Modeling a Linear Relationship

Lecture 46
Section 13.3.2

Robb T. Koether

Hampden-Sydney College

Tue, Apr 15, 2008

Outline

Introduction

- We are now ready to calculate the least-squares regression line.
- The formulas are a bit daunting, but the TI-83 will do the heavy lifting for us.
- Once we find the regression line, we are then in a position to use it to make predictions.

The Least Squares Regression Line

- The equation of the regression line is of the form

$$\hat{y} = a + bx.$$

- b is the slope of the regression line.
- a is the y-intercept.
- We need to find the coefficients a and b from the data.

The Least Squares Regression Line

- The formula for b is

$$b = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sum (x - \overline{x})^2}$$

or

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - \left(\sum x\right)^2}.$$

- The formula for a is

$$a = \overline{y} - b\overline{x}.$$

The Least Squares Regression Line

- It so happens that the regression line passes through the center $(\overline{x}, \overline{y})$ of the scatterplot.
- That means that the point $(\overline{x}, \overline{y})$ satisfies the equation $\hat{y} = a + bx$.
- That is,

$$\overline{y} = a + b\overline{x}.$$

- Thus,

$$a = \overline{y} - b\overline{x}.$$

The First Formula

- Consider again the data set

x	y	$x - \overline{x}$	$y - \overline{y}$
1	8		
3	12		
4	9		
5	14		
8	16		
9	20		
11	17		
15	24		

The First Formula

- Compute the x and y deviations.

x	y	$x - \overline{x}$	$y - \overline{y}$
1	8	-6	-7
3	12	-4	-3
4	9	-3	-6
5	14	-2	-1
8	16	1	1
9	20	2	5
11	17	4	2
15	24	8	9

The First Formula

- Compute the squared deviations.

$x - \overline{x}$	$y - \overline{y}$	$(x - \overline{x})^2$	$(y - \overline{y})^2$	$(x - \overline{x})(y - \overline{y})$
-6	-7			
-4	-3			
-3	-6			
-2	-1			
1	1			
2	5			
4	2			
8	9			

The First Formula

- Compute the squared deviations.

$x - \overline{x}$	$y - \overline{y}$	$(x - \overline{x})^2$	$(y - \overline{y})^2$	$(x - \overline{x})(y - \overline{y})$
-6	-7	36	49	42
-4	-3	16	9	12
-3	-6	9	36	18
-2	-1	4	1	2
1	1	1	1	1
2	5	4	25	10
4	2	16	4	8
8	9	64	81	72

The First Formula

- Find the sums.

$x - \overline{x}$	$y - \overline{y}$	$(x - \overline{x})^2$	$(y - \overline{y})^2$	$(x - \overline{x})(y - \overline{y})$
-6	-7	36	49	42
-4	-3	16	9	12
-3	-6	9	36	18
-2	-1	4	1	2
1	1	1	1	1
2	5	4	25	10
4	2	16	4	8
8	9	64	81	72
0	0	150	206	165

The First Formula

- Compute the coefficients from the first formula.
- Compute b:

$$b = \frac{165}{150} = 1.1.$$

- Then compute a:

$$a = 15 - (1.1)(7) = 7.3.$$

- The equation is

$$\hat{y} = 7.3 + 1.1x.$$

The First Formula

- Compute the coefficients from the first formula.
- Compute b:

$$b = \frac{165}{150} = 1.1.$$

- Then compute a:

$$a = 15 - (1.1)(7) = 7.3.$$

- The equation is

$$\hat{y} = 7.3 + 1.1x.$$

The First Formula

- Compute the coefficients from the first formula.
- Compute b:

$$b = \frac{165}{150} = 1.1.$$

- Then compute a:

$$a = 15 - (1.1)(7) = 7.3.$$

- The equation is

$$\hat{y} = 7.3 + 1.1x.$$

The First Formula

- Compute the coefficients from the first formula.
- Compute b:

$$b = \frac{165}{150} = 1.1.$$

- Then compute a:

$$a = 15 - (1.1)(7) = 7.3.$$

- The equation is

$$\hat{y} = 7.3 + 1.1x.$$

The Second Formula

- Consider yet again the data set

x	y	x^2	y^2	xy
1	8			
3	12			
4	9			
5	14			
8	16			
9	20			
11	17			
15	24			

The Second Formula

- Square x and y and find xy.

x	y	x^2	y^2	xy
1	8	1	64	8
3	12	9	144	36
4	9	16	81	36
5	14	25	196	70
8	16	64	256	128
9	20	81	400	180
11	17	121	289	187
15	24	225	576	360

The Second Formula

- Add up the columns.

x	y	x^2	y^2	xy
1	8	1	64	8
3	12	9	144	36
4	9	16	81	36
5	14	25	196	70
8	16	64	256	128
9	20	81	400	180
11	17	121	289	187
15	24	225	576	360
56	120	542	2006	1005

The Second Formula

- Compute the coefficients from the second formula.
- Compute b:

$$b = \frac{(8)(1005) - (56)(120)}{(8)(542) - (56)^2} = \frac{1320}{1200} = 1.1.$$

- Then compute a as before:

$$a = 15 - (1.1)(7) = 7.3.$$

- The equation is

$$\hat{y} = 7.3 + 1.1x.$$

The Second Formula

- Compute the coefficients from the second formula.
- Compute b:

$$b = \frac{(8)(1005) - (56)(120)}{(8)(542) - (56)^2} = \frac{1320}{1200} = 1.1.$$

- Then compute a as before:

$$a = 15 - (1.1)(7) = 7.3.$$

- The equation is

$$\hat{y} = 7.3 + 1.1x.$$

The Second Formula

- Compute the coefficients from the second formula.
- Compute b:

$$b = \frac{(8)(1005) - (56)(120)}{(8)(542) - (56)^2} = \frac{1320}{1200} = 1.1.$$

- Then compute a as before:

$$a = 15 - (1.1)(7) = 7.3.$$

- The equation is

$$\hat{y} = 7.3 + 1.1x.$$

The Second Formula

- Compute the coefficients from the second formula.
- Compute *b*:

$$b = \frac{(8)(1005) - (56)(120)}{(8)(542) - (56)^2} = \frac{1320}{1200} = 1.1.$$

- Then compute *a* as before:

$$a = 15 - (1.1)(7) = 7.3.$$

- The equation is

$$\hat{y} = 7.3 + 1.1x.$$

Example

- The second method is usually easier if you are doing it by hand.
- By either method, we get the equation

$$\hat{y} = 7.3 + 1.1x.$$

The Regression Line on the TI-83

- On the TI-83, we could use `2-Var Stats` to get the basic summations.
  - Enter `2-Var Stats` $\mathtt{L}_1, \mathtt{L}_2$.
  - Press `ENTER`.
- The calculator reports that
  - $n = 8$
  - $\Sigma x = 56$
  - $\Sigma x^2 = 542$
  - $\Sigma y = 120$
  - $\Sigma y^2 = 2006$
  - $\Sigma xy = 1005$
- Then use the formulas.

The Regression Line on the TI-83

- Or we can use the `LinReg` function.
  - Put the *x* values in L_1.
  - Put the *y* values in L_2.
  - Select `STAT > CALC > LinReg(a+bx)` (item #8).
  - Press `Enter`. `LinReg(a+bx)` appears in the display.
  - Enter L_1, L_2.
  - Press `Enter`.

The Regression Line on the TI-83

- The following appear in the display.
  - The title `LinReg`.
  - The equation `y=a+bx`.
  - The value of a.
  - The value of b.
  - The value of r^2 (to be discussed later).
  - The value of r (to be discussed later).

The Regression Line on the TI-83

- To graph the regression line along with the scatterplot, after selecting `LinReg(a+bx)`,
  - Enter L_1, L_2, Y_1. (That is, add Y_1.)
  - Press `Enter`.
  - Press `ZOOM > ZoomStat` to draw the graph.
- To see the regression equation,
  - Press `Y=`.

Free Lunch Participation vs. Graduation Rate

- Find the equation of the regression line for the school-district data on the free-lunch participation rate vs. the graduation rate.
- Let x be the free-lunch participation.
- Let y be the graduation rate.

Free Lunches vs. Graduation Rates

District	Free Lunch	Grad. Rate
Amelia	41.2	68.9
Caroline	40.2	62.9
Charles City	45.8	67.7
Chesterfield	22.5	80.5
Colonial Hgts	25.7	73.0
Cumberland	55.3	63.9
Dinwiddie	45.2	71.4
Goochland	23.3	76.3
Hanover	13.7	90.1
Henrico	30.2	81.1
Hopewell	63.1	63.4

Free Lunches vs. Graduation Rates

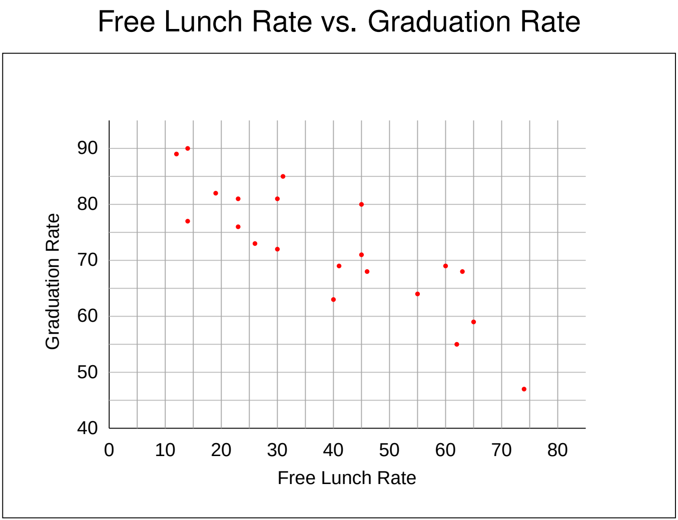
District	Free Lunch	Grad. Rate
King and Queen	59.9	64.1
King William	27.9	67.0
Louisa	44.9	80.1
New Kent	13.9	77.0
Petersburg	61.6	54.6
Powhatan	12.2	89.3
Prince George	30.9	85.0
Richmond	74.0	46.9
Sussex	74.8	59.0
West Point	19.1	82.0

Free Lunch Participation vs. Graduation Rate

- The regression equation is

$$\hat{y} = 91.047 - 0.494x.$$

Scatter Plot



Scatter Plot with Regression Line

Free Lunch Rate vs. Graduation Rate

Predicting y

- What graduation rate would we predict in a district if we knew that the free-lunch participation rate was 50%?
- Calculate

$$\hat{y}(50) = 91.047 - 0.494(50) = 66.347.$$

- We predict a graduation rate of 66.3%.

Predicting y

- What graduation rate would we predict in a district if we knew that the free-lunch participation rate was 50%?
- Calculate

$$\hat{y}(50) = 91.047 - 0.494(50) = 66.347.$$

- We predict a graduation rate of 66.3%.

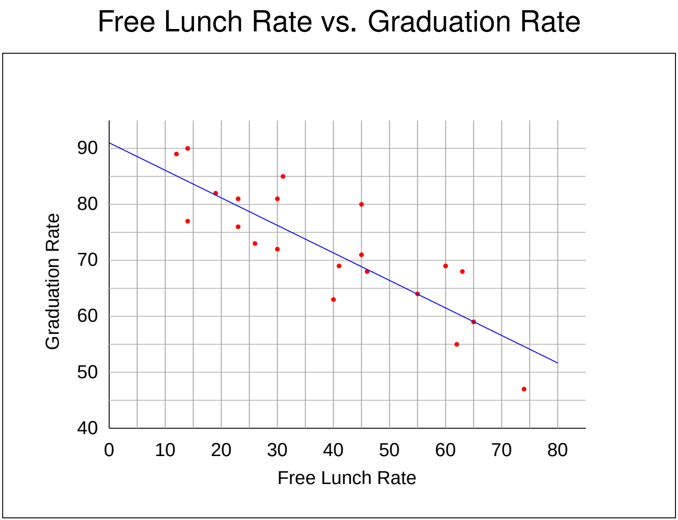
Predicting y

- What graduation rate would we predict in a district if we knew that the free-lunch participation rate was 50%?
- Calculate

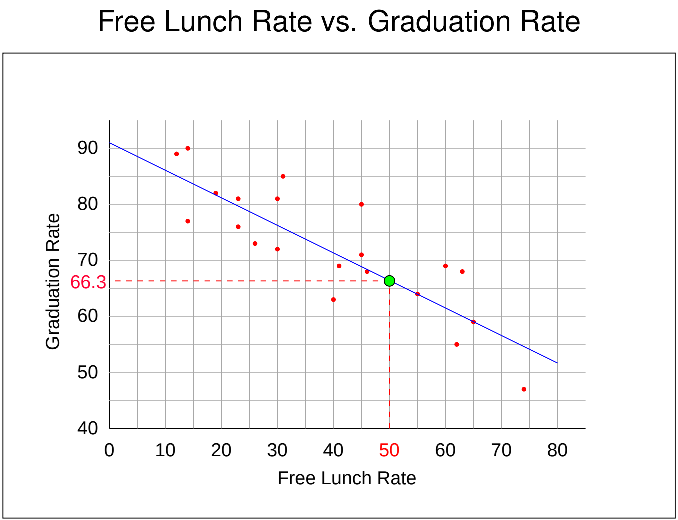
$$\hat{y}(50) = 91.047 - 0.494(50) = 66.347.$$

- We predict a graduation rate of 66.3%.

Predicting y



Predicting y



The Slope of the Regression Line

- The first formula for the slope *b* of the regression line is

$$b = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sum (x - \overline{x})^2}.$$

- Let us consider the numerator

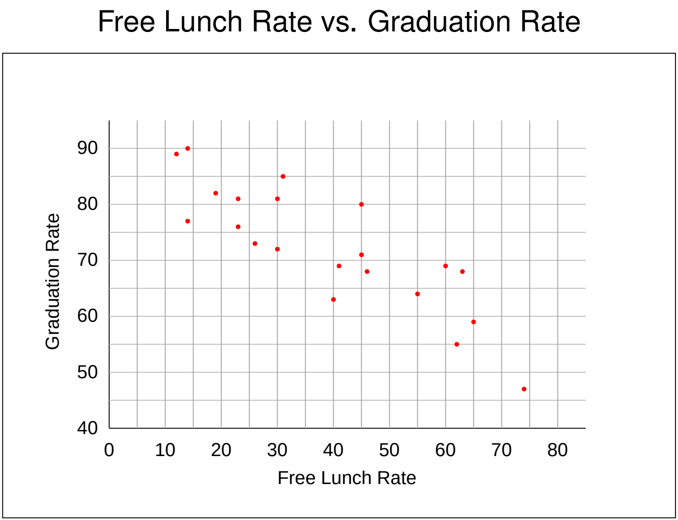
$$\sum (x - \overline{x})(y - \overline{y}).$$

- We will see why it is positive when the trend is upwards, negative when the trend is downwards, and zero when there is no trend.

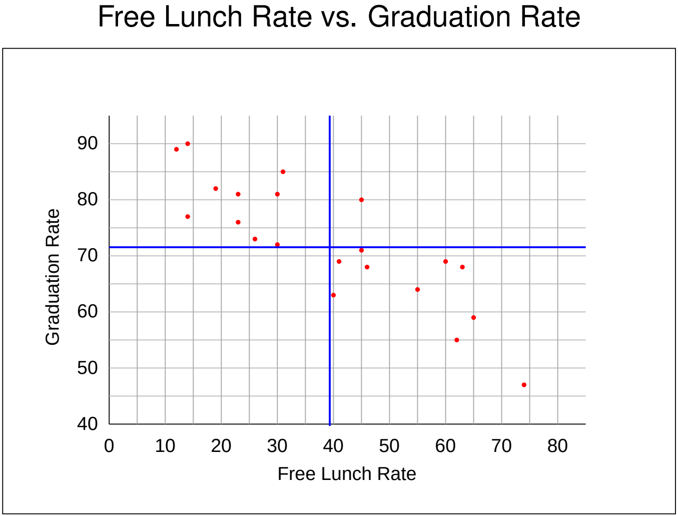
The Slope of the Regression Line

- Consider the free-lunch (x) vs. graduation rate (y) data.
- The average of x is 39.3.
- The average of y is 71.6.
- Use these values to divide the scatterplot into four quadrants.

Scatter Plot



Scatter Plot



The Slope of the Regression Line

- These two lines meet in the center of the scatterplot.
- For every point to the right of the center,

$$x - \overline{x} > 0$$

and for every point to the left,

$$x - \overline{x} < 0.$$

The Slope of the Regression Line

- Similarly, for every point above of the center,

$$y - \overline{y} > 0$$

and for every point below the center,

$$y - \overline{y} < 0.$$

The Slope of the Regression Line

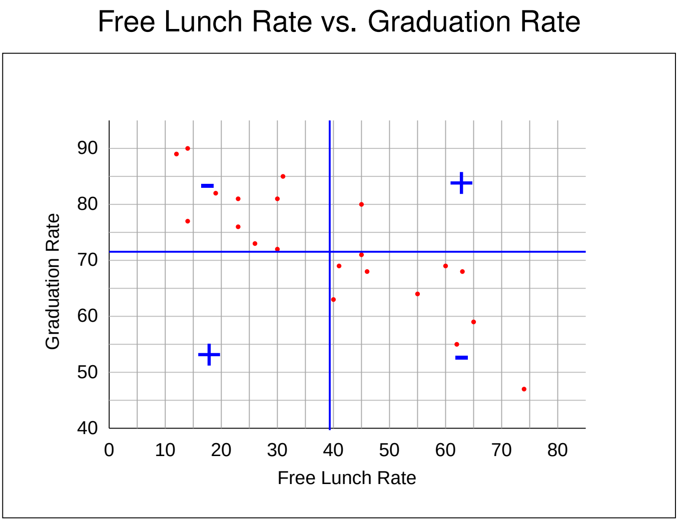
- Therefore, if a point is in the upper-right quadrant or the lower-left quadrant, then

$$(x - \overline{x})(y - \overline{y}) > 0.$$

- If a point is in the upper-left quadrant or the lower-right quadrant, then

$$(x - \overline{x})(y - \overline{y}) < 0.$$

Scatter Plot



The Slope of the Regression Line

- In this example, it is clear that the negative values dominate.
- Therefore,

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

and the regression line has negative slope.

The Slope of the Regression Line

- Had the positive values dominated, then

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

and slope would be negative.

- Had the positive and negative values balanced, then

$$\sum (x - \overline{x})(y - \overline{y}) = 0$$

and slope would be zero.

Variation in the Model

- There is a very simple relationship between the variation in the observed y values and the variation in the predicted y values.

Observed y and Predicted y

- SST $=$ Variation in the Observed y
- SSR $=$ Variation in the Predicted y

Variation in Observed y

- The variation in the observed y is measured by SST (same as SSY).
- For graduation rate data ($\mathtt{L}_2$),

$$\text{SST} = 2598.18.$$

Variation in Predicted y

- The variation in the predicted y is measured by SSR.
- For predicted graduation rate data, let `L3 = Y1(L1)`.
- $\text{SSR} = 1896.67$.
- $\text{SSE} =$ Residual Sum of Squares
- It turns out that

$$\text{SST} = \text{SSE} + \text{SSR}.$$

- That is,

Sum Squared Error

- In the example,

$$\text{SST} - \text{SSR} = 2598.18 - 1896.67 = 701.51.$$

- If we compute the sum of the squared residuals directly, we get

$$\text{SSE} = 701.52.$$

Explaining the Variability

- In the equation

$$SST = SSE + SSR,$$

- SSR is the amount of variability in y that is explained by the model.
- SSE is the amount of variability in y that is not explained by the model.

Explaining the Variability

- In the last example, how much variability in graduation rate is explained by the model (by free-lunch participation)?

Summary

Summary

- The formulas for the regression line $\hat{y} = a + bx$ are

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

and

$$a = \overline{y} - b\overline{x}.$$

- The TI-83 function `LinReg(a+bx)` will do the calculations.
- If we store the regression equation in $\mathtt{Y}_1$, then we can also draw it on the TI-83 and use the TI-83 to make predictions.