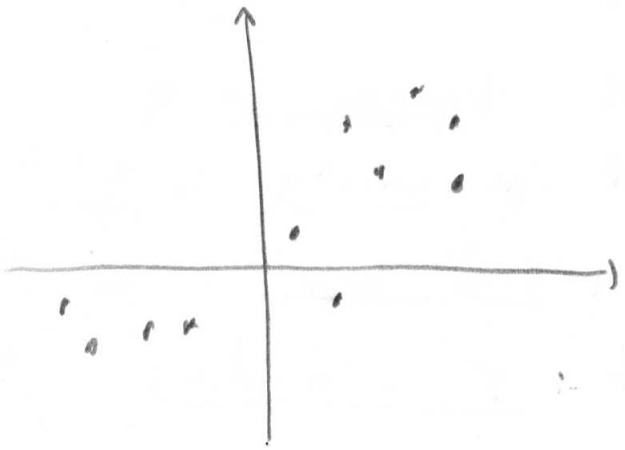
1/29/15 263 Lecture 5

Announcements

- Chap 2 webassign due tomorrow
- Exam I Next Thurs: Chapters 1, 2
- Quiz Today
- Excel HW 2 DUE TUE 2/3

Last time

* Data Set w/ 2 Quant Vars X & Y
* Scatter Plot: Plot $\{(x_i, y_i)\}$.



* Correlation: Compute $\bar{x}$, $\bar{y}$, s_x, s_y, z-scores for all data pts

$$z_{x_i} = \frac{x_i - \bar{x}}{s_x} \qquad z_{y_i} = \frac{y_i - \bar{y}}{s_y}$$

Then $$r = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} z_{y_i}$$

* $-1 \leq r \leq 1$

* Least Squares Line:

$$\hat{Y} = mX + b \qquad m = r\frac{s_Y}{s_X}$$

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

* Least Squares Line minimizes

$$\sum_{i=1}^{n} r_i^2 \qquad \text{"Residuals"} \quad r_i = Y_i - \hat{Y}(x_i)$$

(Y_i: Observed; $\hat{Y}(x_i)$: Predicted)

* Slope: an increase of 1 unit in X results in a change of m units in Y.
-OR- an increase of 1 std dev in X results in a change of r std devs in Y

* r^2 is the fraction of variation in Y "explained" by variation in X.

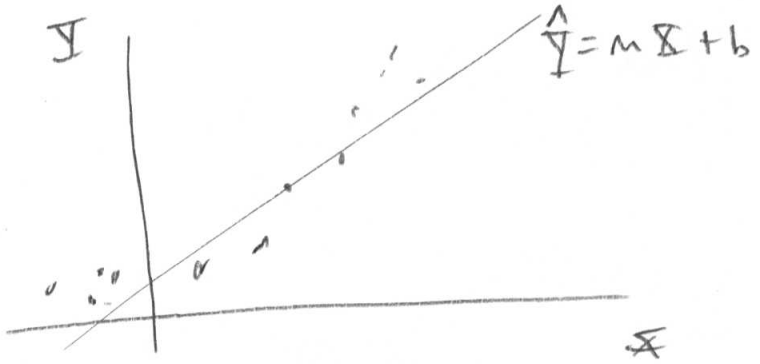
$$r^2 = \frac{\text{Variance of Predicted Vals}}{\text{Variance of Observed Vals}}$$

Residual Plots

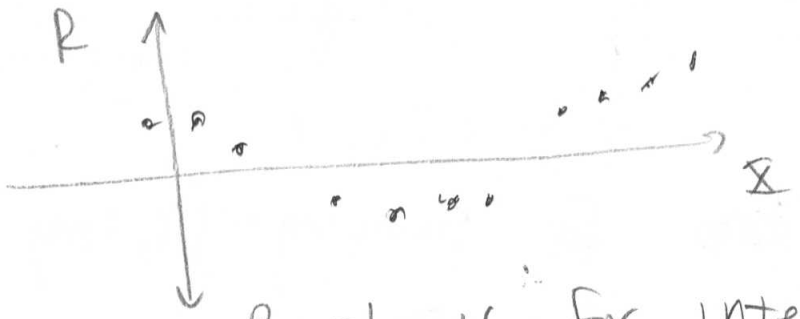
* Residual is Observed - Predicted

$$= Y_i - \hat{Y}_i$$

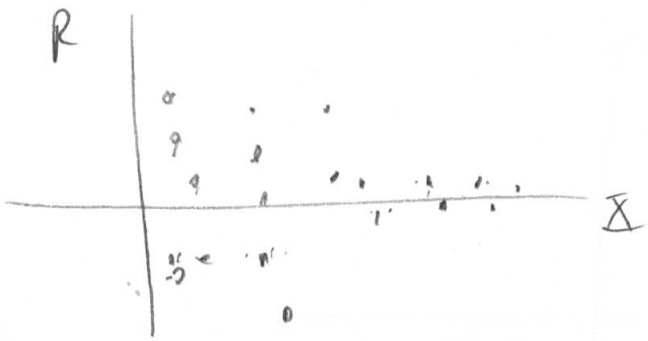
* Patterns in resids are bad



Residual:



* Look @ Residuals for Internet vs. Birth rate data
* From the Residual plot, what do you think the data is doing?



* Sometimes, a log can help:

$$Y = ab^X$$

$$\ln(Y) = X\ln(b) + \ln(a)$$

So A Plot of $(X, \ln(Y))$ should be a Straight line

$$Y = aX^p$$

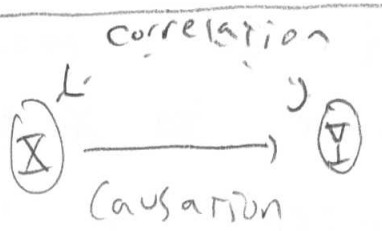
$$\ln(Y) = p\ln(X) + \ln(a)$$

So a Plot of $(\ln(X), \ln(Y))$ will be Straight.

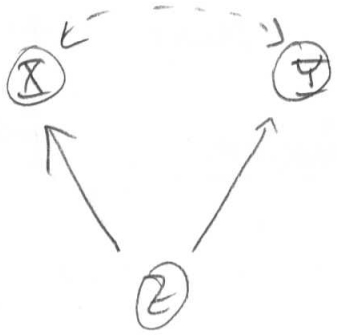
* Outliers: look @ data for Smoking Vs. lung Cancer. Outiers can be <u>influential</u>.

* Lurking Variables: Maybe both X & Y depend on Z? Examples.

Correlation vs. Causation



* Correlation is Necessary for causation, but but not Sufficient.



Common Response

Confounding

Example

Aspirin prevents Heart attacks

(See link on website)

* How do we establish Causation?
  - Perform Experiments
  - Eliminate lurking vars / Confounding
  - ("Control the influence of all vars")

Criteria for Causation

Sometimes an experiment is not practical or ethical. Example Exposure to X-Rays increases chance of cancer in human subjects. Can't do the experiment we want, which would be to expose many people to different levels of X-Rays (controlling for other factors)

So, we use existing data: Patients in hospitals:

X-ray exp Level (mGy)	# deaths from Cancer

- Strong Association between dose level & # Cancer deaths

- Consistent across many Hospitals / Region (controlling for environ. factors)

- Higher dose, stronger response

- Causality is logical in time
  - Can we eliminate those patients who already had cancer?

- The cause is plausible
  - Supporting evidence from e.g. bomb survivors, experiments on animals, etc.

Chapter 3: Producing data.

How do we collect data to answer statistical questions?

Conjectural / anecdotal data:

- I think that an increase in Vitamin C intake will decrease rate of common cold because I have heard it is TRUE.

- I have heard that individuals on welfare are more likely to do drugs

- I am told that vaccines cause autism.

Nullius in Verba motto of the Royal society - "take nobody's word for it"

Where do we get data from?

* "Available" data eg. US Census, NAEP, data.gov

* "Gathered" data: either observational

or Experimental

We observe & measure
E.g. Sunlight vs. Size of Plants

↓

We impose a treatment on cases/individuals to modify a variable

e.g. Artificial sunlight vs. Size of Plants., Asprin Dose, etc.

For Tuesday

- Read section 3.2
- Read MMR/Autism Paper Posted on website
- Bring Q's for Exam (Half of Class)