Chapter 4.2

Linear Regression and the Coefficient of Determination

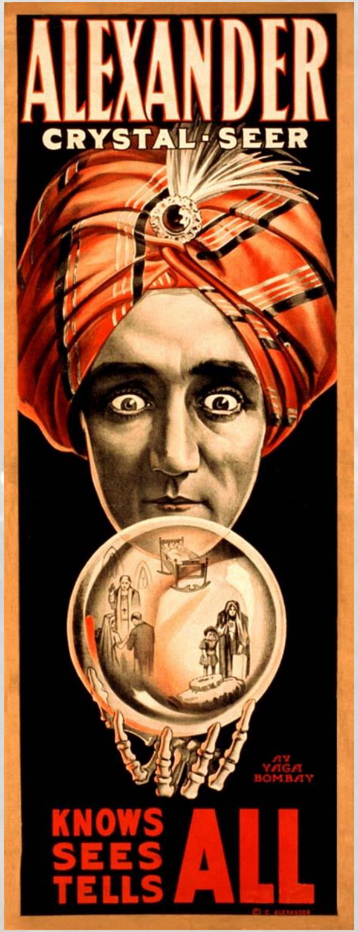
Learning Objectives

At the end of this lecture, the student should be able to:

- Explain what the "least-squares line" is
- Identify and describe the components of the least-squares line equation
- Explain how to calculate the residuals
- Calculate and interpret the coefficient of determination (CD)

Introduction

- Least-squares line
- Least-squares line equation
- Dealing with prediction using the least-squares line
- Coefficient of Determination



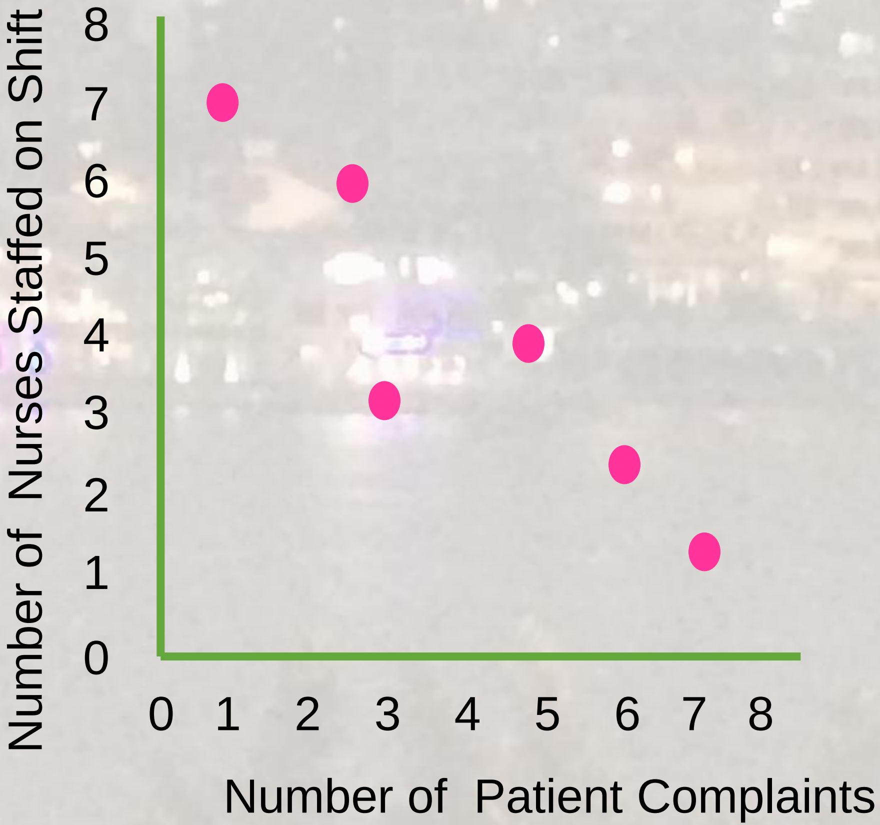
Painting in public domain

Least-Squares Criterion

What this means

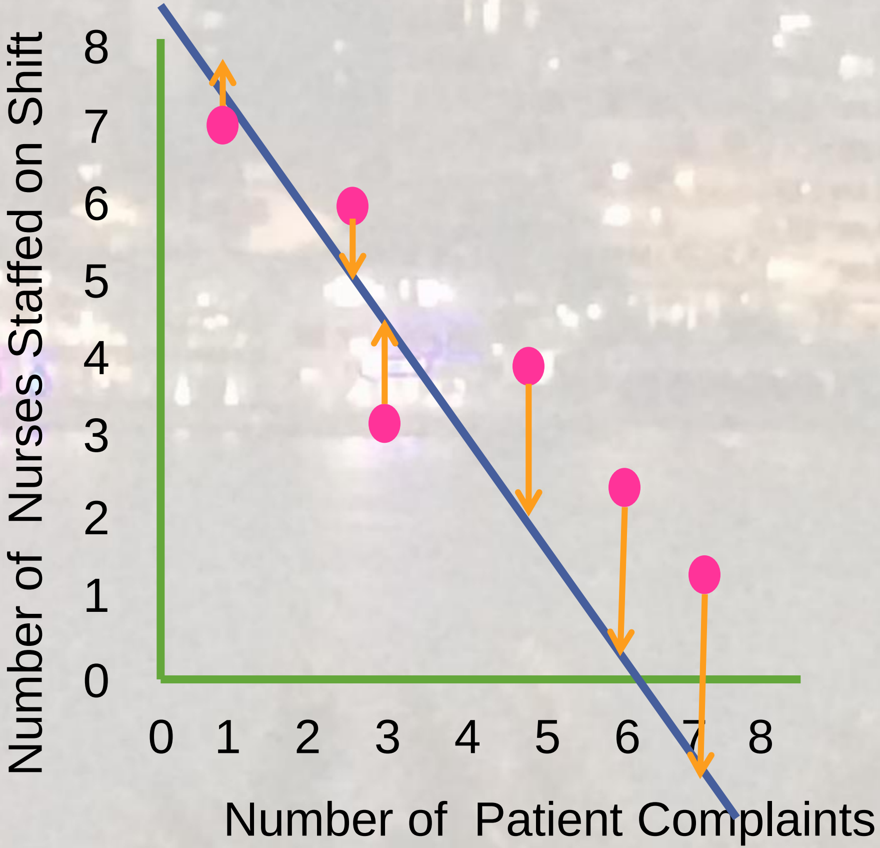
Where Does the Line Go?

- In the last chapter, we plotted scattergrams.
- I just drew a line for demonstration – but there is an official rule as to where this line goes.
- The rule is that the line has to meet the "least squares criterion"



Where Does the Line Go?

- “least-squares line”
  - The vertical distances between the dot and line are squared to get rid of negative sign
  - These are called “squares”
- The line belongs where it would cause the smallest sum of squares for the whole dataset.



Where Does the Line Go?

- If you figure out where the line goes, you can draw it on a scatterplot. But how do you know exactly where it belongs on the graph?
- And what if you don't have a visual? How do you describe the line?
- You use an equation!

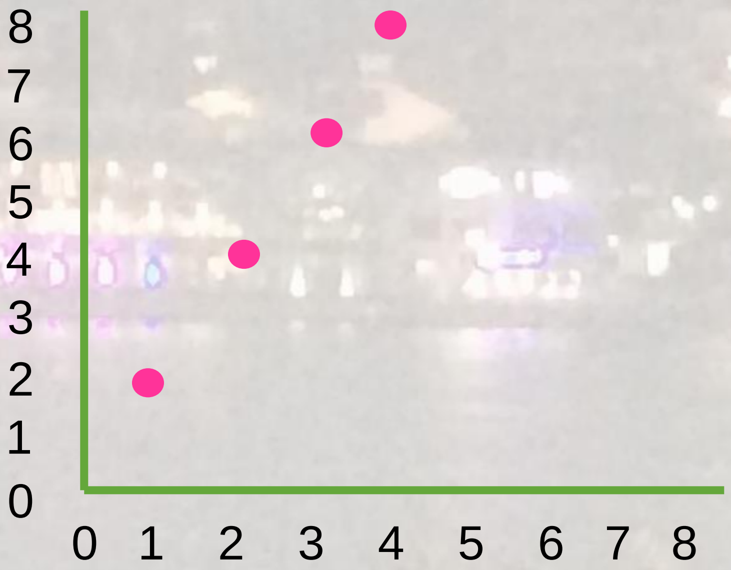
Picture courtesy of Tulane Public Relations

Least-Squares Line Equation

How to Find this Equation

Remember Algebra?

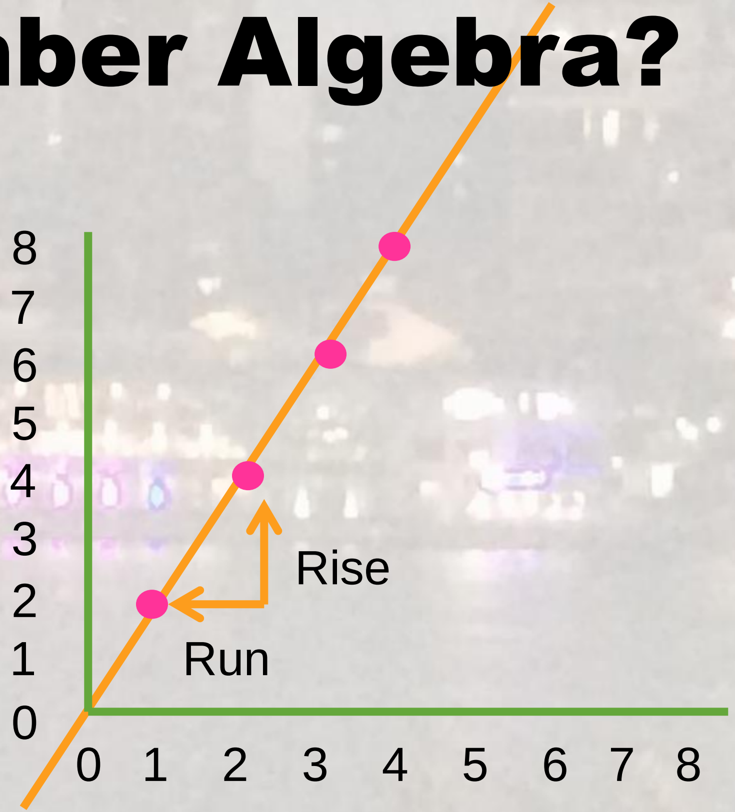
x	y
1	2
2	4
3	6
4	8



Remember Algebra?

x	y
1	2
2	4
3	6
4	8

y = bx + a

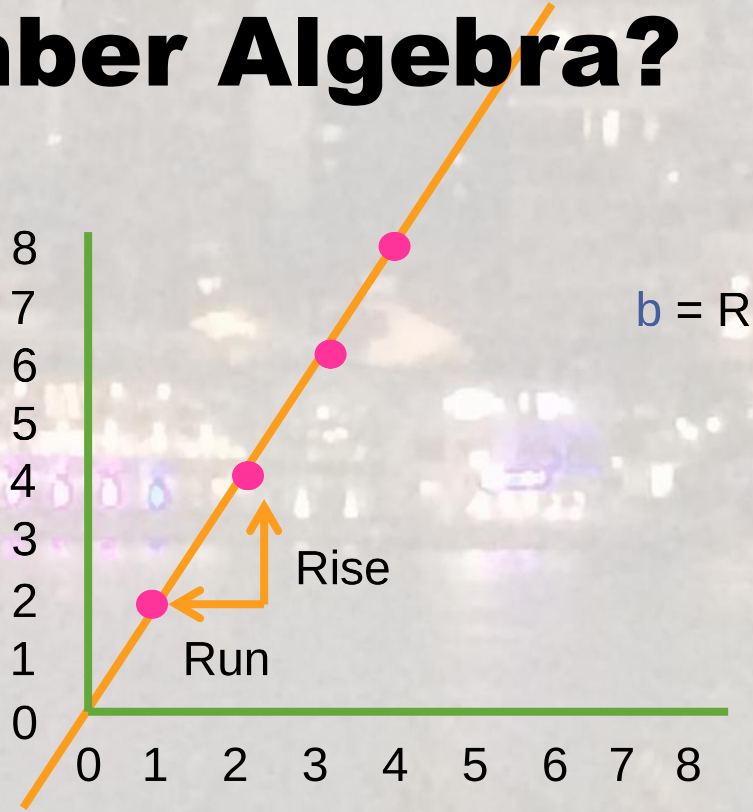


Remember Algebra?

x	y
1	2
2	4
3	6
4	8

y = bx + a

Slope



b = Rise/Run

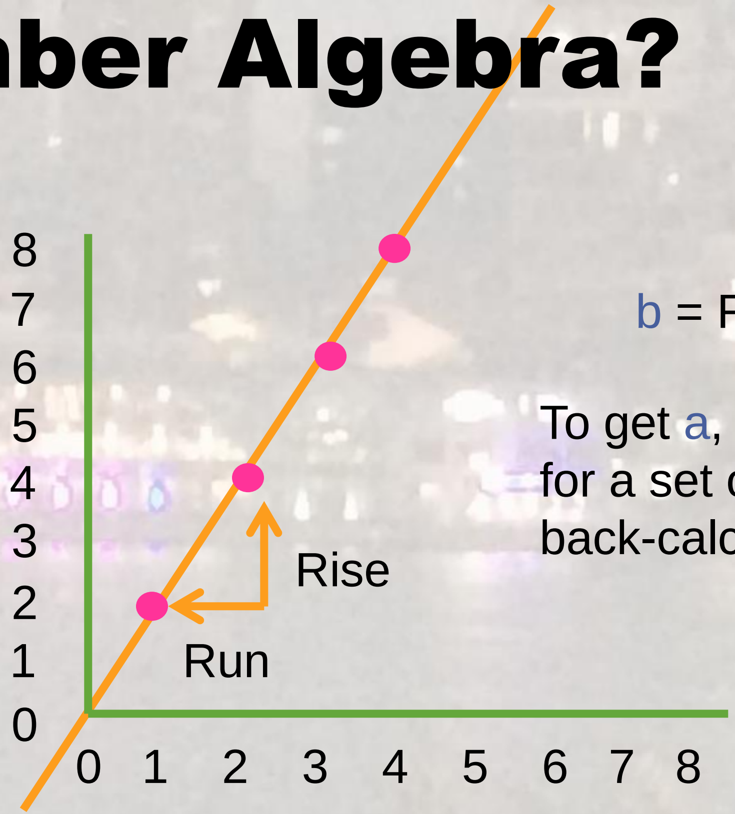
Remember Algebra?

x	y
1	2
2	4
3	6
4	8

$y = bx + a$

Slope

y-intercept

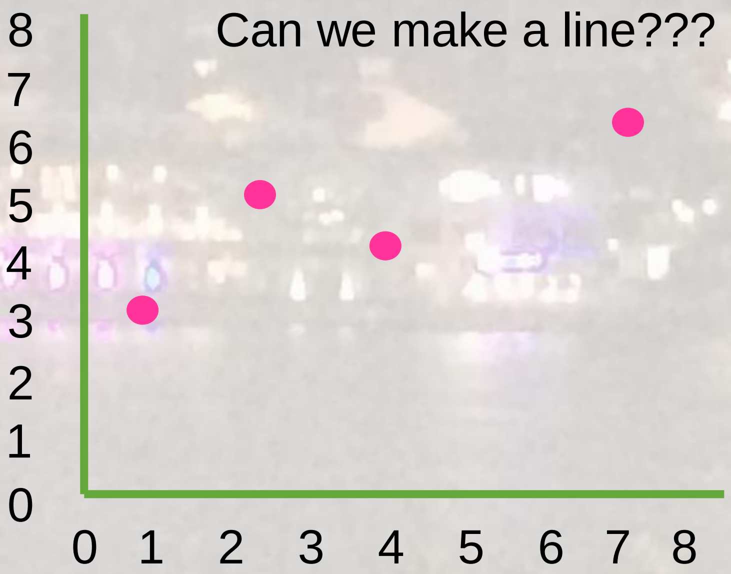


b = Rise/Run

To get a, fill in x and y for a set of x,y pairs and back-calculate.

Okay, now Statistics!

x	y
1	3
3	5
4	4
7	6



Okay, now Statistics!

x	y
1	3
3	5
4	4
7	6

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

Hat (estimate)

Slope

y-intercept

Where Does the Line Go?

Software Approach

- Feed all the x,y pairs you have into the software.
- The software prints out the results in the form of an equation.
  - The "b" – slope
  - And the "a" – the y-intercept

Manual Approach (This Class!)

- Plug all the x,y pairs into an equation to get "b".
- Calculate x-bar and y-bar.
- Plug b, x-bar (for x), and y-bar (for y-hat) into the linear equation to back-calculate a.

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

Hat (estimate) → $\hat{y}$; Slope → b; y-intercept → a

Recycling!

- Least-squares line is usually done along with r
- *SAVE YOUR CALCULATIONS* from r to recycle when calculating b:
  - Σx, Σy, $Σx^2$, and Σxy
- Also – save your r! You will need it later for the Coefficient of Determination.
- NOTE: You will need to calculate x-bar and y-bar – this was not done in r

Photograph by Patrick Nylin

x=DBP, y=# of Appointments

#	x	y	x^2	y^2	xy
1	70	3	4,900	9	210
2	115	45	13,225	2,025	5,175
3	105	21	11,025	441	2,205
4	82	7	6,724	49	574
5	93	16	8,649	256	1,488
6	125	62	15,625	3,844	7,750
7	88	12	7,744	144	1,056
	Σx = 678	Σy = 166	Σx^2 = 67,892	Σy^2 = 6,768	Σxy = 18,458

NOTE: Formula I'm using for b

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

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

GOAL: Fill in b and a so you have the least-squares line equation.

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

x=DBP, y=# of Appointments

#	x	y	x^2	y^2	xy
1	70	3	4,900	9	210
2	115	45	13,225	2,025	5,175
3	105	21	11,025	441	2,205
4	82	7	6,724	49	574
5	93	16	8,649	256	1,488
6	125	62	15,625	3,844	7,750
7	88	12	7,744	144	1,056
	Σx = 678	Σy = 166	Σx^2 = 67,892	Σy^2 = 6,768	Σxy = 18,458

NEW! x-bar = 678/7 = 96.9 y-bar = 166/7 = 23.7

NOTE: Formula I'm using for b

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

$$a = \bar{y} - b\bar{x}$$

GOAL: Fill in b and a so you have the least-squares line equation.

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

x=DBP, y=# of Appointments

$n = 7$

$\Sigma xy = 18{,}458$

$\Sigma x = 678$

$\bar{x} = 96.9$

$\Sigma y = 166$

$\Sigma x^2 = 67{,}892$

$\Sigma y^2 = 6{,}768$

$\bar{y} = 23.7$

NOTE: Formula I'm using for b

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

$$a = \bar{y} - b\bar{x}$$

GOAL: Fill in b and a so you have the least-squares line equation.

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

x=DBP, y=# of Appointments

$n = 7$

$\Sigma xy = 18{,}458$

$\Sigma x = 678$

$\bar{x} = 96.9$

$\Sigma y = 166$

$\Sigma x^2 = 67{,}892$

$\Sigma y^2 = 6{,}768$

$\bar{y} = 23.7$

$$b = \frac{(7)(18{,}458) - (678)(166)}{(7)(67{,}892) - (678)^2}$$

$$b = \frac{16{,}658}{15.560} = 1.1$$

NOTE: Formula I'm using for b

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

$$a = \bar{y} - b\bar{x}$$

GOAL: Fill in b and a so you have the least-squares line equation.

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

x=DBP, y=# of Appointments

$n = 7$

$\Sigma xy = 18{,}458$

$\Sigma x = 678$

$\bar{x} = 96.9$

$\Sigma y = 166$

$\Sigma x^2 = 67{,}892$

$\Sigma y^2 = 6{,}768$

$\bar{y} = 23.7$

$b = 1.1$

$a = \bar{y} - b\bar{x}$

$a = 23.7 - (1.1 * 96.9)$

$a = -80.0$

NOTE: Formula I'm using for b

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

$$a = \bar{y} - b\bar{x}$$

GOAL: Fill in b and a so you have the least-squares line equation.

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

x=DBP, y=# of Appointments

$n = 7$

$\Sigma xy = 18{,}458$

$\Sigma x = 678$

$\bar{x} = 96.9$

$\Sigma y = 166$

$\Sigma x^2 = 67{,}892$

$\Sigma y^2 = 6{,}768$

$\bar{y} = 23.7$

$b = 1.1$

$a = \bar{y} - b\bar{x}$

$a = 23.7 - (1.1 * 96.9)$

$a = -80.0$

CHECK! (1.1*96.9) – 80.0 should = 23.7!

NOTE: Formula I'm using for b

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

$$a = \bar{y} - b\bar{x}$$

GOAL: Fill in b and a so you have the least-squares line equation.

$\hat{y} = 1.1x - 80.0$

Predicting with the Least Squares Line Equation

Different Ways to Use the Equation

Rule About Least Squares Line

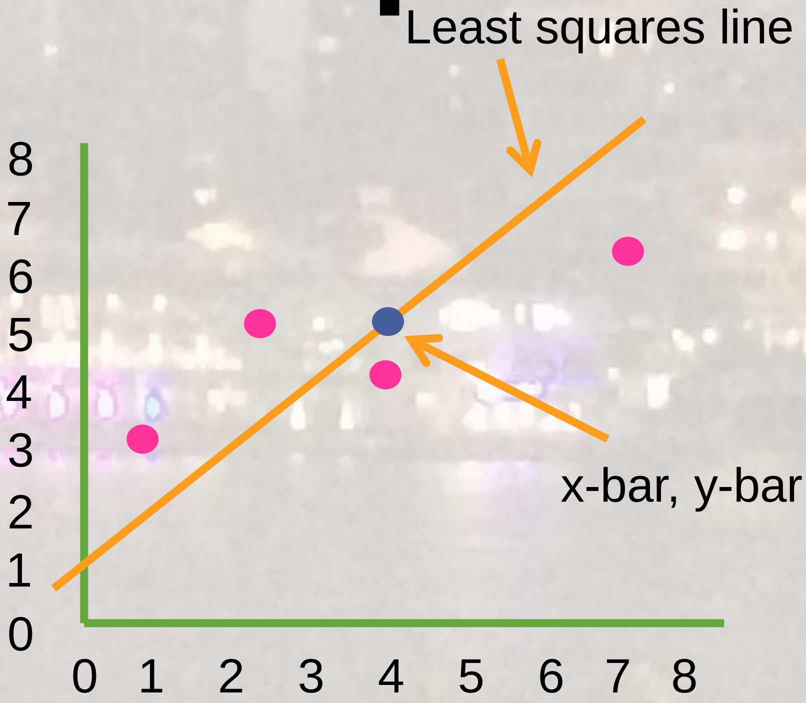
x	y
1	3
3	5
4	4
7	6

$\hat{y} = bx + a$

Hat (estimate)

Slope

y-intercept



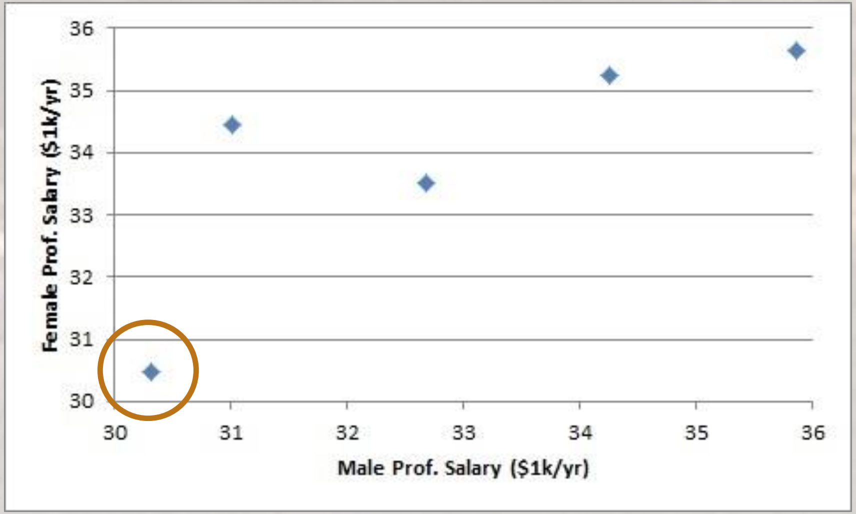
x-bar and y-bar *always* fall on the least squares line – but other points may or may not

Facts About the Slope (b)

- The slope (b) of the least-squares line tells us how many units the response variable (y) is expected to change for each 1 unit of change in the explanatory variable (x).
- For our example: $\hat{y} = 1.1x - 80.0$
  - x=DBP, y=# of Appointments
  - For each increase in 1 mmHg of DBP (x), there is a 1.1 increase in the number of appointments the patient had over the past year (y)
- The number of units change in the y for each unit change in x is called the “marginal change” in the y.

Influential Points

- Like with r, if a point is an outlier, it can drastically influence the least squares line equation.
- An extremely high x or extremely low x can do that.
- Always check the scattergram first for outliers!



What is the "Residual"?

- Once the equation is there, you can plug each x in, and get a y-hat out.
- Patient #1:
  - (1.1*70) – 80.0 = -3
- Patient #2:
  - (1.1*115) – 80.0 = 46.5

$\hat{y} = 1.1x - 80.0$

#	x	y
1	70	3
2	115	45

Residual is y minus y-hat

Patient #1: 3 – (-3) = 6

Patient #2: 45 – 46.5 = -1.5

Bottom Line: You don't want big residuals, because that would mean the line didn't fit very well.

Using Least Squares Line Equation for Prediction

- Let's say you knew someone's DBP and you wanted to predict how many appointments s/he would have next year
- You can plug the DBP in as x, and get y-hat out, and say that's your prediction
- If you use an x within the range of the original equation (70-125), this type of prediction is called interpolation.
- If you use an x from outside the range (such as 65, or 130), it is extrapolation – not a great idea.

#	x	y
1	70	3
2	115	45
3	105	21
4	82	7
5	93	16
6	125	62
7	88	12
	Σx = 678	Σy = 166

Example of Interpolation

The patient in your study has a DBP of 80. That is within the range of your x's. Let's predict how many appointments he will have next year. Here's the equation:

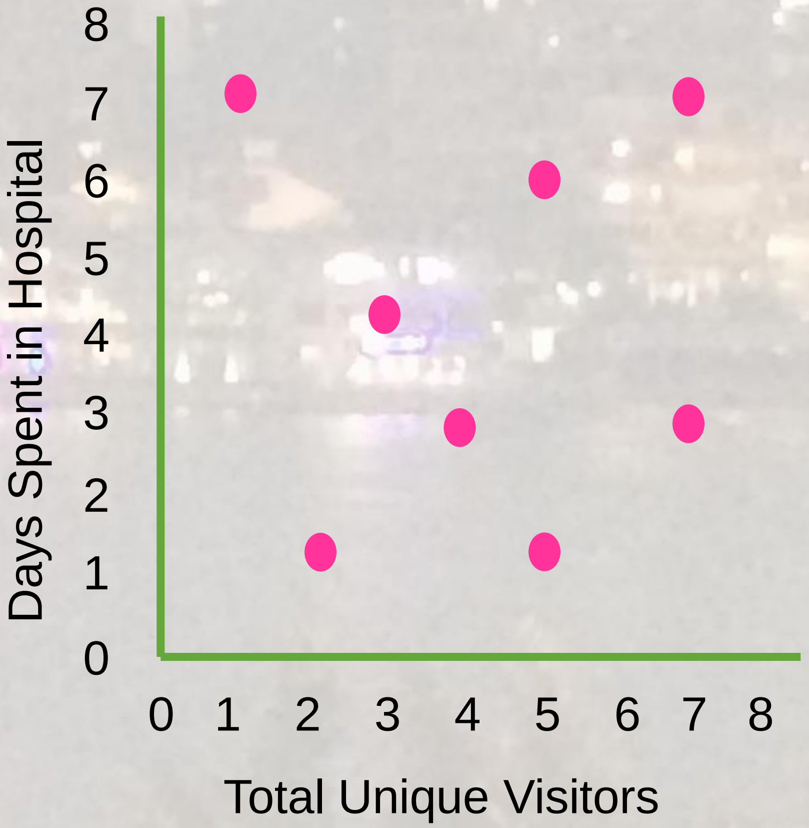
$\hat{y} = 1.1x - 80.0$

(1.1 * 80) – 80 = 8, so we predict this patient will come to 8 appointments next year.

#	x	y
1	70	3
2	115	45
3	105	21
4	82	7
5	93	16
6	125	62
7	88	12
	Σx = 678	Σy = 166

Is it Really This Easy to Make Predictions Using the Least Squares Line?

- No. You can make a linear equation out of any x,y pairs.
- If there is no linear correlation, though, the line is meaningless for prediction.
- Imagine a line for this scatter plot – would that really work for prediction?
- To evaluate if our least-squares line equation is should be used for interpretation, we use the Coefficient of Determination



Coefficient of Determination

Get out the r!

The Coefficient of Determination (CD)

- This is r^2 (in other words, r times r)
  - Then, like CV, we turn it into a %
- In the example, our r=0.95
- 0.95 * 0.95 = .90
- CD = 90%
- 90% = explained variation in y (by the linear equation)
- 100% - 90% = 10% unexplained variation
- "90% of the variation in the number of appointments is explained by DBP."
- "10% of the variation in the number of appointments is NOT explained by DBP."
- What happens if the CD is low?
  - CD should be better than at least 50% (random)
  - The higher, the better
  - If it is low, it means other variables might be needed to explain more of the variation

Chapter 4 Summary

- We started with quantitative x,y pairs
- We made a scatterplot to look at the linear relationship between x and y, and look at outliers
- We calculated r to see if our correlation was positive or negative, and weak, moderate or strong
- We calculated b and a to come up with the least-squares line equation
    - Notice: the sign on b will always match the sign on r (negative or positive)
    - Also notice: Strong correlations will give you high CDs
- We used the linear equation to calculate residuals
- We used r to calculate the CD to decide if we wanted to use the linear equation for prediction
- We decided it was good for prediction at 90%

Conclusion

- Least-squares criterion and calculating the least-squares line
- Reviewing issues with prediction using the least-squares line
- Coefficient of Determination (CD)

Photo courtesy of Fluzwup