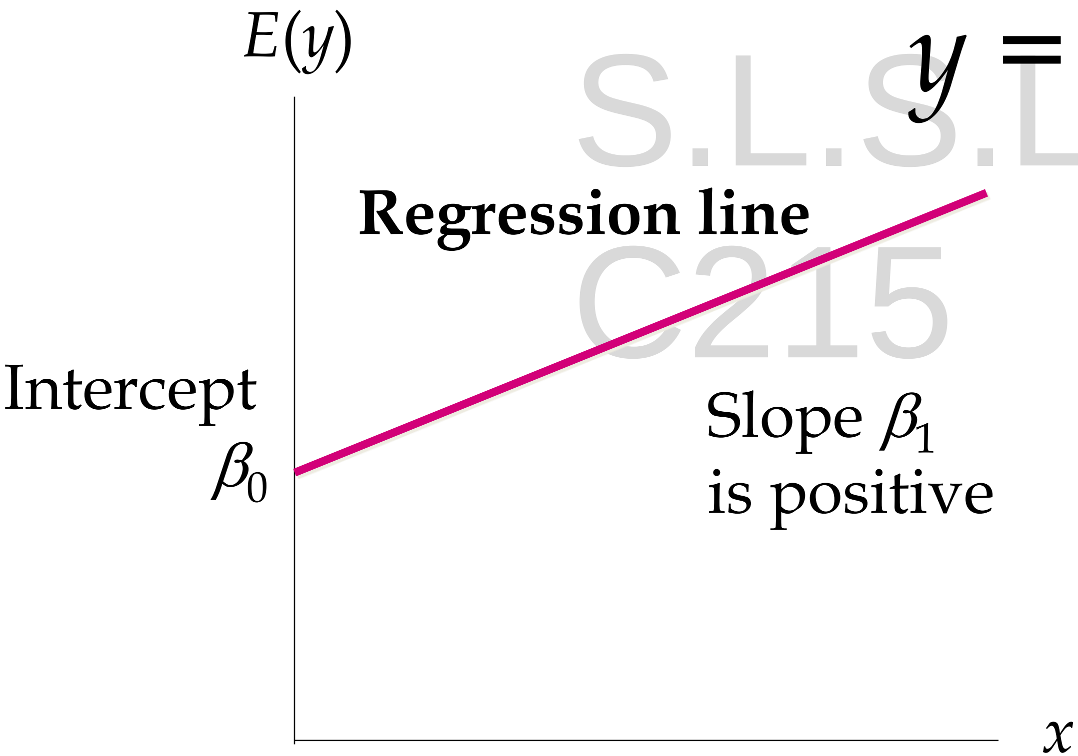
- No Relationship



SIMPLE LINEAR REGRESSION MODEL

S.L.S.Ly
C215

$$y = \beta_0 + \beta_1 x + \varepsilon$$



β_o and β_1 – parameters of the model

ε is a random variable referred to as the error term.

ε – variability in y that cannot be explained

Bowerman, et al. (2017) pp. 534

CHARACTERISTICS OF THE ERROR TERM

S.L.S.Ly
C215

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The more variables you add
The small ε will become
because now you know more!

$$y=\beta_0+\beta_1 x_1+ \beta_2 x_2+ \beta_3 x_3+ \beta_4 x_4 + \beta_5 x_5+ \beta_6 x_6+ \beta_7 x_7+ \beta_8 x_8 \ldots + \beta_i x_i$$

CHARACTERISTICS OF THE ERROR TERM

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Cannot be explained!

Cannot be calculated!

Cannot.. Just don't ask sigh…

S.L.S.Ly C215

The more variables you add
The small ε will become
because now you know more!

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7+\beta_8x_8\ldots+\beta_ix_i$$

Bowerman, et al. (2017) pp. 533

REGRESSION EQUATION

Describes how the expected value of y, E(y) is related to x

$$E(y) = \beta_0 + \beta_1 x$$

SLSly
C215

- Graph of the regression equation is a straight line.
- β_0 is the y intercept of the regression line.
- β_1 is the slope of the regression line.
- $E(y)$ is the expected value of y for a given x value.

Bowerman, et al. (2017) pp. 535

ESTIMATED SIMPLE LINEAR REGRESSION EQUATION

$$E(y) = \beta_0 + \beta_1 x$$

Sample statistics are

S.L.S.Ly
C215

$$\hat{y} = b_0 + b_1 x$$

- The graph is called the estimated regression line.
- b_0 is the y intercept of the line.
- b_1 is the slope of the line.
- $\hat{y}$ is the estimated value of y for a given x value.

Bowerman, et al. (2017) pp. 535

PROBLEM #10.1

In the linear regression equation, $y = b_0 + b_1x_1$, why is the term at the left given as $\hat{y}$ instead of simply y?

because it is an estimated value for the dependent variable given a value of x.

S.L.S.Ly
C215

- **Least Square Method**
- **Coefficient of Determination**
- **Model Assumptions**
- **Testing for Significance**
- **Covariance & Coefficient of Correlation**
- **Using the Estimated Regression –Equation for Estimation and Prediction**

2. LEAST SQUARE METHOD

Is a procedure for using sample data to find the estimated regression equation.

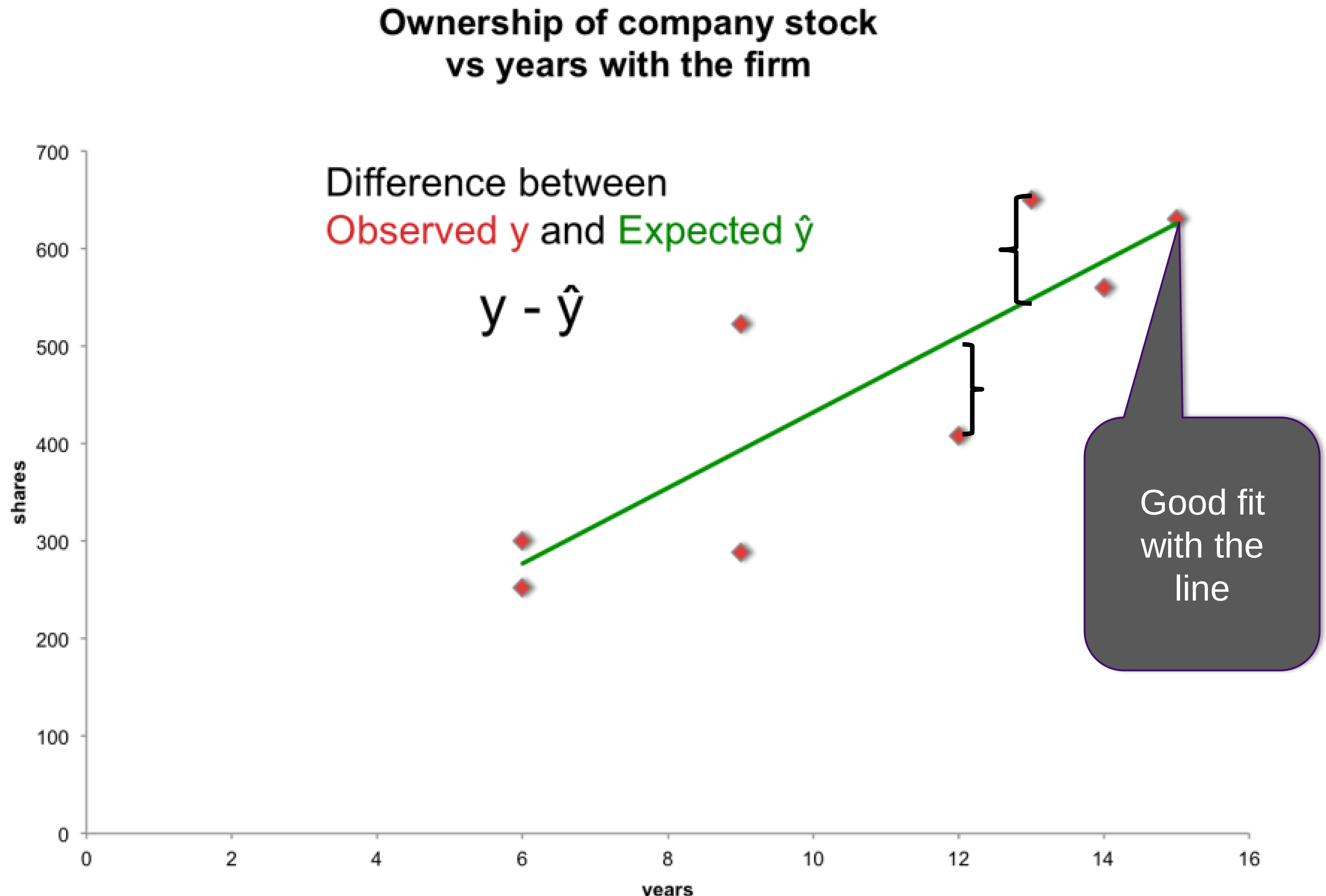
$$\hat{y} = \boxed{b_0} + \boxed{b_1} x$$

The goal to using the least square method

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}_i)}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sum x_i^2 - (\sum x_i)^2 / n}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Bowerman, et al. (2017) pp. 537



Ownership of company stock
vs years with the firm
Difference between
Observed y and Expected ŷ
y - ŷ
Good fit with the line
shares
years
700
600
500
400
300
200
100
0
0
2
4
6
8
10
12
14
16

LEAST SQUARES METHOD

Least Squares Criterion

$$\min \sum (y_i - \hat{y}_i)^2$$

where:

y_i = observed value of the dependent variable for the ith observation

$\hat{y}_i$ = estimated value of the dependent variable for the ith observation

Bowerman, et al. (2017) pp. 536

COMM215 First the Foundation, then Innovation

PROBLEM # 10.2

A scatter diagram includes the data points (x=2,y=10), (x=3, y=12), (x=4, y=20), and (x=5, y=16). Two regression lines are proposed: (1) *y* = 10 + x, and (2) *y* = 8 + 2x. Using the least-squares criterion, which of these regression lines is the better fit to the data? Why?

S.L.S.Ly
C215

$\hat{y} = 10 + x$

x	y	$\hat{y}$	$(y-\hat{y})$	$(y-\hat{y})^2$
2	10	12	-2	4
3	12	13	-1	1
4	20	14	6	36
5	16	15	1	1
				42

$$\hat{y} = 10 + \boxed{} = 10 + 2 = 12$$

PROBLEM # 10.2

A scatter diagram includes the data points (x=2,y=10), (x=3, y=12), (x=4, y=20), and (x=5, y=16). Two regression lines are proposed: (1) $y = 10 + x$, and (2) $y = 8 + 2x$. Using the least-squares criterion, which of these regression lines is the better fit to the data? Why?

S.L.S.Ly
C215

$\hat{y} = 8 + 2x$

x	**y**	**ŷ**	**(y-ŷ)**	**(y-ŷ)²**
2	10	12	-2	4
3	12	14	-2	4
4	20	16	4	16
5	16	18	-2	4
				28

PROBLEM # 10.2

A scatter diagram includes the data points (x=2,y=10), (x=3, y=12), (x=4, y=20), and (x=5, y=16). Two regression lines are proposed: (1) $y = 10 + x$, and (2) $y = 8 + 2x$. Using the least-squares criterion, which of these regression lines is the better fit to the data? Why?

S.L.S.Ly C215

For $\hat{y} = 10 + x$, the least square criterion = 42

For $\hat{y} = 8 + 2x$, the least square criterion = 28

LEAST SQUARES METHOD

Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

C215

where:

x_i = value of independent variable for *i*th observation

y_i = value of dependent variable for *i*th observation

$\bar{x}$ = mean value for independent variable

$\bar{y}$ = mean value for dependent variable

Bowerman, et al. (2017) pp. 536

LEAST SQUARES METHOD

- *y*-Intercept for the Estimated Regression Equation

$$b_0 = \bar{y} - b_1 \bar{x}$$

S. C215

PROBLEM # 10.7

For a sample of 8 employees, a personnel director has collected the following data on ownership of company stock versus years with the firm.

X= years	Y = shares
6	300
12	408
14	560
6	252
9	288
13	650
15	630
9	522

a. Determine the least-squares regression line and interpret its slope.

b. For an employee who has been with the firm 10 years, what is the predicted number of shares of stock owned?

S L S L y C215

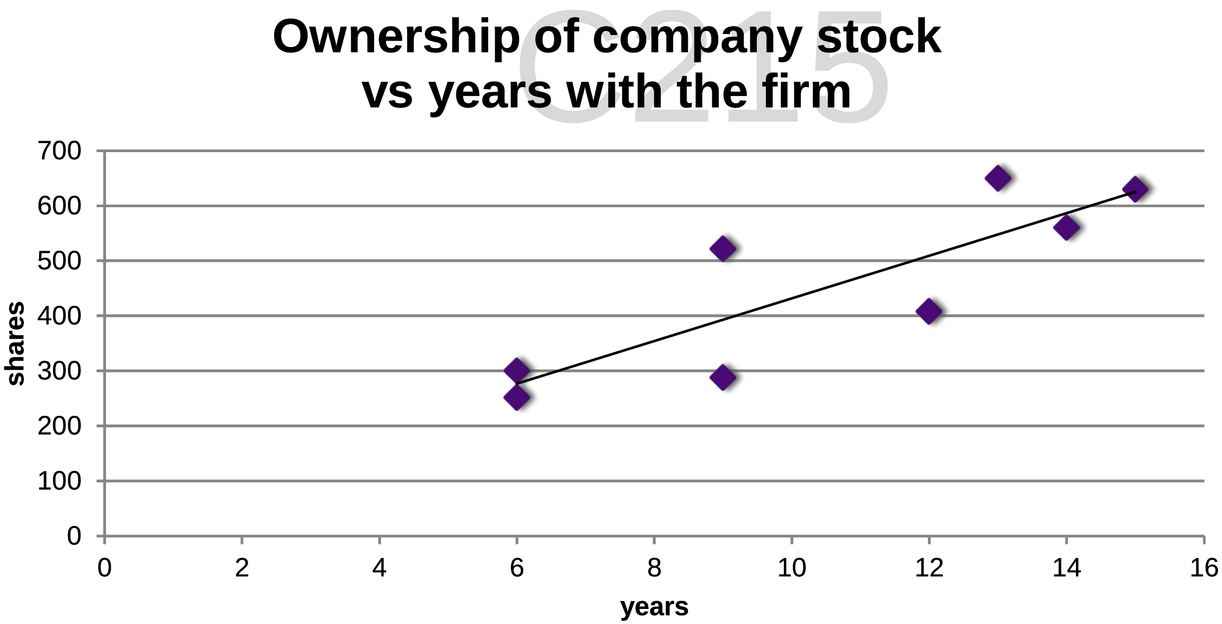
PROBLEM # 10.7

For a sample of 8 employees, a personnel director has collected the following data on ownership of company stock versus years with the firm.

a. Determine the least-squares regression line and interpret its slope.

S.L.S.Ly C215

$$\hat{y} = b_0 + b_1 x$$



X= years	Y = shares
6	300
12	408
14	560
6	252
9	288
13	650
15	630
9	522

PROBLEM # 10.7

X= years	Y = shares
6	300
12	408
14	560
6	252
9	288
13	650
15	630
9	522

For a sample of 8 employees, a personnel director has collected the following data on ownership of company stock versus years with the firm.

a. Determine the least-squares regression line and interpret its slope.

S.L.S.Ly
C215

$$\hat{y} = b_0 + b_1 x$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}_i)}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sum x_i^2 - (\sum x_i)^2 / n}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

PROBLEM # 10.7

For a sample of
collected the foll
versus years wit

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}_i)}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sum x_i^2 - (\sum x_i)^2 / n}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

X= years	Y = shares
6	300
12	408
14	560
6	252
9	288
13	650
15	630
9	522

a. Determine the least-squares regression line and interpret its slope.

x	y	xy	x^2
6.00	300.00	1800.00	36.00
12.00	408.00	4896.00	144.00
14.00	560.00	7840.00	196.00
6.00	252.00	1512.00	36.00
9.00	288.00	2592.00	81.00
13.00	650.00	8450.00	169.00
15.00	630.00	9450.00	225.00
9.00	522.00	4698.00	81.00
Σ x = 84.00	Σ y = 3610.00	Σ xy = 41,238.00	Σ x^2 = 968.00

PROBLEM # 10.7

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y}_i)}{\sum(x_i - \bar{x})^2} = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sum x_i^2 - (\sum x_i)^2 / n}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

For a sample of 8 employees, a personnel director has collected the following data on ownership of company stock versus years with the firm.

a. Determine the least-squares regression line and interpret its slope.

$$\hat{y} = b_0 + b_1 x$$

S.L.S.Ly
C215

Σ x = 84.00	**Σ y = 3610.00**	**Σ xy = 41,238.00**	**Σ x^2 = 968.00**

Do your calculations! Find bo and b1!

PROBLEM # 10.7

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}_i)}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sum x_i^2 - (\sum x_i)^2 / n}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

For a sample of 8 employees, a personnel director has collected the following data on ownership of company stock versus years with the firm.

a. Determine the least-squares regression line and interpret its slope.

$$\hat{y} = b_0 + b_1 x$$

S.L.S.Ly C215

Σ x = 84.00	**Σ y = 3610.00**	**Σ xy = 41,238.00**	**Σ x^2 = 968.00**

$$b_1 = \frac{\Sigma xy - [\Sigma x \, \Sigma y / n]}{\Sigma x^2 - [(\Sigma x)^2 / n]} = \frac{41{,}238 - [(84)\,(3610)/8]}{(968) - [(84)^2 / 8]}$$

$$= \frac{3333}{86} = 38.756$$

PROBLEM # 10.7

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}_i)}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sum x_i^2 - (\sum x_i)^2 / n}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

For a sample of 8 employees, a personnel director has collected the following data on ownership of company stock versus years with the firm.

a. Determine the least-squares regression line and interpret its slope. $\hat{y} = b_0 + b_1 x$

S.L.S.Ly

Σ x = 84.00	**Σ y = 3610.00**	**Σ xy = 41,238.00**	**Σ x^2 = 968.00**

C215

$b_0 = y_{mean} - b_1\, x_{mean} = 451.25 - 38.756\,(10.5)$ $= 44.314$

$y_{mean} = \Sigma\, y / n = 3610.00/8 = 451.25$

$x_{mean} = \Sigma\, x / n = 84 /8 = 10.5$

PROBLEM # 10.7

X= years	Y = shares
6	300
12	408
14	560
6	252
9	288
13	650
15	630
9	522

For a sample of 8 employees, a personnel director has collected the following data on ownership of company stock versus years with the firm.

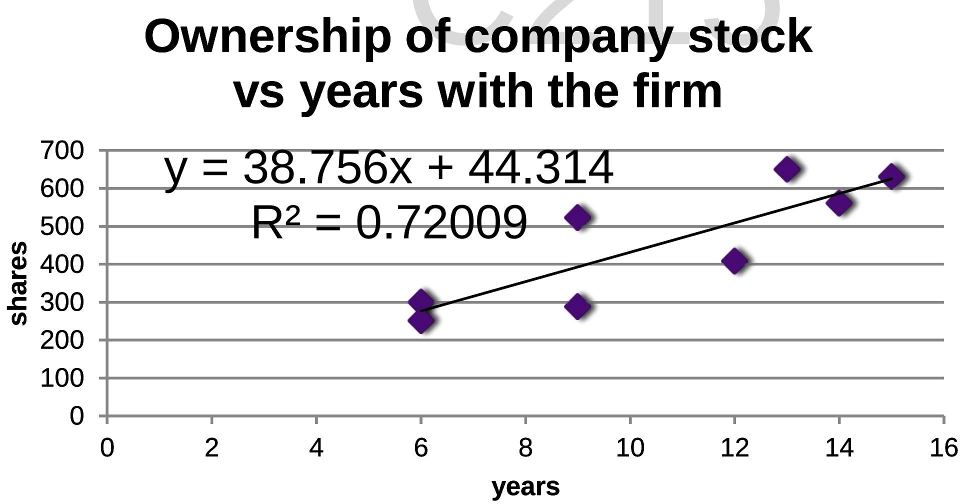
a. Determine the least-squares regression line and interpret its slope.

S.L.S.Ly
C215

$$\hat{y} = b_0 + b_1 x$$

b_1= 38.756

b_o= 44.314



For every additional year in the company, the employee will receive an additional 38.756 shares. When the employee starts working, they will receive a base of 44.312 shares.

PROBLEM # 10.7

For a sample of 8 employees, a personnel director has collected the following data on ownership of company stock versus years with the firm.

X= years	Y = shares
6	300
12	408
14	560
6	252
9	288
13	650
15	630
9	522

S.L.S.Ly
C215

$$\hat{y} = b_0 + b_1 x$$

	C	D	E	F	G	H	I
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.849					
5	R Square	0.720					
6	Adjusted R Square	0.673					
7	Standard Error	91.479					
8	Observations	8					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	129173.13	129173.13	15.436	0.008	
13	Residual	6	50210.37	8368.40			
14	Total	7	179383.50				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	44.3140	108.51	0.408	0.697	-221.197	309.825
18	Years	38.7558	9.86	3.929	0.008	14.618	62.893

PROBLEM # 10.7

X= years	Y = shares
6	300
12	408
14	560
6	252
9	288
13	650
15	630
9	522

For a sample of 8 employees, a personnel director has collected the following data on ownership of company stock versus years with the firm.

S.L.S.Ly

b. For an employee who has been with the firm

10 years, what is the predicted number of shares of stock owned?

C215

$$\hat{y} = 44.314 + 38.756\, x$$

$$\hat{y} = 44.314 + 38.756\,(10)$$

$$\hat{y} = 431.9$$

the predicted value of Shares will be 431.9

- Coefficient of Determination
- Model Assumptions
- Testing for Significance
- Covariance & Coefficient of Correlation
- Using the Estimated Regression –Equation for Estimation and Prediction

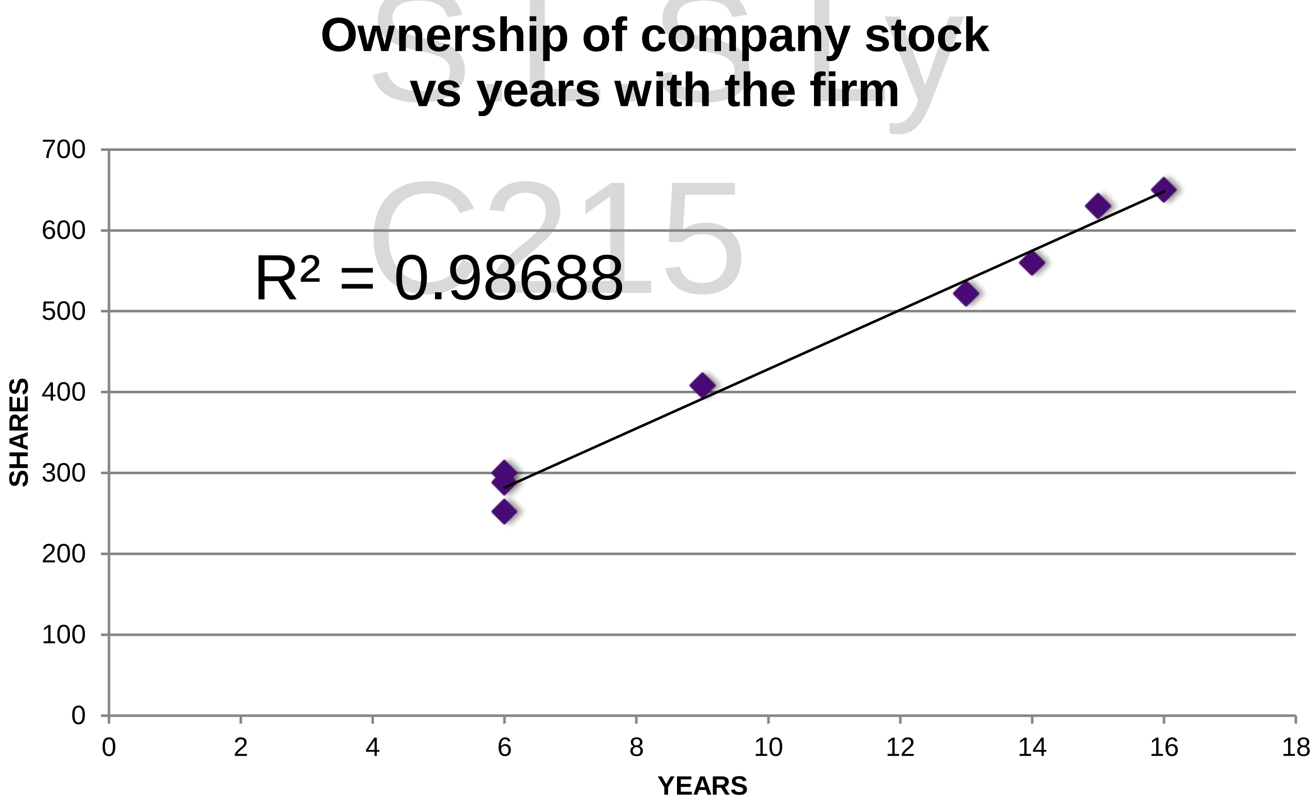
3. COEFFICIENT OF DETERMINATION

CORRELATION COEFFICIENT

Bowerman, et al. (2017) pp. 543

COEFFICIENT OF DETERMINATION

x	y
6	300
9	408
14	560
6	252
6	288
16	650
15	630
13	522



SLS Ly C215

COEFFICIENT OF DETERMINATION

PROVIDES A MEASURE OF THE GOODNESS OF FIT FOR THE ESTIMATED REGRESSION EQUATION

S.L.S.Ly
C215

IN OTHER WORDS

DOES THE INDEPENDENT VARIABLE

EXPLAIN

THE DEPENDENT VARIABLE WELL?

Explain 40% of the grade

FINAL COMM 215 Grade (y) =

Explain 15% of the grade

Study Time (x_1) +

These 4 Variables explain 79% of your grade!

Work Time (x_2) +

# of hours spent on Facebook (x_3) +

Explain 9% of the grade

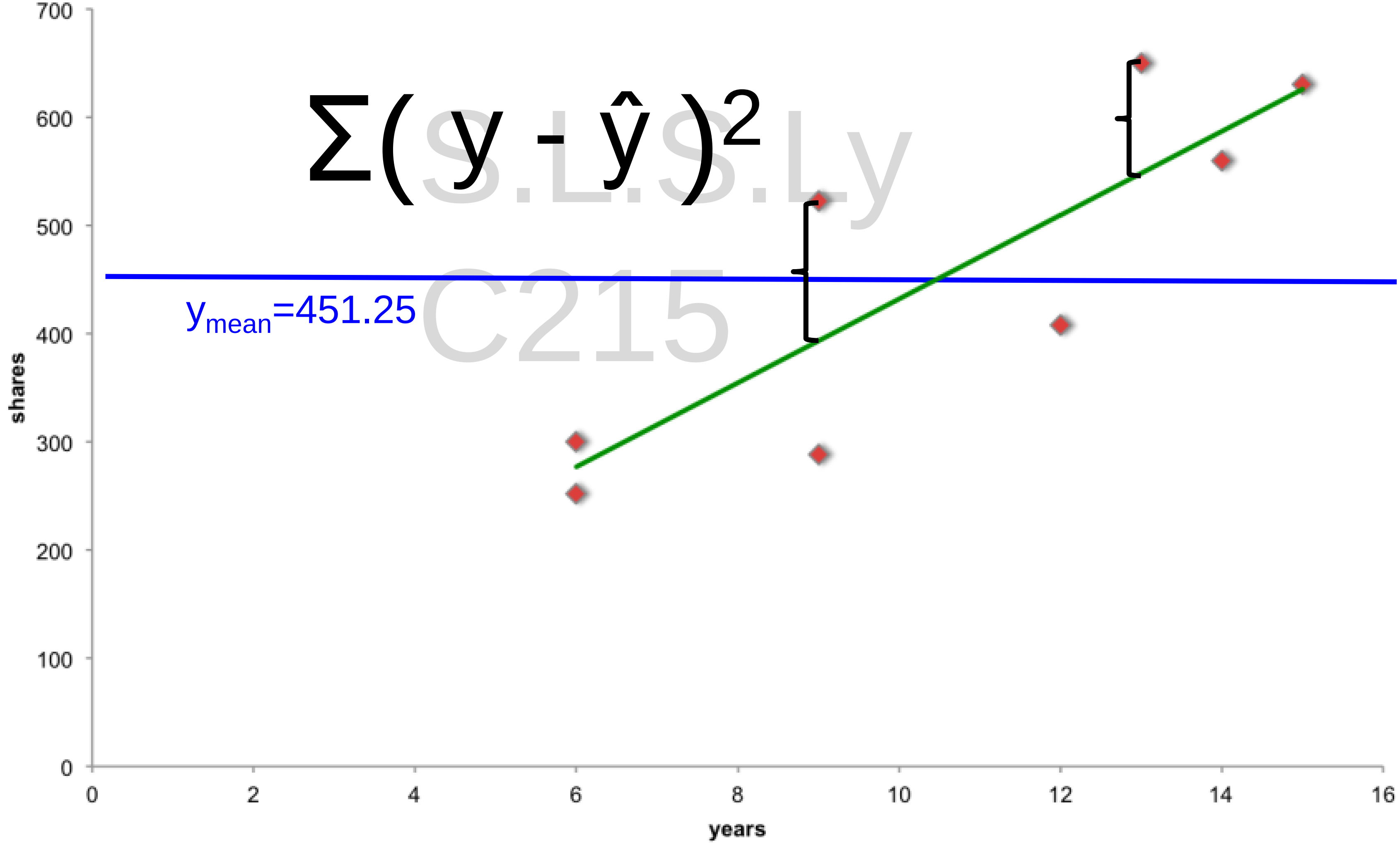
Family Time (x_4) +

anything else you can think of ($x_{...}$) …

Explain 15% of the grade

Sum of Squares due to Error (SSE)

Difference between Observed y and Expected ŷ

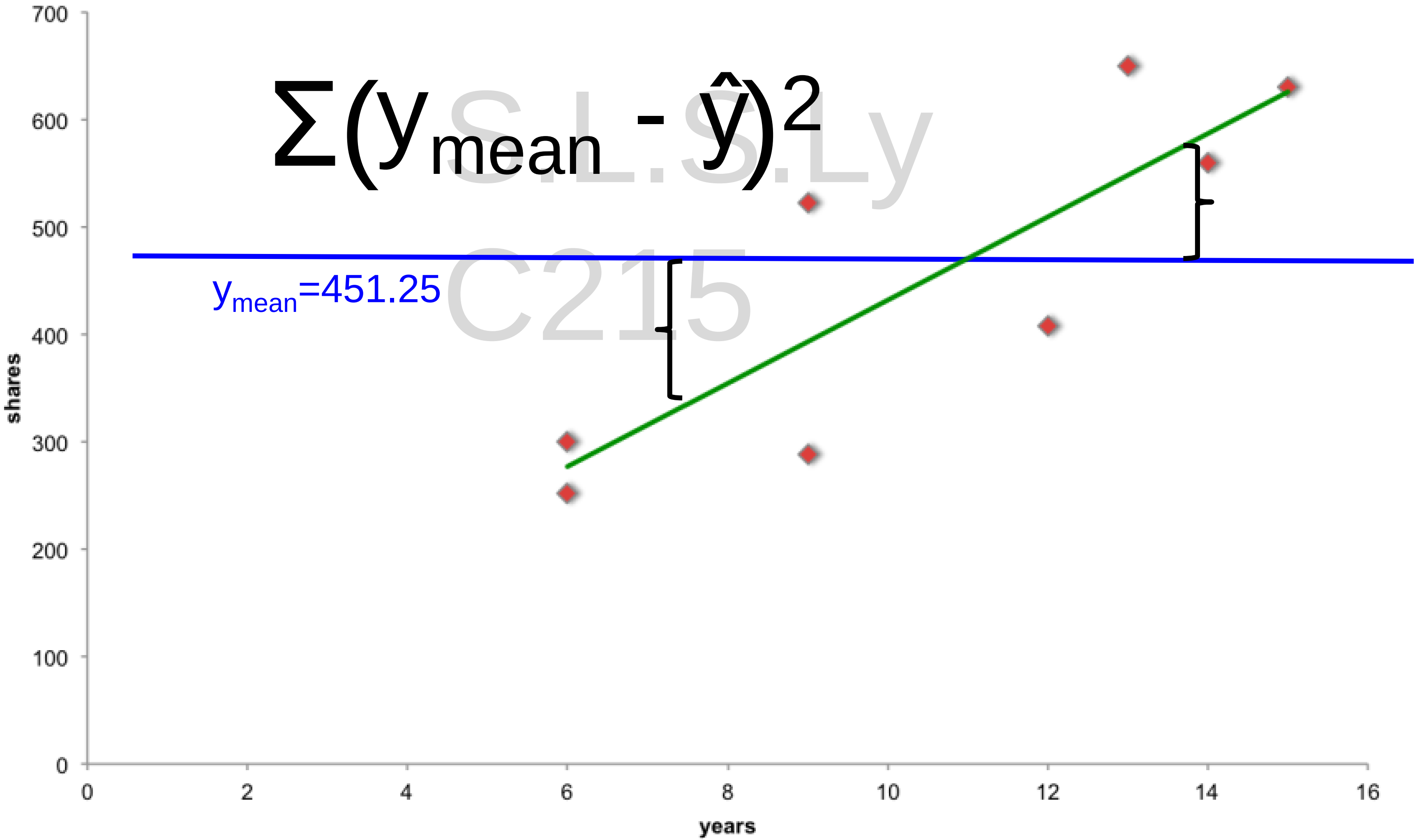


S.L.S.Ly
C215

Bowerman, et al. (2017) pp. 546

Sum of Squares due to Regression (SSR)

Difference between Mean of y and Expected ŷ



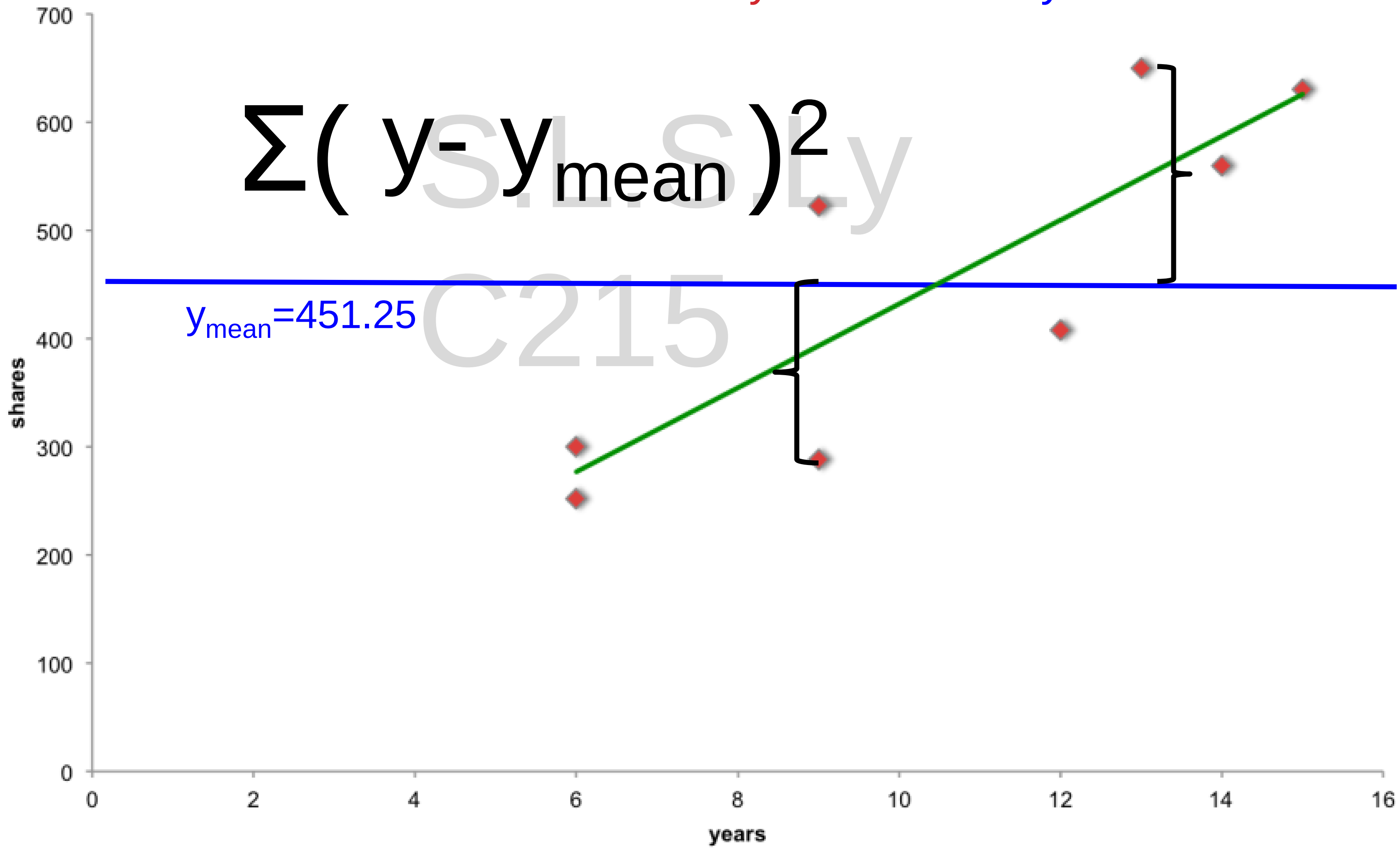
Bowerman, et al. (2017) pp. 546

COMM215 First the Foundation, then Innovation

Total Sum of Squares (SST)

Difference between Observed y and Mean of y

$$\Sigma(y - y_{mean})^2$$



Bowerman, et al. (2017) pp. 546

COEFFICIENT OF DETERMINATION

Relationship Among SST, SSR, SSE

$$SST = SSR + SSE$$

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

where:

SST = total sum of squares

SSR = sum of squares due to regression

SSE = sum of squares due to error

COEFFICIENT OF DETERMINATION

- The <u>coefficient of determination</u> is:

$$r^2 = \text{SSR}/\text{SST}$$

where:

SSR = sum of squares due to regression
= explained variation

SST = total sum of squares
= total variation

Bowerman, et al. (2017) pp. 545

COEFFICIENT OF DETERMINATION (R^2)

Expresses proportion of the variation in the dependent variable (y) that is explained by the regression line:

$\hat{y} = b_0+b_1x_1$

COEFFICIENT OF CORRELATION (R)

Describes both the direction and the strength of the linear relationship between two variables

$r = (\text{sign of } b_1) \sqrt{r^2}$

Bowerman, et al. (2017) pp. 546

PROBLEM #10.9

For a set of data, the total variation or sum of squares for y is

SST = 143.0, and error sum of squares is SSE = 24.0. What proportion of the variation in y is explained by the regression equation?

S.L.S.Ly
C215

PROBLEM #10.9

We also know SST = SSR + SSE

For a set of data, the total variation or sum of squares for y is

SST = 143.0, and error sum of squares is SSE = 24.0. What proportion of the variation in y is explained by the regression equation?

S.L.S.Ly C215

We are asked to find the coefficient of determination :

R^2 = SSR/SST

PROBLEM #10.4

We also know SST = SSR + SSE

For a set of data, the total variation or sum of squares for y is

SST = 143.0, and error sum of squares is SSE = 24.0. What proportion of the variation in y is explained by the regression equation?

S.L.S.Ly

C215

SST – SSE = SSR
143.0 – 24.0 = 119

We are asked to find the coefficient of determination :

119 /143.0 = 0.832 = 83.2%

R^2 = SSR/SST

Interpretation:
83.2% of the variation in y is explained by x
83.2% of the variation in ____DV is explained by IV, IV, IV

- **Model Assumptions**
- **Testing for Significance**
- **Covariance & Coefficient of Correlation**
- **Using the Estimated Regression –Equation for Estimation and Prediction**

4. MODEL ASSUMPTIONS

Before conducting regression analysis…

S.L.S.Ly
C215

What determines an appropriate model for the relationship between the dependent and the independent variable(s).

Even if the data fits well , the estimated regression equation <u>should not be used</u> until further analysis of

how appropriate the assumed model is.

One way to determine is to test for significance.

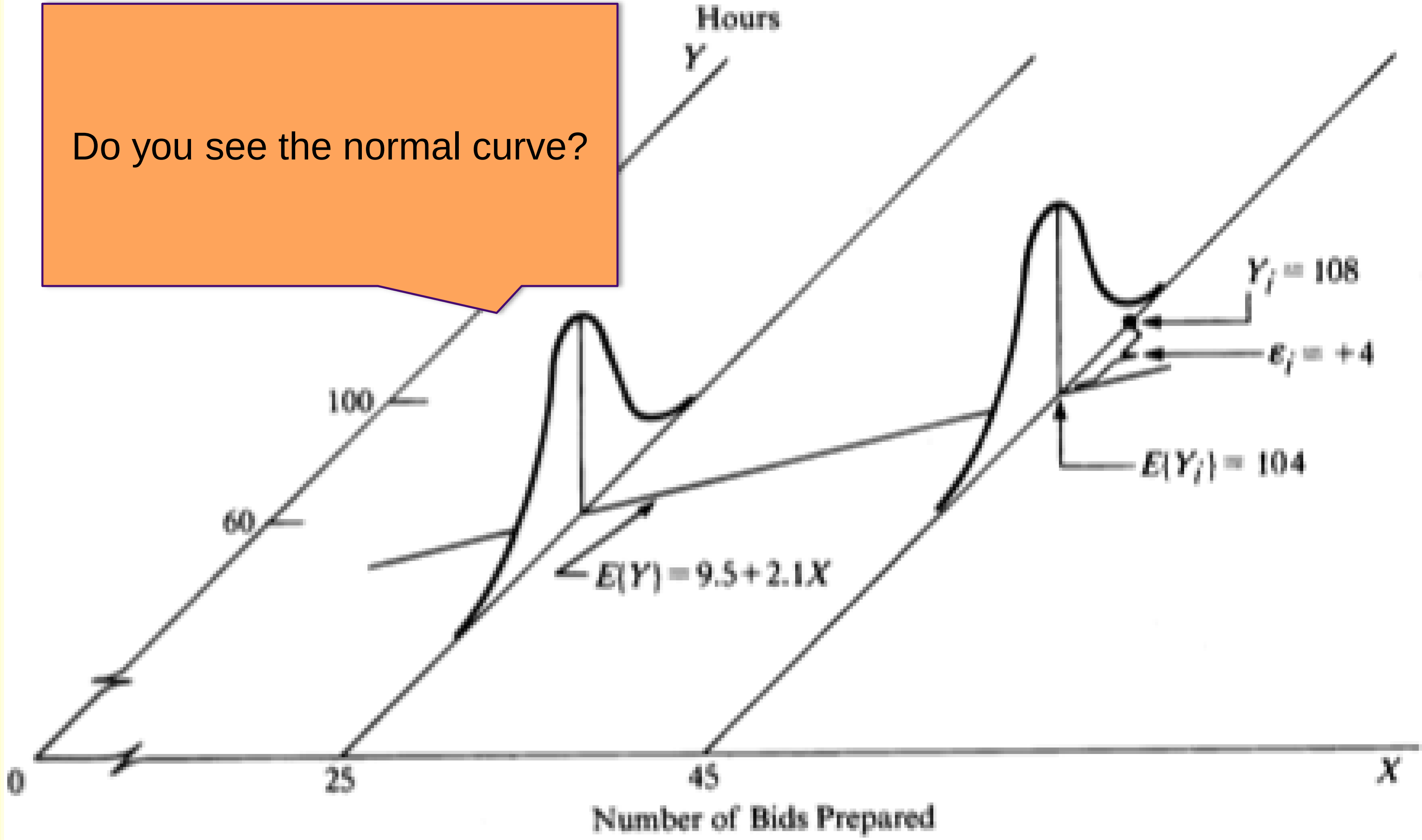
S.L.S.Ly
C215

ASSUMPTIONS- ERROR TERM

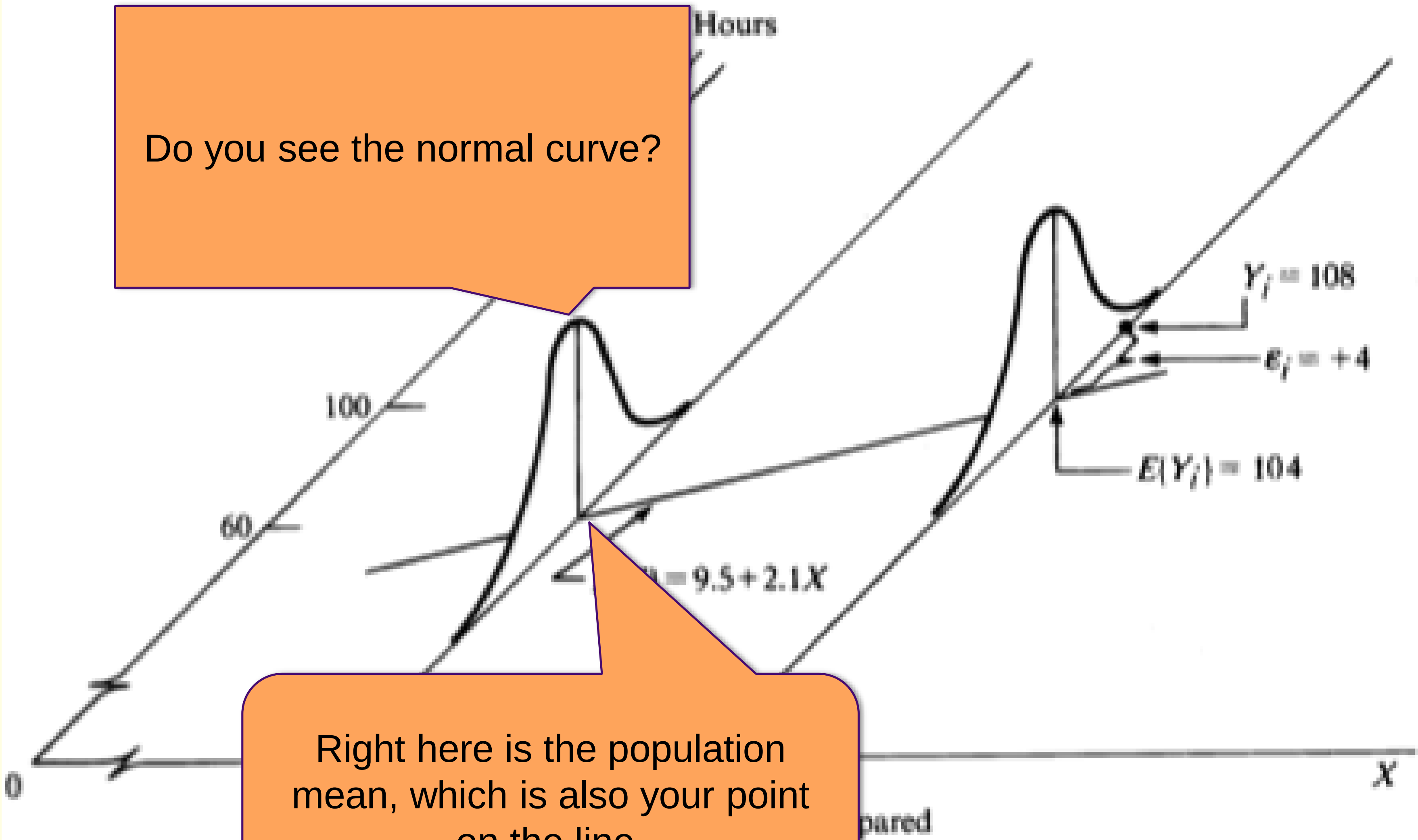
1. The error term ε is a random variable with an expected value of zero. In estimating an element that is unpredictable, it is best to assume it to be zero.
2. The variance of ε, denoted by σ^2 is the same for all values of x.
3. The values of error are independent.
4. The error term is normally distributed.

Bowerman, et al. (2017) pp. 548

Do you see the normal curve?

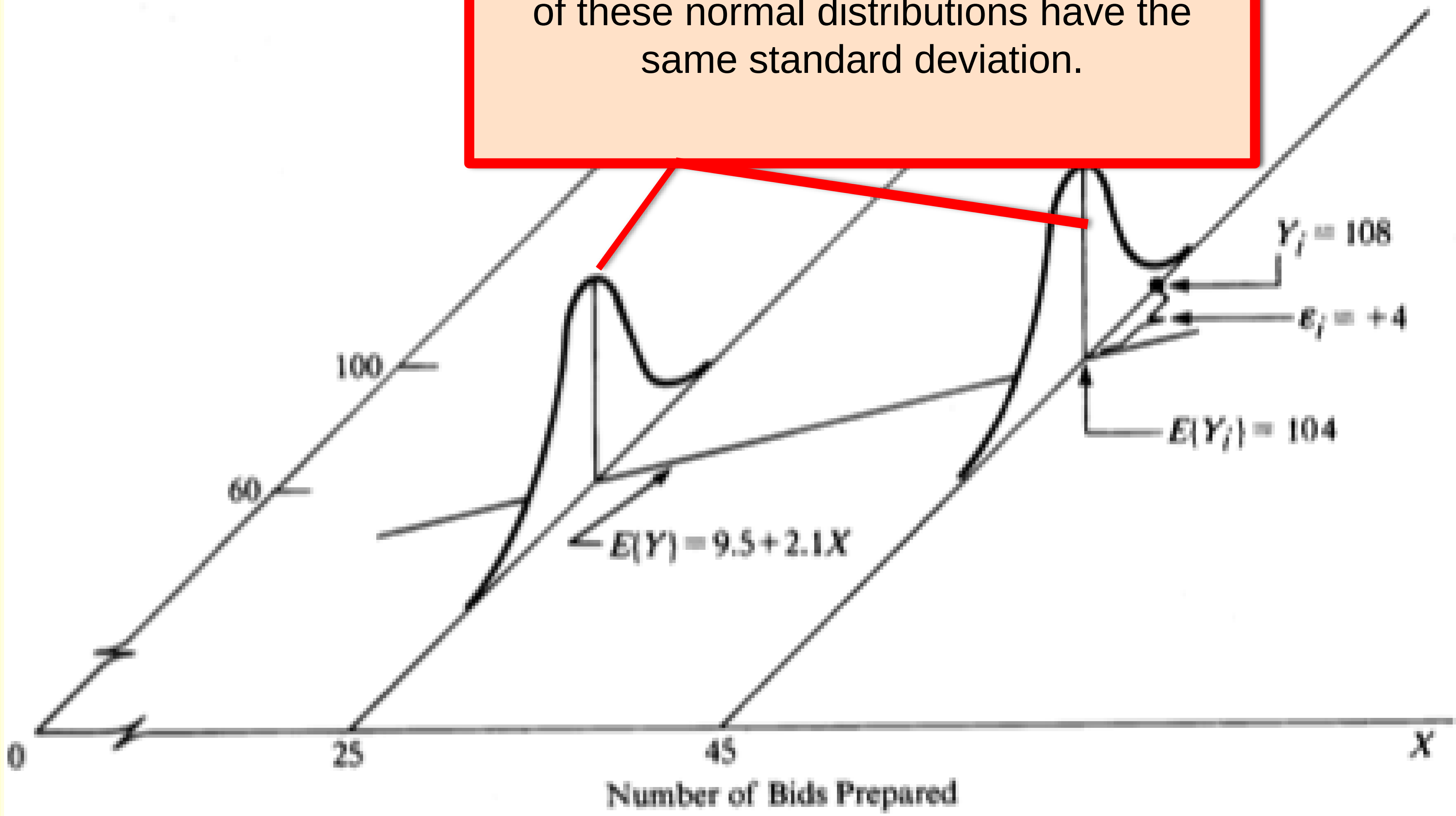


Do you see the normal curve?



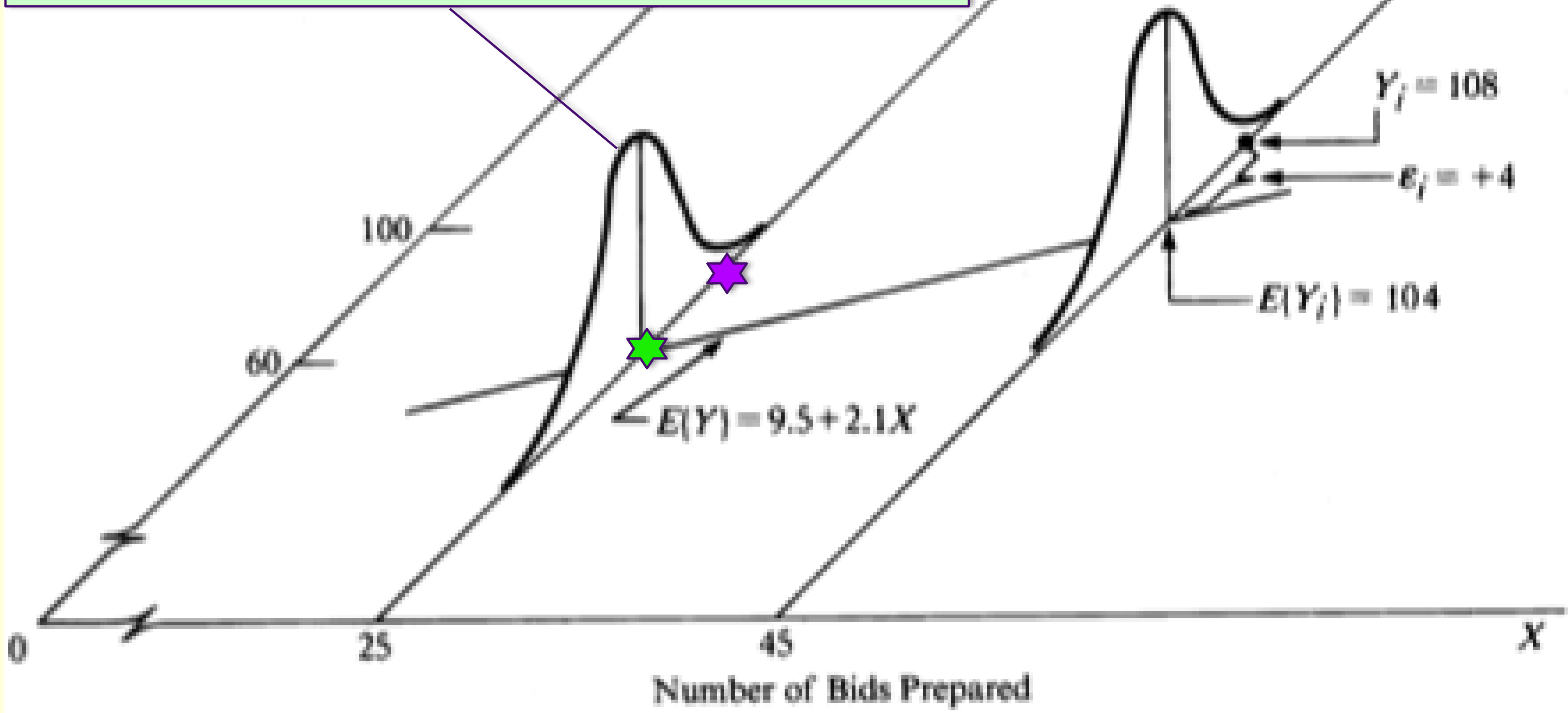
Right here is the population mean, which is also your point on the line.

Assumption # 2: The variance/ standard deviation for each value x is the same. All of these normal distributions have the same standard deviation.



Assumptions # 3 & 4: the error terms are independent and are normally distributed.

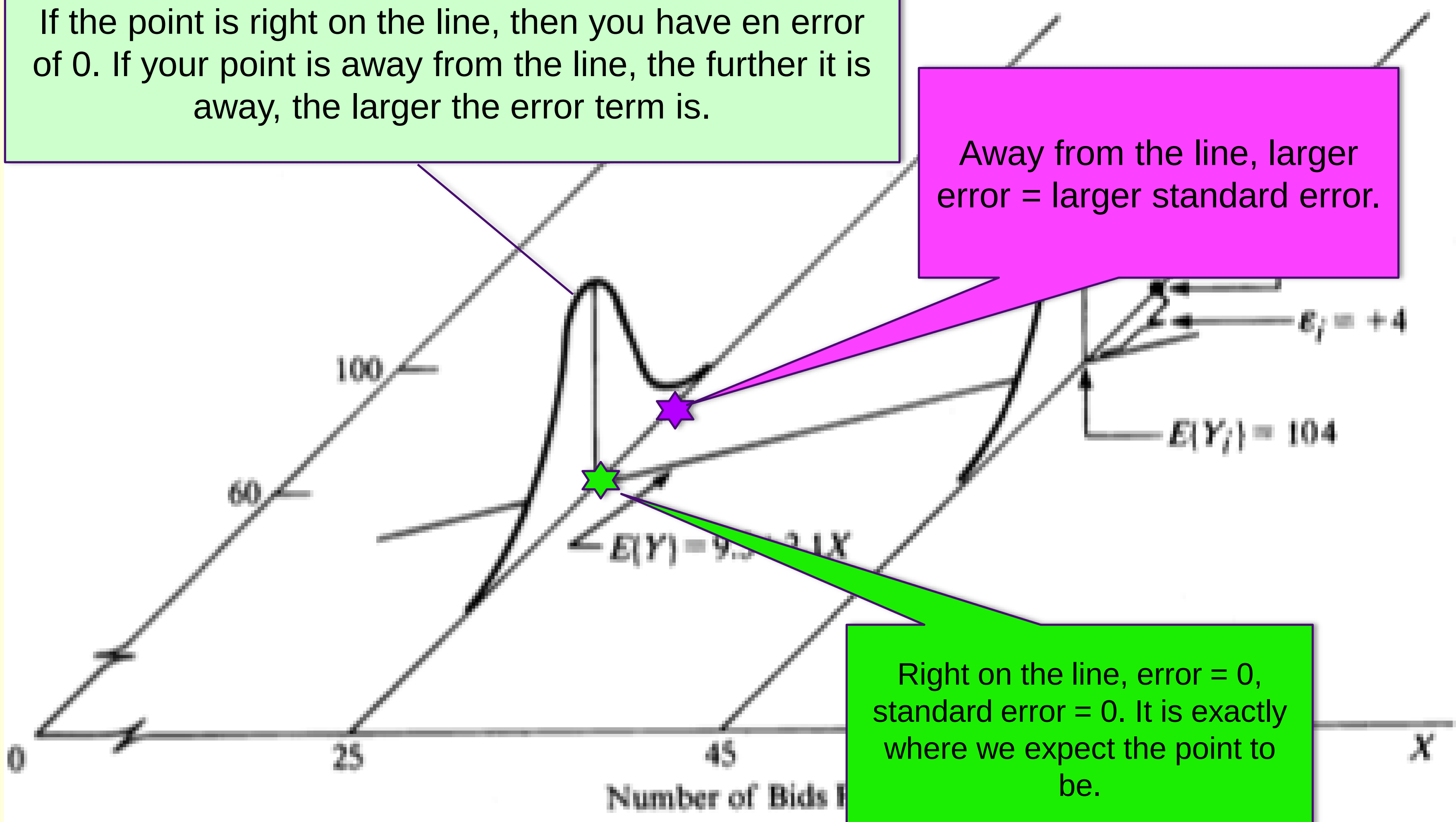
If the point is right on the line, then you have en error of 0. If your point is away from the line, the further it is away, the larger the error term is.



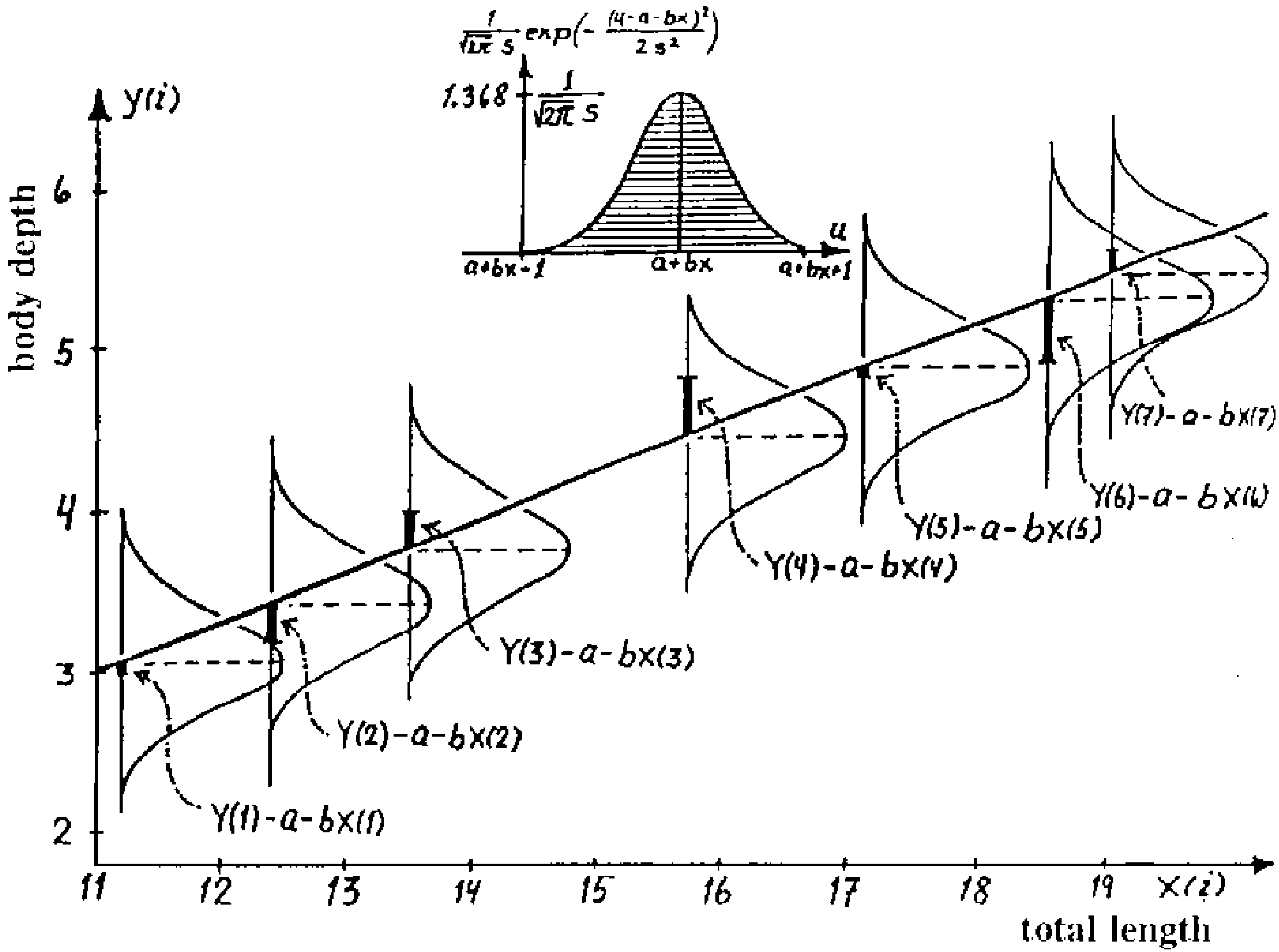
Assumptions # 3 & 4: the error terms are independent and are normally distributed.

If the point is right on the line, then you have en error of 0. If your point is away from the line, the further it is away, the larger the error term is.

Away from the line, larger error = larger standard error.



Right on the line, error = 0, standard error = 0. It is exactly where we expect the point to be.



$\frac{1}{\sqrt{2\pi}\,s} \exp\left(-\frac{(u-a-bx)^2}{2s^2}\right)$
1.368
$\frac{1}{\sqrt{2\pi}\,s}$
u
a+bx-1
a+bx
a+bx+1
y(i)
body depth
6
5
4
3
2
11
12
13
14
15
16
17
18
19
x(i)
total length
Y(1)-a-bx(1)
Y(2)-a-bx(2)
Y(3)-a-bx(3)
Y(4)-a-bx(4)
Y(5)-a-bx(5)
Y(6)-a-bx(6)
Y(7)-a-bx(7)

- Testing for Significance
- Covariance & Coefficient of Correlation
- Using the Estimated Regression –Equation for Estimation and Prediction

5. TESTING FOR SIGNIFICANCE

ESTIMATE σ^2

TESTING FOR SIGNIFICANCE

CONFIDENCE INTERVAL FOR B_1

S.L.S.Ly
C215

TESTING FOR SIGNIFICANCE

An Estimate of σ^2

- The mean square error (MSE) provides the estimate of σ^2, and the notation s^2 is also used.

$$s^2 = \text{MSE} = \text{SSE}/(n-2)$$

- where:

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

$$\text{SSE} = \sum y_i^2 - b_0 \sum y_i - b_1 \sum x_i y_i$$

Bowerman, et al. (2017) pp. 551

TESTING FOR SIGNIFICANCE

An Estimate of σ

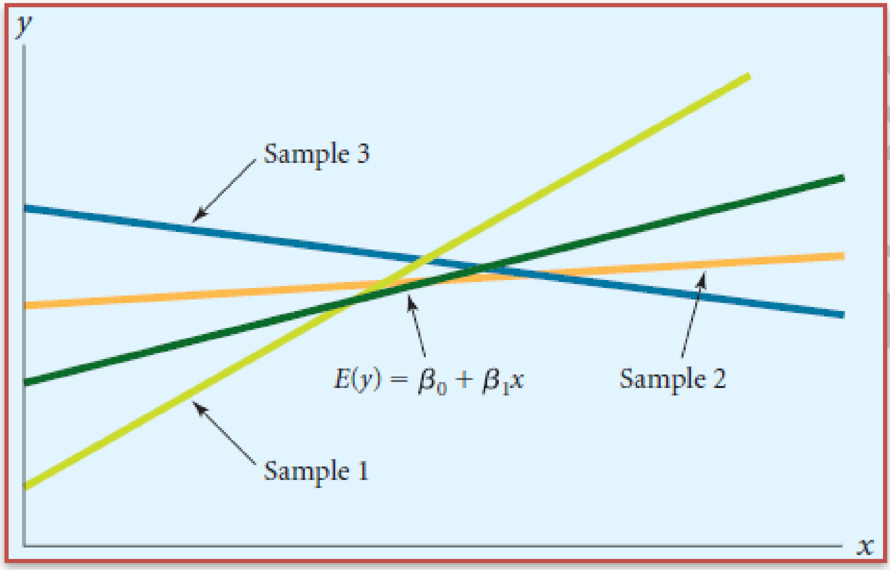
- To estimate σ we take the square root of σ^2.
- The resulting s is called the <u>standard error of the estimate</u>.

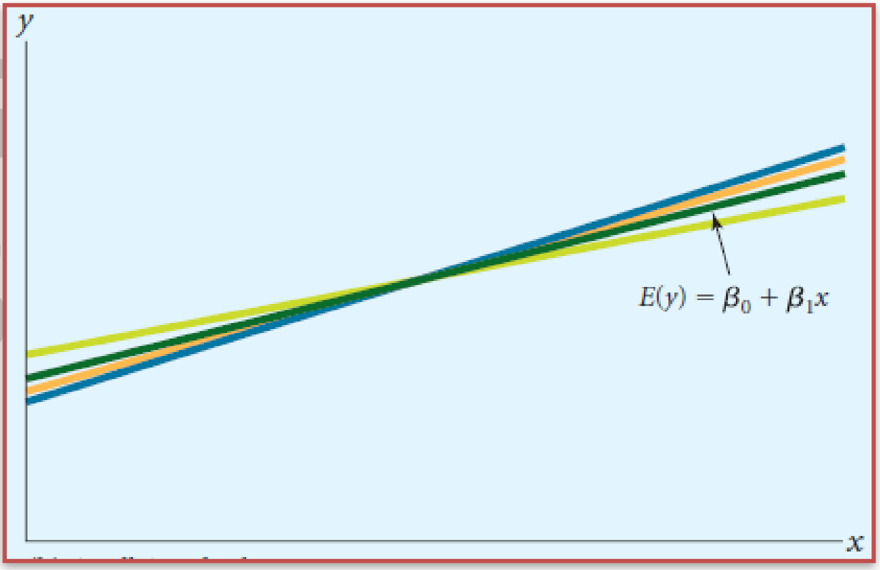
S.L.S.Ly

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}}$$

Why is this n-2?, it actually is n-k-1.
In a simple linear regression, since you always have 1 independent variable (k=1), therefore automatically it becomes n-1-1.

STANDARD ERROR OF THE ESTIMATE





Large Standard Error

Small Standard Error

TESTING FOR SIGNIFICANCE

- To test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of β_1 is zero.

- Two tests are commonly used:

  t Test and F Test

- Both the t test and F test require an estimate of σ^2, the variance of ε in the regression model.

TESTING FOR SIGNIFICANCE: *T* TEST

Hypotheses

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

S. C215

Test Statistic

$$t = \frac{b_1}{s_{b_1}}$$ where $$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}}$$

Bowerman, et al. (2017) pp. 552

TESTING FOR SIGNIFICANCE: *T* TEST

1. Set up Hypotheses. $H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$
2. What is the appropriate test statistic to use?. $t = \dfrac{b_1}{s_{b_1}}$
3. Calculate the test statistic value.
4. Find the critical value for the test statistic. $\alpha = .05$
5. Define the decision rule
6. Make your decision
7. Interpret the conclusion in context

CONFIDENCE INTERVAL FOR β_1

The form of a confidence interval for β_1 is:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

b_1 is the point estimator

$t_{\alpha/2} s_{b_1}$ is the margin of error

where $t_{\alpha/2}$ is the t value providing an area of $\alpha/2$ in the upper tail of a t distribution with n - 2 degrees of freedom

S.E.C.Ly
C215

CONFIDENCE INTERVAL FOR β_1

- We can use a 95% confidence interval for β_1 to test the hypotheses just used in the t test.
- H_0 is rejected if the hypothesized value of β_1 is not included in the confidence interval for β_1.

S.L.S.Ly
C215

CONFIDENCE INTERVAL FOR β_1

- Rejection Rule

  Reject H_0 if 0 is not included in the confidence interval for β_1.

- 95% Confidence Interval for β_1

$$b_1 \pm t_{\alpha/2} s_{b_1} = 5.0 \pm 2.048(2.25) = 5.0 \pm 4.608$$

or 0.392 to 9.608

- Conclusion

  0 is not included in the confidence interval. Reject H_0

S.L.S.Ly C215

SOME CAUTIONS ABOUT THE INTERPRETATION OF SIGNIFICANCE TESTS

SLSLy C215

- Rejecting H_0: $\beta_1 = 0$ and concluding that the relationship between x and y is significant does not enable us to conclude that a cause-and-effect relationship is present between x and y.

- Just because we are able to reject H_0: $\beta_1 = 0$ and demonstrate statistical significance does not enable us to conclude that there is a linear relationship between x and y.

7 SECTIONS

1. ~~SIMPLE LINEAR REGRESSION MODEL 14.2~~
2. ~~LEAST SQUARE METHOD 14.2~~
3. ~~COEFFICIENT OF DETERMINATION 14.2~~
4. ~~MODEL ASSUMPTIONS 14.2~~
5. ~~TESTING FOR SIGNIFICANCE 14.3~~
6. COVARIANCE & COEFFICIENT OF CORRELATION 14.1
7. USING THE ESTIMATED REGRESSION 11.7 EQUATION FOR ESTIMATION AND PREDICTION

S.L.S.Ly
C215

What is a standard error of the estimate?

What is a standard error of the slope?

S.L.S.Ly
C215

$$s = \sqrt{\mathrm{MSE}} = \sqrt{\frac{\mathrm{SSE}}{n-2}}$$

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}}$$

PROBLEM # 10.15

At 5% level significance. The manager of Colonial Furniture has been reviewing weekly advertising expenditures. During the past 6 months, all advertisements for the store have appeared in the local newspaper. The number of ads per week has varied from one to seven. The store's sales staff has been tracking the number of customers who enter the store each week. The number of ads and the number of customers per week for the past 26 weeks were recorded.

a. Determine the sample regression line

b. Interpret the coefficients.

c. Can the manager infer that the larger the number of ads, the larger the number of customers?

d. Find and interpret the coefficient of determination.

e. In your opinion, is it worthwhile exercise to use the regression equation to predict the number of customers who will enter the store, given that Colonial intends to advertise five times in the newspaper? If so, find the 95% prediction interval. If not, explain why not.

COMM215 First the Foundation, then Innovation

PROBLEM # 10.5

Ads	Customer
5	353
6	319
3	440
2	332
4	172
2	331
4	344
2	483
4	329
2	532
7	496
5	393
4	376
7	372
2	512
5	254
5	459
2	153
1	426
6	566
6	596
5	395
6	676
3	194
2	135
7	367

At 5% level significance. The manager of Colonial Furniture has been reviewing weekly advertising expenditures. During the past 6 months, all advertisements for the store have appeared in the local newspaper. The number of ads per week has varied from one to seven. The store's sales staff has been tracking the number of customers who enter the store each week. The number of ads and the number of customers per week for the past 26 weeks were recorded.

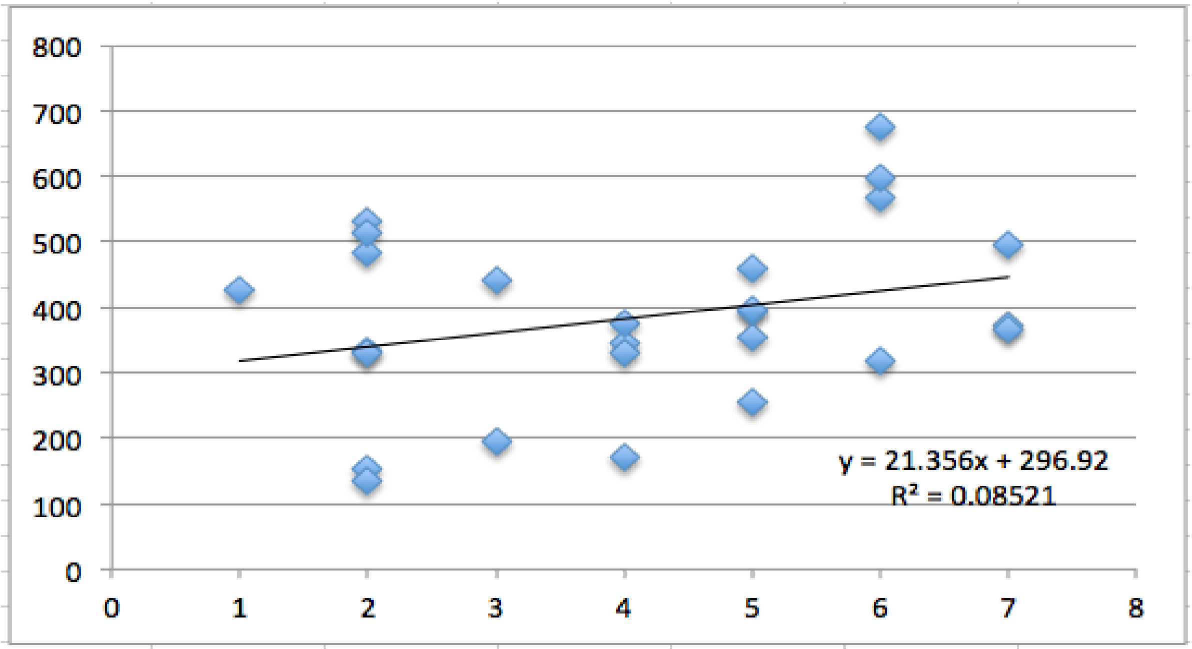
a. Determine the sample regression line

b. Interpret the coefficients.

$$\hat{y} = 296.92 + 21.356x$$

On average each additional ad generates 21.36 customers.

S.L.S.Ly C215



PROBLEM # 10.5

c. Can the manager infer that the larger the number of ads, the larger the number of customers?

S.L.S.Ly
C215

1. Set up the hypotheses:

H_o: $\beta_1 = 0$; H_a: $\beta_1 > 0$

2. What is the appropriate test statistics to use?

One tail t-test, $\alpha=0.05$

Testing for Significance: t Test

1. Set up Hypotheses.
2. What is the appropriate test statistic to use?. $$t = \frac{b_1}{s_{b_1}}$$
3. Calculate the test statistic value.
4. Find the critical value for the test statistic. $\alpha = .05$
5. Define the decision rule
6. Make your decision
7. Interpret the conclusion in context

3. Calculate the test statistic value.

5

$$t = \frac{b_1}{s_{b_1}}$$

4

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}}$$

3

$$\Sigma(x_i - \bar{x})^2 = (\sum x_i^2) - \frac{(\sum x_i)^2}{n} = SS_{xx}$$

2

$$s_e = \sqrt{\frac{SSE}{n-k-1}}$$

Standard Error of Estimate (s_ε)

1

SSE

1

PROBLEM # 10.5

At 5% level significance. The manager of Colonial Furniture has been reviewing weekly advertising expenditures. During the past 6 months, all advertisements for the store have appeared in the local newspaper. The number of ads per week has varied from one to seven. The store's sales staff has been tracking the number of customers who enter the store each week. The number of ads and the number of customers per week for the past 26 weeks were recorded.

c. Can the manager infer that the larger the number of ads, the larger the number of customers?

$$SSE = \Sigma(y-\hat{y})^2 = 424281.17$$

Ads x	Customer y	$\hat{y} = 296.92 + 21.356x$	$y-\hat{y}$	$(y-\hat{y})^2$
5.00	353.00	403.70	-50.70	2570.49
6.00	319.00	425.06	-106.06	11247.88
3.00	440.00	360.99	79.01	6242.90
2.00	332.00	339.63	-7.63	58.25
4.00	172.00	382.34	-210.34	44244.60
2.00	331.00	339.63	-8.63	74.51
4.00	344.00	382.34	-38.34	1470.26
2.00	483.00	339.63	143.37	20554.38
4.00	329.00	382.34	-53.34	2845.58
2.00	532.00	339.63	192.37	37005.45
7.00	496.00	446.41	49.59	2458.97
5.00	393.00	403.70	-10.70	114.49
4.00	376.00	382.34	-6.34	40.25
7.00	372.00	446.41	-74.41	5537.15
2.00	512.00	339.63	172.37	29710.73
5.00	254.00	403.70	-149.70	22410.09
5.00	459.00	403.70	55.30	3058.09
2.00	153.00	339.63	-186.63	34831.50
1.00	426.00	318.28	107.72	11604.46
6.00	566.00	425.06	140.94	19865.21
6.00	596.00	425.06	170.94	29221.85
5.00	395.00	403.70	-8.70	75.69
6.00	676.00	425.06	250.94	62972.89
3.00	194.00	360.99	-166.99	27884.99
2.00	135.00	339.63	-204.63	41874.26
7.00	367.00	446.41	-79.41	6306.27
				424281.17

$$SSE = \Sigma(y-\hat{y})^2$$

S.L.S C215

PROBLEM # 10.5

At 5% level significance. The manager of Colonial Furniture has been reviewing weekly advertising expenditures. During the past 6 months, all advertisements for the store have appeared in the local newspaper. The number of ads per week has varied from one to seven. The store's sales staff has been tracking the number of customers who enter the store each week. The number of ads and the number of customers per week for the past 26 weeks were recorded.

c. Can the manager infer that the larger the number of ads, the larger the number of customers?

$$SSE = \Sigma(y-\hat{y})^2 = 424281.17$$

$$s = \sqrt{\frac{SSE}{n-k-1}}$$

$$s = \sqrt{\frac{424281.17}{26-1-1}}$$

$$s = 132.96$$

S.L.S.Ly C215

PROBLEM # 10.5

At 5% level significance. The manager of Colonial Furniture has been reviewing weekly advertising expenditures. During the past 6 months, all advertisements for the store have appeared in the local newspaper. The number of ads per week has varied from one to seven. The store's sales staff has been tracking the number of customers who enter the store each week. The number of ads and the number of customers per week for the past 26 weeks were recorded.

c. Can the manager infer that the larger the number of ads, the larger the number of customers?

Ads x	x-xbar	(x-xbar)²
5.00	0.88	0.7744
6.00	1.88	3.5344
3.00	-1.12	1.2544
2.00	-2.12	4.4944
4.00	-0.12	0.0144
2.00	-2.12	4.4944
4.00	-0.12	0.0144
2.00	-2.12	4.4944
4.00	-0.12	0.0144
2.00	-2.12	4.4944
7.00	2.88	8.2944
5.00	0.88	0.7744
4.00	-0.12	0.0144
7.00	2.88	8.2944
2.00	-2.12	4.4944
5.00	0.88	0.7744
5.00	0.88	0.7744
2.00	-2.12	4.4944
1.00	-3.12	9.7344
6.00	1.88	3.5344
6.00	1.88	3.5344
5.00	0.88	0.7744
6.00	1.88	3.5344
3.00	-1.12	1.2544
2.00	-2.12	4.4944
7.00	2.88	8.2944
4.12		**86.6544**

4

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} \quad s = 132.96$$

$$s_{b_1} = \frac{132.96}{\sqrt{86.6544}} = 14.28$$

3

$$SS_{xx} = \Sigma(x_i - \bar{x})^2$$

PROBLEM # 10.5

At 5% level significance. The manager of Colonial Furniture has been reviewing weekly advertising expenditures. During the past 6 months, all advertisements for the store have appeared in the local newspaper. The number of ads per week has varied from one to seven. The store's sales staff has been tracking the number of customers who enter the store each week. The number of ads and the number of customers per week for the past 26 weeks were recorded.

c. Can the manager infer that the larger the number of ads, the larger the number of customers?

S.L.S.Ly C215

$$t = \frac{b_1}{s_{b_1}}$$

= + 21.356

= 14.28

$$t = \quad 1.4955$$

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} = 14.28$$

ŷ = 296.92 + 21.356x

PROBLEM # 10.5

3. Calculate the test statistic value.

At 5% level significance. The manager of Colonial Furniture has been reviewing weekly advertising expenditures. During the past 6 months, all advertisements for the store have appeared in the local newspaper. The number of ads per week has varied from one to seven. The store's sales staff has been tracking the number of customers who enter the store each week. The number of ads and the number of customers per week for the past 26 weeks were recorded.

c. Can the manager infer that the larger the number of ads, the larger the number of customers?

S.L.S.Ly
C215

$$t = \frac{b_1 - \beta_1}{s_{b1}} = \frac{21.356 - 0}{14.28} = 1.496$$

PROBLEM # 10.5

4. Find the critical value of the test statistics

$t_{\alpha,\ n\text{-}k\text{-}1} = t_{0.05,24} = 1.711$

S.L.S.Ly
C215

5. Define the decision rule

Reject Ho, if $t_{observed} > t_{critical}$, otherwise do not reject.

6. Make your decision

Since $t_{observed} = 1.4955 < t_{critical}$, then we do not reject Ho

7. Interpret in the context

There is not enough evidence to conclude that the larger the number of ads the larger the number of customers.

- Covariance & Coefficient of Correlation

Using the Estimated Regression –Equation for Estimation and Prediction

6. COVARIANCE & COEFFICIENT OF CORRELATION

Covariance

Interpretation of the covariance

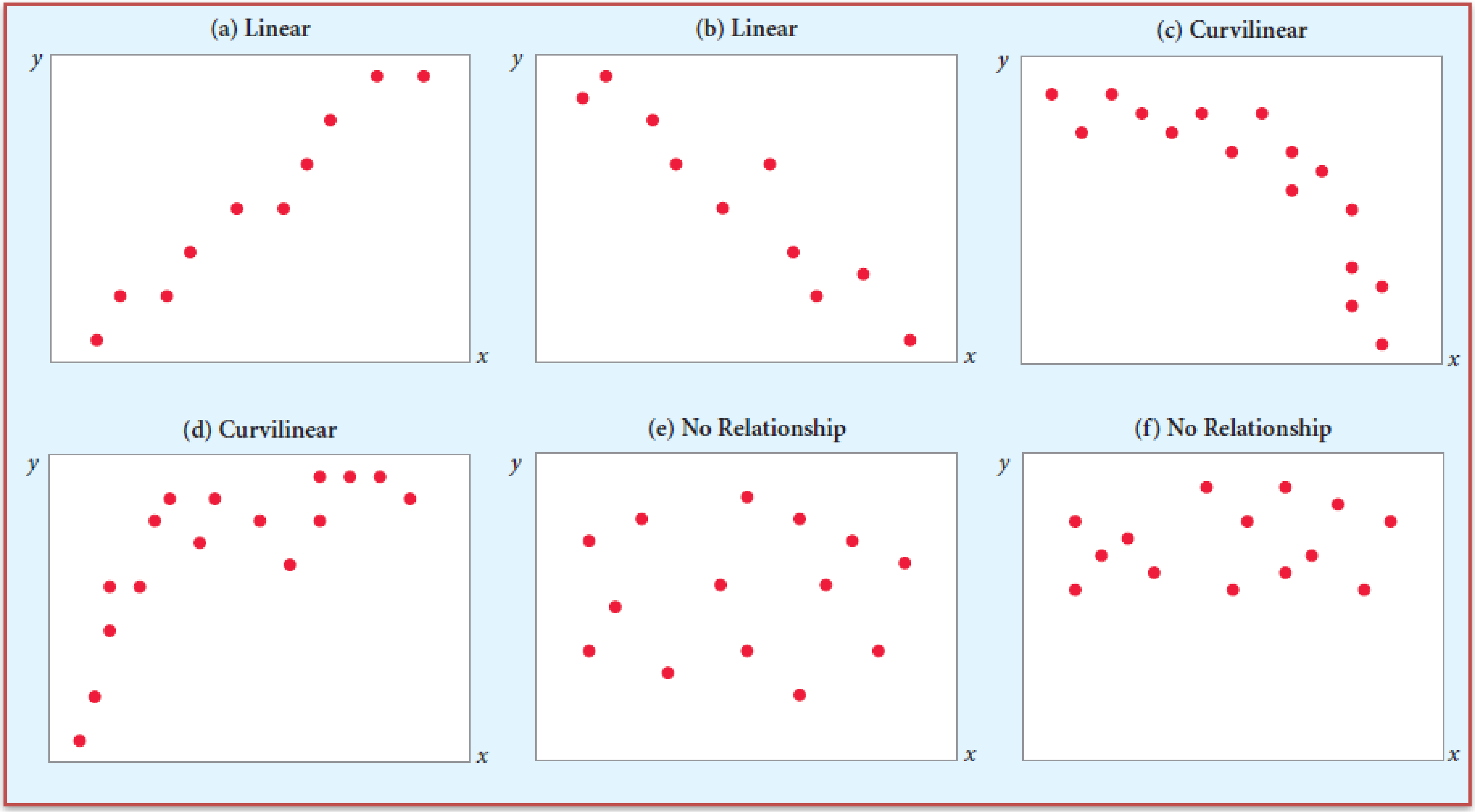
Correlation coefficient

S.L.S.Ly
C215

MEASURES OF ASSOCIATION BETWEEN TWO VARIABLES

- Thus far we have examined numerical methods used to summarize the data for one variable at a time.
- Often a manager or decision maker is interested in the relationship between two variables.
- Two descriptive measures of the relationship between two variables are covariance and correlation coefficient.

2 VARIABLE RELATIONSHIPS



COVARIANCE

- The covariance is a measure of the linear association between two variables.
- Positive values indicate a positive relationship.
- Negative values indicate a negative relationship.

Bowerman, et al. (2017) pp. 161

COVARIANCE

- The covariance is computed as follows:
  - $$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$ for samples
  - $$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$ for populations

CORRELATION COEFFICIENT

- Correlation is a measure of linear association and not necessarily causation.
- Just because two variables are highly correlated, it does not mean that one variable is the cause of the other.

CORRELATION COEFFICIENT

- The correlation coefficient is computed as follows:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

for samples

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

for populations

Pearson Product Moment Correlation Coefficient.

CORRELATION COEFFICIENT

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

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

r - Sample correlation coefficient
n - Sample size
x - Value of the independent variable
y - Value of the dependent variable

CORRELATION COEFFICIENT

- The coefficient can take on values between -1 and +1.
- Values near -1 indicate a <u>strong negative linear relationship</u>.
- Values near +1 indicate a <u>strong positive linear relationship</u>.
- The closer the correlation is to zero, the weaker the relationship.

Covariance and Correlation Coefficient

	x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
	277.6	69	10.65	-1.0	-10.65
	259.5	71	-7.45	1.0	-7.45
	269.1	70	2.15	0	0
	267.0	70	0.05	0	0
	255.6	71	-11.35	1.0	-11.35
	272.9	69	5.95	-1.0	-5.95
Average	267.0	70.0		Total	-35.40
Std. Dev.	8.2192	.8944			

Ads x	Customer y	x-xmean	(x-xmean)²	y-ymean	(y-ymean)²	(x-xmean)(y-ymean)
5.00	353.00	0.88	0.77	-31.81	1011.88	-27.993
6.00	319.00	1.88	3.53	-65.81	4330.96	-123.723
3.00	440.00	-1.12	1.25	55.19	3045.94	-61.813
2.00	332.00	-2.12	4.49	-52.81	2788.90	111.957
4.00	172.00	-0.12	0.01	-212.81	45288.10	25.537
2.00	331.00	-2.12	4.49	-53.81	2895.52	114.077
4.00	344.00	-0.12	0.01	-40.81	1665.46	4.897
2.00	483.00	-2.12	4.49	98.19	9641.28	-208.163
4.00	329.00	-0.12	0.01	-55.81	3114.76	6.697
2.00	532.00	-2.12	4.49	147.19	21664.90	-312.043
7.00	496.00	2.88	8.29	111.19	12363.22	320.227
5.00	393.00	0.88	0.77	8.19	67.08	7.207
4.00	376.00	-0.12	0.01	-8.81	77.62	1.057
7.00	372.00	2.88	8.29	-12.81	164.10	-36.893
2.00	512.00	-2.12	4.49	127.19	16177.30	-269.643
5.00	254.00	0.88	0.77	-130.81	17111.26	-115.113
5.00	459.00	0.88	0.77	74.19	5504.16	65.287
2.00	153.00	-2.12	4.49	-231.81	53735.88	491.437
1.00	426.00	-3.12	9.73	41.19	1696.62	-128.513
6.00	566.00	1.88	3.53	181.19	32829.82	340.637
6.00	596.00	1.88	3.53	211.19	44601.22	397.037
5.00	395.00	0.88	0.77	10.19	103.84	8.967
6.00	676.00	1.88	3.53	291.19	84791.62	547.437
3.00	194.00	-1.12	1.25	-190.81	36408.46	213.707
2.00	135.00	-2.12	4.49	-249.81	62405.04	529.597
7.00	367.00	2.88	8.29	-17.81	317.20	-51.293
4.12	**384.81**		**86.65**		**463802.04**	**1850.577**
			1.86		136.21	74.023
			3.466176		18552.081544	

S.L.S.Ly
C215

Ads x	Customer y	x-xmean	$(x\text{-}xmean)^2$	y-ymean	$(y\text{-}ymean)^2$	(x-xmean) (y-ymean)
5.00	353.00	0.88	0.77	-31.81	1011.88	-27.993
...	....	...	...	...	...	...
6.00	566.00	1.88	3.53	181.19	32829.82	340.637
6.00	596.00	1.88	3.53	211.19	44601.22	397.037
5.00	395.00	0.88	0.77	10.19	103.84	8.967
6.00	676.00	1.88	3.53	291.19	84791.62	547.437
3.00	194.00	-1.12	1.25	-190.81	36408.46	213.707
2.00	135.00	-2.12	4.49	-249.81	62405.04	529.597
7.00	367.00	2.88	8.29	-17.81	317.20	-51.293
4.12	**384.81**		**86.65**		**463802.04**	**1850.577**
			1.86		136.21	74.023

S.L.S.Ly
C215

Ads x	Customer y	x-xmean	(x-xmean)²	y-ymean	(y-ymean)²	(x-xmean)(y-ymean)
5.00	353.00	0.88	0.77	-31.81	1011.88	-27.993
			3.53	-65.81	4330.96	-123.723
			1.25	55.19	3045.94	-61.813
			4.49	-52.81	2788.90	111.957
			0.01	-212.81	45288.10	25.537
			4.49	-53.81	2895.52	114.077
			0.01			4.897
2.00	483.00	-2.12	4.49			-208.163
4.00	329.00	-0.12	0.01			6.697
2.00	532.00	-2.12	4.49			-312.043
7.00	496.00	2.88	8.29			320.227
5.00	393.00	0.88	0.77			7.207
4.00	376.00	-0.12	0.01	-8.81	77.62	1.057
7.00	372.00			12.81		
2.00	512.0			27.19		
5.00	254.0			30.81		
5.00	459.0			74.19		
2.00	153.0			31.81		
1.00	426.0			41.19	1696.62	-128.513
6.00	566.00	1.88	3.53	181.19	32829.82	340.637
6.00	596.00	1.88	3.53	211.19	44601.22	397.037
5.00	395.00	0.88	0.77	10.19	103.84	8.967
6.00	676.00	1.88	3.53	291.19	84791.62	547.437
3.00	194.00	-1.12	1.25	-190.81	36408.46	213.707
2.00	135.00	-2.12	4.49	-249.81	62405.04	529.597
7.00	367.00	2.88	8.29	-17.81	317.20	-51.293
4.12	**384.81**		**86.65**		**463802.04**	**1850.577**
			1.86		136.21	74.023

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$$

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

S.L.S C215

PROBLEM # 10.5

At 5% level significance. The manager of Colonial Furniture has been reviewing weekly advertising expenditures. During the past 6 months, all advertisements for the store have appeared in the local newspaper. The number of ads per week has varied from one to seven. The store's sales staff has been tracking the number of customers who enter the store each week. The number of ads and the number of customers per week for the past 26 weeks were recorded.

d. Find and interpret the coefficient of determination.

e. In your opinion, is it worthwhile exercise to use the regression equation to predict the number of customers who will enter the store, given that Colonial intends to advertise five times in the newspaper? If so, find the 95% prediction interval. If not, explain why not.

S.L.S.Ly
C215

Ads x	Customer y	x-xmean	(x-xmean)^2	y-ymean	(y-ymean)^2	(x-xmean)(y-ymean)
5.00	353.00	0.88	0.77	-31.81	1011.88	-27.993
			...	...	...	...
			3.53	-65.81	4330.96	-123.723
			4.49	-231.81	53735.88	491.437
			9.73	41.19	1696.62	-128.513
			3.53	181.19	32829.82	340.637
			3.53	211.19	44601.22	397.037
			0.77	10.19	103.84	8.967
6.00	676.00	1.88	3.53	291.19	84791.62	547.437
3.00	194.00	-1.12	1.25	-190.81	36408.46	213.707
2.00	135.00	-2.12	4.49	-249.81	62405.04	529.597
7.00	367.00	2.88	8.29	-17.81	317.20	-51.293
4.12	**384.81**		**86.65**		**463802.04**	**1850.577**
			1.86		136.21	74.023

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

$$s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$$

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

PROBLEM # 10.5

At 5% level significance. The manager of Colonial Furniture has been reviewing weekly advertising expenditures. During the past 6 months, all advertisements for the store have appeared in the local newspaper. The number of ads per week has varied from one to seven. The store's sales staff has been tracking the number of customers who enter the store each week. The number of ads and the number of customers per week for the past 26 weeks were recorded.

d. Find and interpret the coefficient of determination.

e. In your opinion, is it worthwhile exercise to use the regression equation to predict the number of customers who will enter the store, given that Colonial intends to advertise five times in the newspaper? If so, find the 95% prediction interval. If not, explain why not.

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{(74.02)^2}{(3.47)(18{,}552)} = .0851$$

There is a weak linear relationship between the number of ads and the number of customers.

S.L.S.Ly C215

PROBLEM # 10.5

At 5% level significance. The manager of Colonial Furniture has been reviewing weekly advertising expenditures. During the past 6 months, all advertisements for the store have appeared in the local newspaper. The number of ads per week has varied from one to seven. The store's sales staff has been tracking the number of customers who enter the store each week. The number of ads and the number of customers per week for the past 26 weeks were recorded.

S.L.S.Ly

d. Find and interpret the coefficient of determination.

e. In your opinion, is it worthwhile exercise to use the regression equation to predict the number of customers who will enter the store, given that Colonial intends to advertise five times in the newspaper? If so, find the 95% prediction interval. If not, explain why not.

C215

The linear relationship is too weak for the model to produce predictions.

Using the Estimated Regression –Equation for Estimation and Prediction

7. ESTIMATION

POINT ESTIMATION

INTERVAL ESTIMATION

CONFIDENCE INTERVAL FOR THE MEAN VALUE OF Y

PREDICTION INTERVAL FOR AN INDIVIDUAL VALUE OF Y

POINT ESTIMATION

If 3 TV ads are run prior to a sale, we expect the mean number of cars sold to be:

$$\hat{y} = 10 + 5(3) = 25 \text{ cars}$$

Using the Estimated Regression Equation for Estimation and Prediction

- Confidence Interval Estimate of $E(y_p)$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$$

- Prediction Interval Estimate of y_p

$$y_p \pm t_{\alpha/2} s_{\text{ind}}$$

where:

confidence coefficient is 1 - α and
$t_{\alpha/2}$ is based on a t distribution
with n - 2 degrees of freedom

Confidence Interval for $E(y_p)$

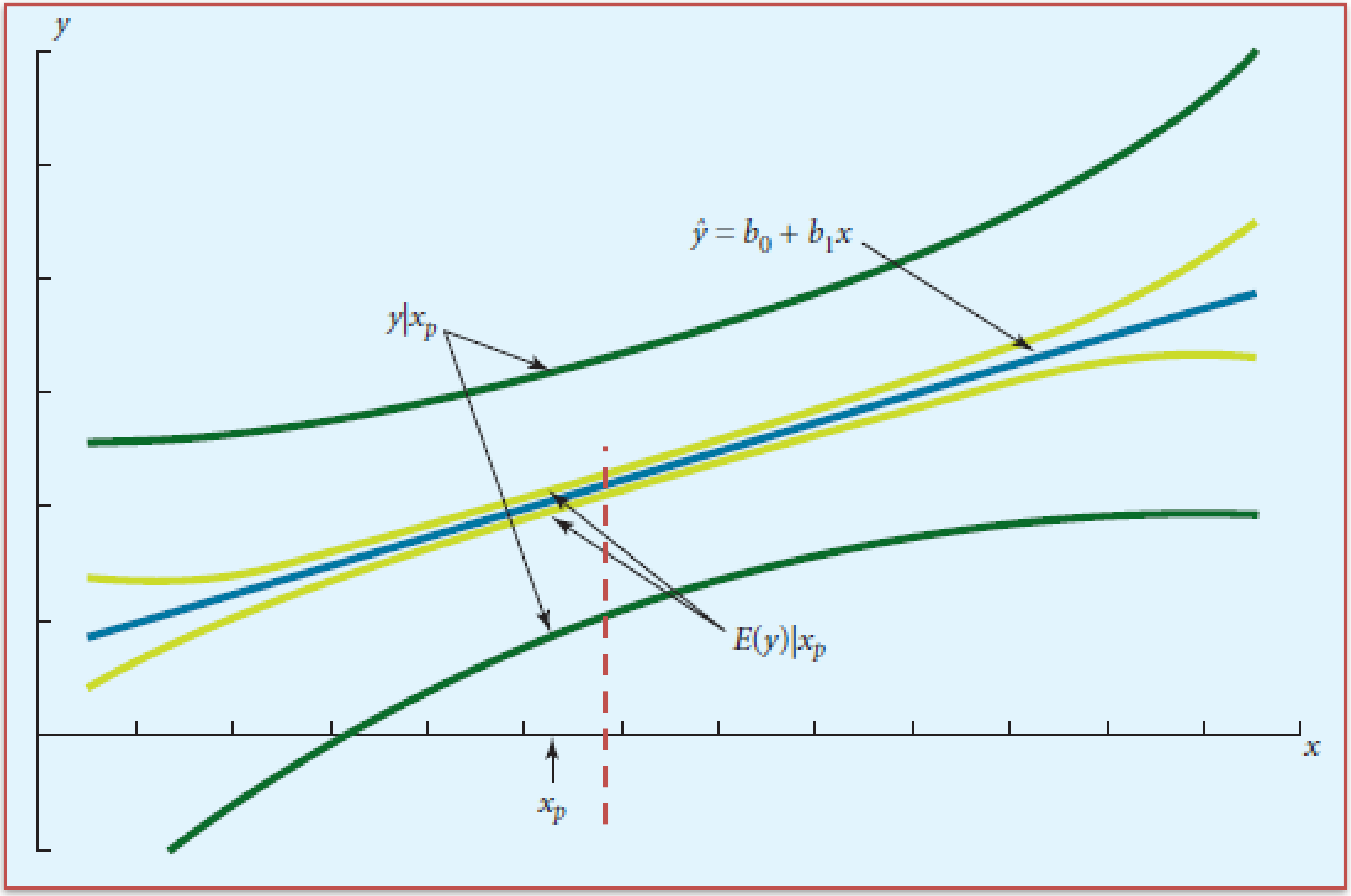
- Estimate of the Standard Deviation of $\hat{y}_p$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Bowerman, et al. (2017) pp. 560

CONFIDENCE VS PREDICTION INTERVAL



COMM215
First the Foundation, then Innovation

PROBLEM #10.6

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}_i)}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sum x_i^2 - (\sum x_i)^2 / n}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

For n=6 data points, the following quantities have been calculated:

$\sum x = 40$ $\quad \sum y = 76$ $\quad \sum xy = 400$

$\sum x^2 = 346$ $\quad \sum y^2 = 1160$ $\quad \sum (y - \hat{y})^2 = 52.334$

a. Determine the least-squares regression line.

S.L.S.Ly
C215

PROBLEM #10.6

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y}_i)}{\sum(x_i - \bar{x})^2} = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sum x_i^2 - (\sum x_i)^2 / n}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

For n=6 data points, the following quantities have been calculated:

$\sum x = 40$ $\sum y = 76$ $\sum xy = 400$

$\sum x^2 = 346$ $\sum y^2 = 1160$ $\sum(y - \hat{y})^2 = 52.334$

S.L.S.Ly
C215

a. Determine the least-squares regression line.

To determine the least squares regression line, we must calculate the slope and y-intercept.

$$b_1 = \frac{400 - [(40)(76)/6]}{346 - [(40)^2/6]} = -1.345$$

$$b_0 = \bar{y} - b_1\bar{x} = 12.67 - (-1.345)(6.67) = 21.641$$

The regression equation is ŷ = 21.641 – 1.345 x

$$s_{y.x} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}}$$

PROBLEM #10.6

For n=6 data points, the following quantities have been calculated:

$\sum x = 40$ $\quad \sum y = 76$ $\quad \sum xy = 400$

$\sum x^2 = 346$ $\quad \sum y^2 = 1160$ $\quad \sum (y - \hat{y})^2 = 52.334$

S.L.S.Ly
C215

b. Determine the standard error of estimate.

PROBLEM #10.6

$$s_{y.x} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

For n=6 data points, the following quantities have been calculated:

$\sum \boldsymbol{x} = \mathbf{40}$ $\sum \boldsymbol{y} = \mathbf{76}$ $\sum \boldsymbol{xy} = \mathbf{400}$

$\sum \boldsymbol{x}^2 = \mathbf{346}$ $\sum \boldsymbol{y}^2 = \mathbf{1160}$ $\sum (\boldsymbol{y} - \hat{\boldsymbol{y}})^2 = \mathbf{52.334}$

S.L.S.Ly
C215

b. Determine the standard error of estimate.

The standard error of the estimate is

$$s_{y.x} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{52.334}{6-2}} = 3.617$$

PROBLEM #10.6

The regression equation is $\hat{y} = 21.641 - 1.345\,(7) = 12.226$

For n=6 data points, the following quantities have been calculated:

$\sum x = 40$ $\quad \sum y = 76$ $\quad \sum xy = 400$

$\sum x^2 = 346$ $\quad \sum y^2 = 1160$ $\quad \sum (y - \hat{y})^2 = 52.334$

c. Construct the 95% confidence interval for the mean of y when x=7.0

S.L.S.Ly C215

$$\hat{y} \pm t s_{y.x} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(\sum x_i^2) - \frac{(\sum x_i)^2}{n}}} = 12.226 \pm 2.776(3.617) \sqrt{\frac{1}{6} + \frac{(7 - 6.67)^2}{(346 - \frac{40^2}{6})}}$$

$$12.226 \pm 10.041 \sqrt{\frac{1}{6} + \frac{0.1089}{79.333}}$$

$$12.226 \pm 10.041 \times 0.4099$$

$$12.226 + 10.041 \times 0.4099 = 16.342$$

$$12.226 - 10.041 \times 0.4099 = 8.11$$

PROBLEM #10.6

The regression equation is $\hat{y} = 21.641 - 1.345\,(7) = 12.226$

For n=6 data points, the following quantities have been calculated:

$\sum x = 40$ $\quad \sum y = 76$ $\quad \sum xy = 400$

$\sum x^2 = 346$ $\quad \sum y^2 = 1160$ $\quad \sum (y - \hat{y})^2 = 52.334$

S.L.S.Ly
C215

c. Construct the 95% confidence interval for the mean of y when x=7.0

We also need the t-value with 6 - 2 = 4 degrees of freedom for a 95% interval; this value is 2.776.

Therefore the 95% confidence interval for the mean value of y when x = 7 is:

$$\hat{y} \pm t s_{y.x} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(\sum x_i^2) - \frac{(\sum x_i)^2}{n}}} = 12.226 \pm 2.776(3.617)\sqrt{\frac{1}{6} + \frac{(7 - 6.67)^2}{(346 - \frac{40^2}{6})}}$$

The confidence interval ranges from 8.110 to 16.342

PROBLEM #10.6

The regression equation is ŷ = 21.641 – 1.345 (9) = 9.536

For n=6 data points, the following quantities have been calculated:

$\sum x = 40$ $\sum y = 76$ $\sum xy = 400$

S.L.S.Ly

$\sum x^2 = 346$ $\sum y^2 = 1160$ $\sum (y - \hat{y})^2 = 52.334$

d. Construct the 95% confidence interval for the mean of y when x=9.0

C215

$$\hat{y} \pm t s_{y.x} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(\sum x_i^2) - \frac{(\sum x_i)^2}{n}}} = 9.536 \pm 2.776(3.617) \sqrt{\frac{1}{6} + \frac{(9 - 6.67)^2}{(346 - \frac{40^2}{6})}}$$

$$9.536 \pm 10.041 \sqrt{\frac{1}{6} + \frac{5.4289}{79.333}}$$

$$9.536 \pm 10.041 \times 0.4849$$

9.536 + 10.041 x 0.4849 = 14.4048

9.536 - 10.041 x 0.4849 = 4.668

PROBLEM #10.6

The regression equation is $\hat{y} = 21.641 - 1.345\ (9) = 9.536$

For n=6 data points, the following quantities have been calculated:

$\sum x = 40$ $\quad \sum y = 76$ $\quad \sum xy = 400$

$\sum x^2 = 346$ $\quad \sum y^2 = 1160$ $\quad \sum (y - \hat{y})^2 = 52.334$

S.L.S.Ly
C215

d. Construct the 95% confidence interval for the mean of y when x=9.0

We also need the t-value with 6 - 2 = 4 degrees of freedom for a 95% interval; this value is 2.776.

Therefore the 95% confidence interval for the mean value of y when x = 9 is:

$$\hat{y} \pm t s_{y.x} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(\sum x_i^2) - \frac{(\sum x_i)^2}{n}}} = 9.536 \pm 2.776(3.617) \sqrt{\frac{1}{6} + \frac{(9 - 6.67)^2}{(346 - \frac{40^2}{6})}}$$

The confidence interval ranges from 4.668 to 14.404

PROBLEM #10.6

For n=6 data points, the following quantities have been calculated:

$\sum x = 40$ $\sum y = 76$ $\sum xy = 400$

$\sum x^2 = 346$ $\sum y^2 = 1160$ $\sum (y - \hat{y})^2 = 52.334$

e. Compare the width of the confidence interval obtained in part (c) with the obtained in part (d). Which is wider and why?

S.L.S.Ly
C215

PROBLEM #10.6

For n=6 data points, the following quantities have been calculated:

$\sum x = 40$ $\quad \sum y = 76$ $\quad \sum xy = 400$

$\sum x^2 = 346$ $\quad \sum y^2 = 1160$ $\quad \sum(y - \hat{y})^2 = 52.334$

e. Compare the width of the confidence interval obtained in part (c) with the obtained in part (d). Which is wider and why?

S.L.S.Ly
C215

The confidence interval in d is wider because 9 is farther from the mean of x than 7.

For x = 7,
8.110 to 16.342

For x = 9,
4.668 to 14.404

Prediction Interval for y_p

- Estimate of the Standard Deviation of an Individual Value of y_p

$$\hat{y}_p \pm t_{\alpha/2} s_{\text{ind}}$$

$$s_{\text{ind}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Bowerman, et al. (2017) pp. 560

PROBLEM #10.7

The regression equation is $\hat{y} = 21.641 – 1.345\ (2) = 18.951$

For the summary data provided in Problem #10.18, construct a 95% prediction interval for an individual y value whenever

a. x=2

S.L.S.Ly
C215

PROBLEM #10.7

The regression equation is $\hat{y} = 21.641 – 1.345 (2) = 18.951$

For the summary data provided in Problem #10.18, construct a 95% prediction interval for an individual y value whenever

a. x=2

We also need the t-value with 6 - 2 = 4 degrees of freedom for a 95% interval; this value is 2.776. Therefore, the 95% prediction interval for an individual y value when x = 2 is:

S.L.S.Ly
C215

$$\hat{y} \pm t s_{y.x}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{(\sum x_i^2)-\frac{(\sum x_i)^2}{n}}} = 18.951 \pm 2.776(3.617)\sqrt{1+\frac{1}{6}+\frac{(2-6.67)^2}{(346-\frac{40^2}{6})}}$$

$$18.951 \pm 10.041\sqrt{1+\frac{1}{6}+\frac{21.8089}{79.333}}$$

$$18.951 \pm 10.041 \times 1.2006$$

18.951 + 10.041 x 1.2006 = 31.006

18.951 - 10.041 x 1.2006 = 6.896

PROBLEM #10.7

The regression equation is ŷ = 21.641 – 1.345 (2) = 18.951

For the summary data provided in Problem #10.18, construct a 95% prediction interval for an individual y value whenever

a. x=2

S.L.S.Ly
C215

We also need the t-value with 6 - 2 = 4 degrees of freedom for a 95% interval; this value is 2.776. Therefore, the 95% prediction interval for an individual y value when x = 2 is:

$$\hat{y} \pm t s_{y.x} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\left(\sum x_i^2\right) - \frac{\left(\sum x_i\right)^2}{n}}} = 18.951 \pm 2.776(3.617) \sqrt{1 + \frac{1}{6} + \frac{(2 - 6.67)^2}{\left(346 - \frac{40^2}{6}\right)}}$$

The prediction interval ranges from 6.895 to 31.007