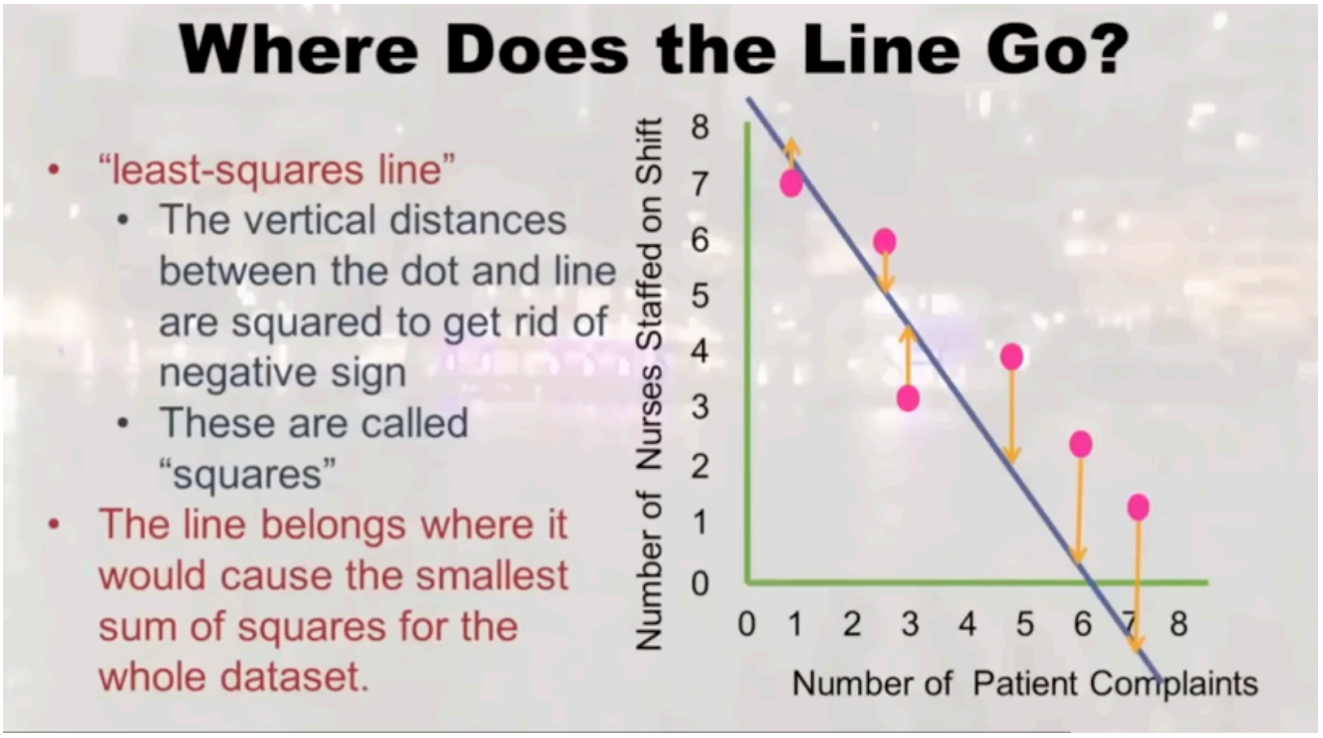
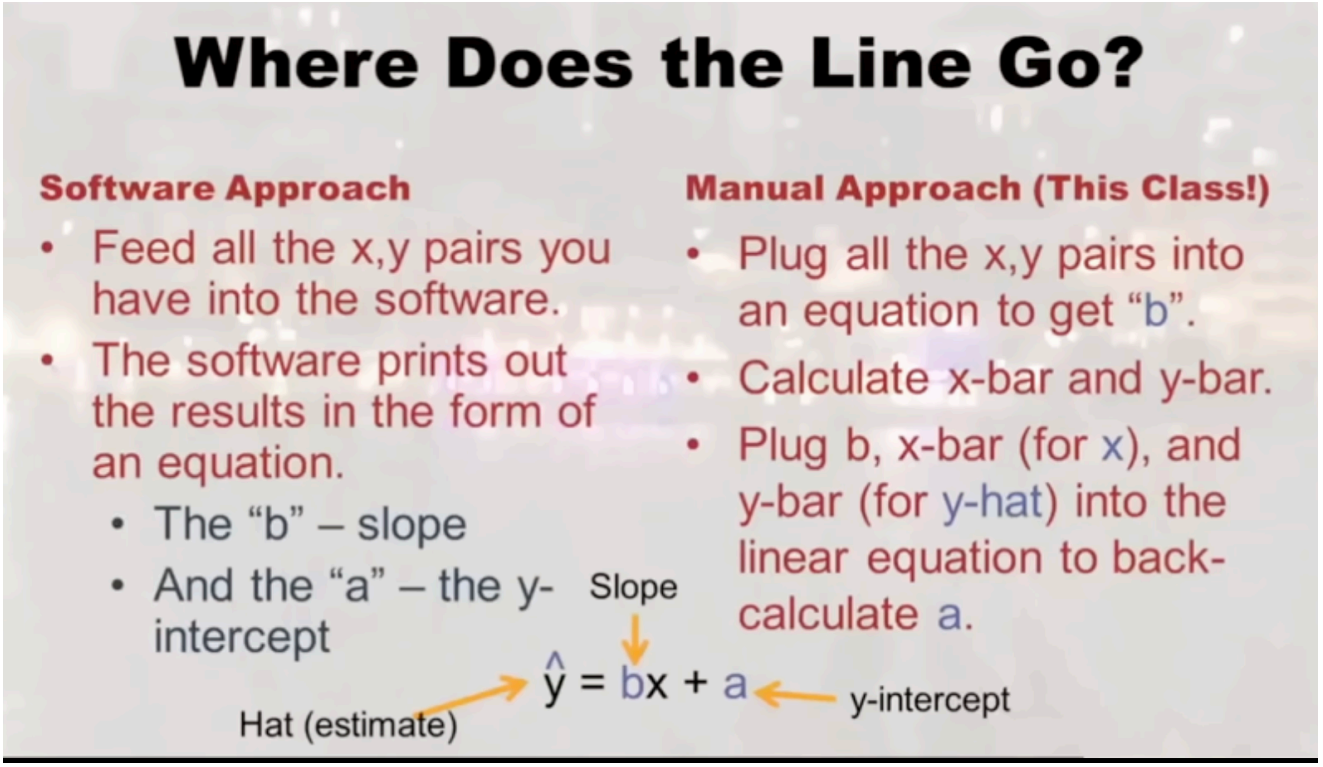
4.2 Linear Regression and the Coefficient of determination

Least-square Criterion



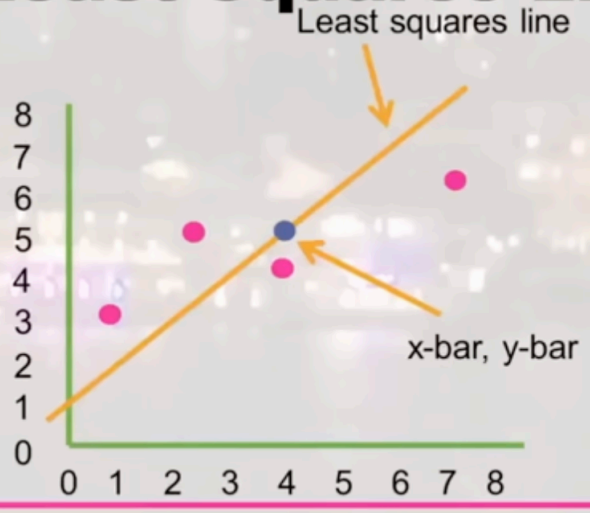


$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

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

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.

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.

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