Biostatistics 201a - Lecture 11 - 10/17/11

Contents

- Correlation recap
- Regression assumptions
- CIs and tests for β_0, β_1

pages = 10

10/17/11

Correlation Recap:

- Measures strength and direction of a linear relationship between X and Y
- Calculated either as $\sqrt{R^2}$ in regression or

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} = \frac{SCP}{\sqrt{SSX \cdot SSY}} = \frac{Cov(X,Y)}{SD(X)SD(Y)}$$

The top piece $SCP = \sum (X_i - \bar{X})(Y_i - \bar{Y})$
= "sum of cross products"

Intuition:
- If X↑ when Y↑ and X↓ when Y↓ you get positive terms in SCP
- If X and Y go in opposite directions you get negative terms
- If X and Y unrelated you get a mix of +, − terms which cancel out and get correlation near 0

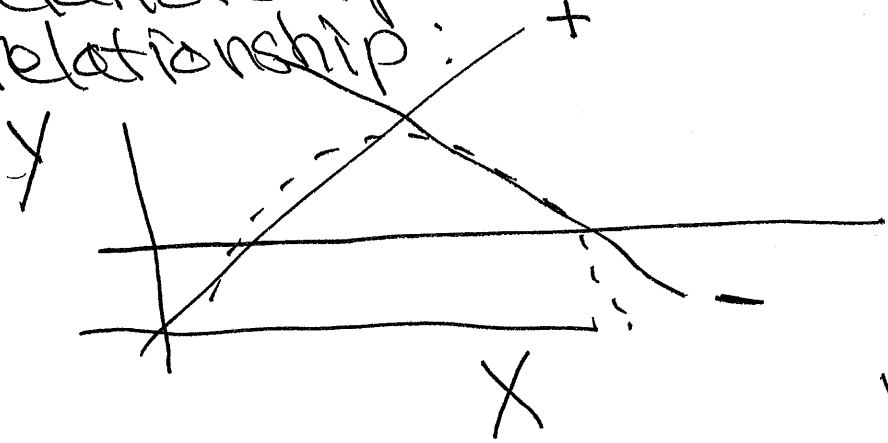
Value SCP is related to a quantity called "covariance" = how X and Y vary together

$$Cov(X,Y) = \frac{1}{n-1} SCP = \frac{1}{n-1} \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

Why not just use that?
Problem is this is unit dependent so hard to tell how big it is. To fix this we standardize the covariance by dividing by the standard deviations of X and Y. ⇒ Resulting properties of correlation

① $r > 0 \Rightarrow$ positive relationship
$r < 0 \Rightarrow$ negative relationship
no relationship $\Rightarrow r = 0$

However $r=0$ does not mean no relationship - it means no linear relationship:



best fitting line flat and $r=0$ but X and Y are beautifully curvilinearly related!

② r is unit free
This means you can compare correlations across data sets.

③ $-1 \leq r \leq 1$

(4) Values near 1 in absolute value ⇒ strong relationships ($r = \pm 1$ means points lie perfectly on a line)
values near 0 mean weak (linear) relationships

(5) Correlation fits our definition of an effect size.

Inference in Regression:

Measures F, RMSE, R^2, r helped measure how well line fits the data points but (except for p-value in F test) they don't depend on the distribution of X and Y. If we formal inference (CIs, tests, etc.) we need to make some assumptions:

Pre-assumptions
(a) A straight line is a reasonable model
(b) X values are "known" or measured without error — not random

The main assumptions we then need are about the errors in the model:

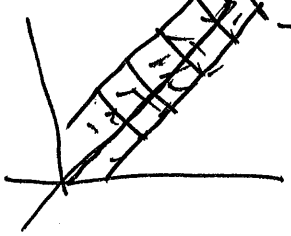
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

ε_i: individual variation about the average Y at a given X.

We assume

① The ε_i's have a normal distribution. This lets us assume b_0 and b_1 are also normal so we can do t-tests and CIs.

② The ε_i's are mean 0 – pictorially means line goes through the middle of data cloud for all X values.



If this is true, b_0 and b_1 are unbiased estimates of β_0 and β_1 ("right on average")

③ The ε_i's are independent – there's nothing systematic about them – like saying we got a simple random sample.

(4) The ε_is are "homoscedastic" - variation of points about the line is the same for all X. Makes variance of b_0 and b_1 much easier to compute and makes RMSE a sensible measure of how good our predictions are.

We'll learn soon how to check these and what to do if they're wrong.... If they are correct, b_0 and b_1 are normally distributed with means β_0 and β_1 and standard errors we can compute ⇒ lets us do CIs and tests about β_0, β_1.

Confidence Intervals: Basic form

$$b_0 \pm t_{\alpha/2,\, n-2} \cdot s_{b_0}$$
$$b_1 \pm t_{\alpha/2,\, n-2} \cdot s_{b_1}$$

standard errors we can get from STATA/SAS - depend on σ^2, n, variation in Xs

d.f. match MSE which is our estimate of σ^2

Radiation Example:

β_0: CI [95.69, 133.74] - on average in areas where there is no radiation exposure, cancer mortality rate is between 96 and 134 deaths / 100,000 people. Since these rates are all above 0 we're sure (95%) that there's cancer even without radiation.

$X = 0$

β_1: CI [5.88, 12.59] - On average every extra point of radiation exposure is associated with an increase in mortality rate of between 5.88 to 12.59 deaths per 100,000. Again whole CI is above 0 $\Rightarrow$ positive relationship between radiation and mortality rate.

Hypothesis tests for β_0, β_1:
Most standard test we do is whether $\beta_1 = 0$ - why?

$Y_i = \beta_0 + \cancel{\beta_1 X} + \varepsilon$ ($\beta_1 X \to 0$) If $\beta_1 = 0$ then X drops of model and there's no linear relationship

$H_0: \beta_1 = 0$ - no linear relationship between X and Y

$H_A: \beta_1 \neq 0$ - there is a relationship...

Test statistic: (b_1 ← sample slope; 0 ← null hypothesis slope)

$$t_{obs} = \frac{b_1 - 0}{s_{b_1}}$$

(s_{b_1} ← standard error of estimate)

= "how far" sample slope is from 0

Large values of t_{obs}, either positive or negative, imply a relationship

$p\text{-value} = 2P(t_{n-2} \geq |t_{obs}|)$

($n-2$ ← d.f. for MSE)

Reject if $p\text{-value} < \alpha$

Can do this for β_0 as well

Computer packages give these automatically.

Cancer Data

slope:

$H_0: \beta_1 = 0$ – no relationship between radiation exposure / mortality

$H_A: \beta_1 \neq 0$ – there is a relationship between radiation / mortality

$$t_{obs} = \frac{b_1 - 0}{s_{b_1}} \overset{\text{STATA}}{=} \frac{9.23 - 0}{1.42} = 6.51$$

$$p\text{-value} = 2P(t_7 \geq |6.51|) \approx 0.000$$

↑ on STATA printout or use tail

Reject H_0 at $\alpha = .05$, conclude there's a relationship

Intercept:

$H_0: \beta_0 = 0$ – when no radiation exposure cancer mortality rate is 0

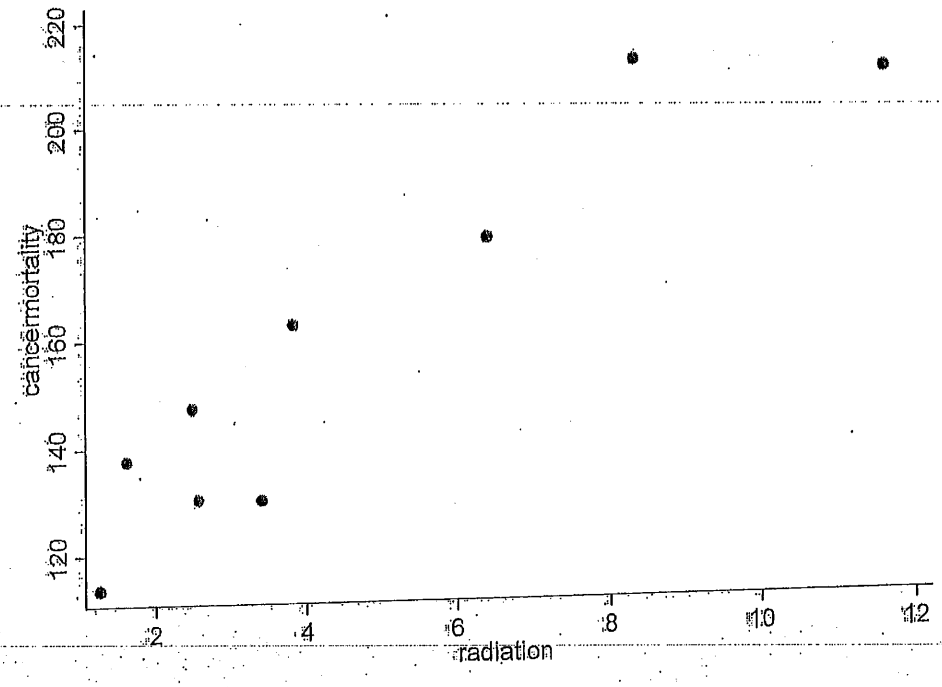
$H_A: \beta_0 \neq 0$ – there is cancer mortality even when no radiation

$$t_{obs} = \frac{b_0 - 0}{s_{b_0}} = \frac{14.7 - 0}{8.05} = 14.26$$

p-val $\approx 0 < \alpha$

$\Rightarrow$ reject H_0, conclude there's cancer mortality even w/ no radiation

Next time: How to test values of β_0, β_1 other than 0 (not on regression printout)



. reg cancermortality radiation

Source	SS	df	MS
Model	8309.55541	1	8309.55541
Residual	1373.94679	7	196.278113
Total	9683.5022	8	1210.43778

Number of obs = 9
F(1, 7) = 42.34
Prob > F = 0.0003
R-squared = 0.8581
Adj R-squared = 0.8378
Root MSE = 14.01

Table for inference on B_0, B_1 same

cancermort~y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
radiation	b_1 = 9.231456	s_{b_1} 1.418787	t_{obs} 6.51	2-sided p-obs 0.000	5.876557	12.58635
cons	b_0 = 114.7156	s_{b_0} 8.045664	14.26	0.000	95.69066	133.7406

X = intercept

95% CIs for B_0, B_1

	X	Y	X^2	Y^2	XY
	2.49	147.1	6.2001	21638.41	366.279
	2.57	130.1	6.6049	16926.01	334.357
	3.41	129.9	11.6281	16874.01	442.959
	1.25	113.5	1.5625	12882.25	141.875
	1.62	137.5	2.6244	18906.25	222.75
	3.83	162.3	14.6689	26341.29	621.609
	11.64	207.5	135.490	43056.25	2415.3
	6.41	177.9	41.0881	31648.41	1140.339
	8.34	210.3	69.5556	44226.09	1753.902
Sum	41.56	1416.1	289.42	232498.97	7439.37
Mean	4.618	157.34			