Lecture #19

6/27/14

Admin
- Comp r.v.'s not on final (p268)

Plan
- review
- rest of ch 6
- ch 19.1 - 19.3 ■
- parts of ch 21 will also be covered

We talked about dependent r.v.'s and non-zero covariance:

$Cov(X,Y) \triangleq$

$E[(X-\mu_X)(Y-\mu_Y)] \neq 0 \Rightarrow X, Y$ dependent

Covariance was a hard metric to conceptualize

$\Rightarrow \rho_{XY} \triangleq Corr(X,Y) \triangleq \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \in [-1,1]$

value -1 is perfect neg. linear association

and +1 is perfect pos. linear association

Bivariate analysis of interval or ratio data.

So (X,Y) is related in pairs (X_1,Y_1), (X_2,Y_2), $\vdots$, (X_n,Y_n)

As usual, we do not know the parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y$ nor ρ. So we use statistics to estimate them (inference).

$$\bar{x} \to \mu_X, \ \bar{y} \to \mu_Y, \ s_X \to \sigma_X, \ s_Y \to \sigma_Y$$

$$s_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} \longrightarrow Cov[X, Y]$$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}} = \frac{s_{XY}}{s_X s_Y} \longrightarrow \rho_{XY}$$

"Simple cov" or just "cov"

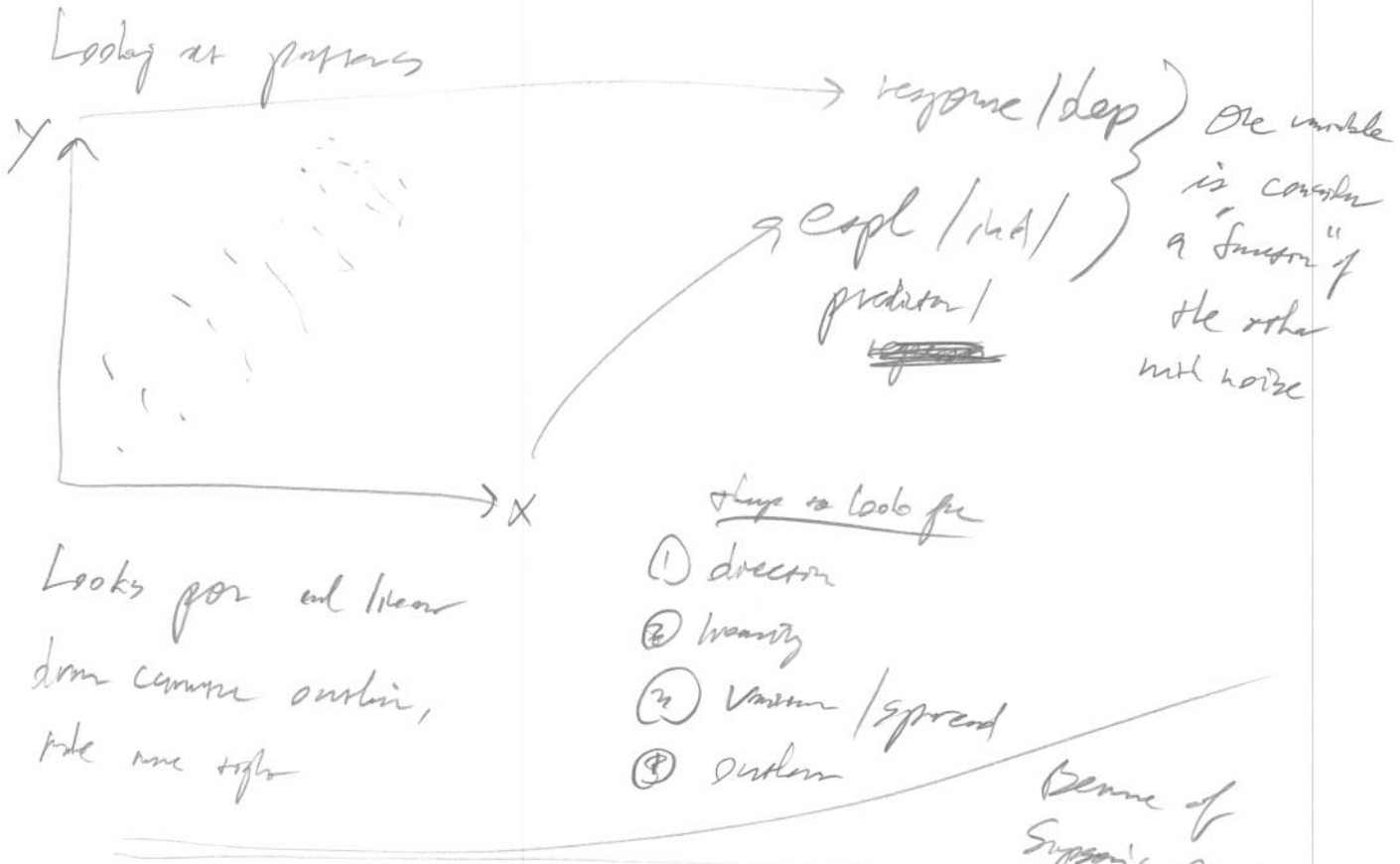
Just like there's hyp tests to test values of μ, there's hypothesis tests to test values of ρ which we won't cover since we're going to do something more powerful.

If there are many internal variables (imagine a table with lots of cols) then a cov matrix lists the single cov for each pair

	Age	Price	Size
Age		0.58	0.67
Price			0.13
Size			

pllt

talk about diagonal and symmetry



Ch 6.4: association with line (also all of ch 19)

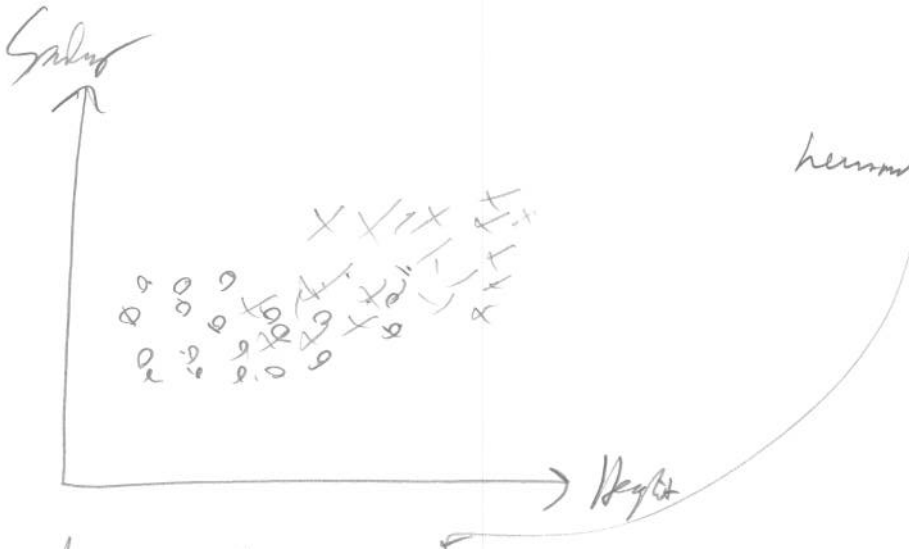
We will now define all these formulas

The method is "regression" — a strange word which will talk about tomorrow... it has two main goals:

① TO best predict a response given the explanatory value

② TO investigate how significant a predictor is in affecting change in the response

Goal #2 you will cover ad-nauseam in Stat 102. For now, our goal is #1.



however what do O's, X's are...

describe this association: linear, pos

Does this mean being taller causes you to earn more money? NO

Correlation $\not\Rightarrow$ Causation

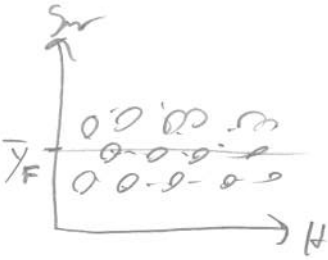
Spurious correlation

Diff between corr and association?

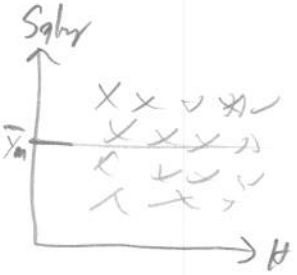
ch 6-5

We learnt before there may be a lurking variable and we can get Simpson's Paradox (when controlling for a lurking variable, association can change)

Let's say O's are women, X's are men

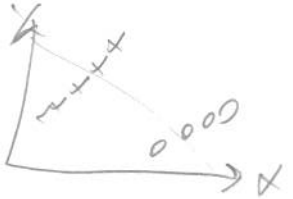


no correlation



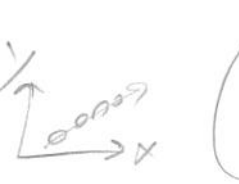
no correlation

What typical to be association? GONE! Why many times an association to begin with? Height tells you something about gender and men make more than women...

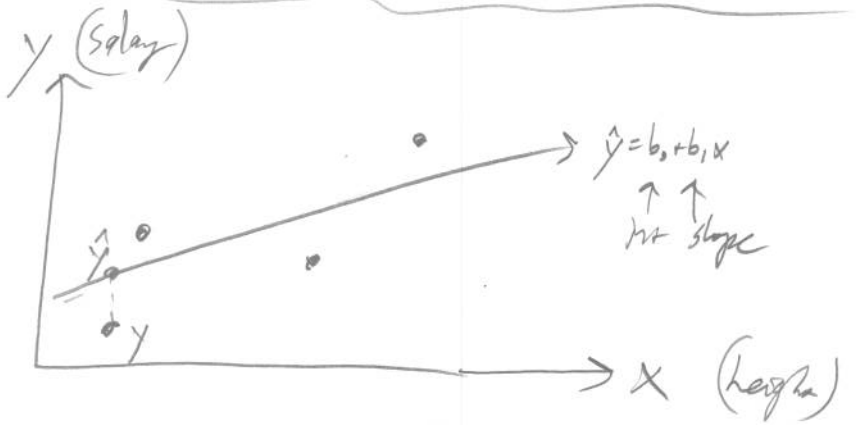


$\Rightarrow$

,



(Wikipedia)



Step I Assume relationship is linear which is why we call it "linear regression"

We want to draw the "best fit line" or "fitted line" or "least square regression line". How do we do this?

Maybe we want to pick some sort of "error" function. Call it err. err should take in our line's value at our actual y value and spit out an error.

→ Our predicted value we call $\hat{y}$.

$$err(\hat{y}, y)$$

What should this be?

Review these:

$$Var[X] = E[(X-\mu)^2] \quad \text{AND} \quad X \sim N(\mu, \sigma^2) \triangleq f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(X-\mu)^2}$$

the squared function

Squaring is very special above the squared function. It just seems natural. In advanced courses you'll see this choice in assumption gives you a lot of benefits.

Now, we must determine the total amount of this error over all pts... So we must determine:

$$\sum_{i=1}^{n} err(y_i, \hat{y}_i) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

$$= \Sigma (y_i - b_0 - b_1 x_i)^2$$

with respect to all possible vals of b_0, b_1. How do we do this? Remember calculus? Let's do b_0 first:

$$\frac{\partial}{\partial b_0} \left[\sum_{i=1}^{n} (y_i^2 + b_0^2 + b_1^2 x_i^2 - 2y_i b_0 - 2y_i b_1 x_i + 2b_0 b_1 x_i) \right] = 0$$

$$\Rightarrow \frac{\partial}{\partial b_0} \left[\Sigma y_i^2 + n b_0^2 + n b_1^2 x_i^2 - 2n\bar{y} b_0 - 2\Sigma y_i b_1 x_i + 2n b_0 b_1 \bar{x} \right] = 0$$

$$\Rightarrow 2n b_0 - 2n\bar{y} + 2n b_1 \bar{x} = 0$$

$$b_0 - \bar{y} + b_1 \bar{x} = 0$$

$$\Rightarrow \boxed{b_0 = \bar{y} - b_1 \bar{x}} \quad \checkmark$$

$$\frac{\partial}{\partial b_1} [\dots] = 0$$

$$\Rightarrow \boxed{b_1 = r \frac{s_y}{s_x}}$$

From summary stat bivariate statistics, you can compute slope + intercept of best fit line.

$$\hat{y}_i = \bar{y} - b_1\bar{x} + b_1 x_i = \bar{y} + b_1(x_i - \bar{x}) = \bar{y} + r\frac{s_y}{s_x}(x_i - \bar{x})$$

$$\hat{y}_i = \bar{y} + r\frac{s_y}{s_x}(x_i - \bar{x})$$

Look at this

$$\hat{y}_i - \bar{y} = r\frac{s_y}{s_x}(x_i - \bar{x})$$

$$\Rightarrow \frac{\hat{y}_i - \bar{y}}{s_y} = r\left(\frac{x_i - \bar{x}}{s_x}\right)$$

$$\Rightarrow \boxed{z_y = r z_x}$$ // done with ch. 6

- talk about residuals
- talk about $\bar{e} = 0$
- talk about $\bar{x}, \bar{y}$
- talk about RMSE