Biostatistics 201a - Lecture 10 - 10/14/11

Contents

- Regression ANOVA table
- F-test
- RMSE
- R^2
- correlation (w/ Bonus material)
- printout w/ F, RMSE, R^2 marked

pages = 13

10/14/11

How do we tell how good a job our SLR is doing?

We do this by, as in ANOVA, breaking the variability in Y into a piece attributable to X and a piece that isn't

⇒ ANOVA table

n = total # subjects
p = # of X variables (for now p=1)

	Source	SS (sum of squares)	df	MS (mean squares)	F	Prob > F
Between	Model	SSR	p = 1	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$	p-value
Within	Error	SSE	$n-p-1$ $= n-2$	$MSE = \frac{SSE}{n-p-1}$		
	Total	SST				

SSR = "sum of squares for regression"

$$= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \rightarrow \text{like } \sum (Y_j - \bar{Y})^2 \text{ in ANOVA}$$

$\hat{Y}_i$: value predicted by model; $\bar{Y}$: average value of Y

= variability in Y explained by X

SSE = "sum of squared errors"

$$= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \longrightarrow \text{just like } (Y_{ij} - \bar{Y}_j)^2$$

Y_i: actual / observed value

$\hat{Y}_i$: predicted value (on line)

piece in ANOVA

= squared distance (summed) from the points to the line

= variability in Y not explained by X / error we make when using X to predict Y

Turns out that

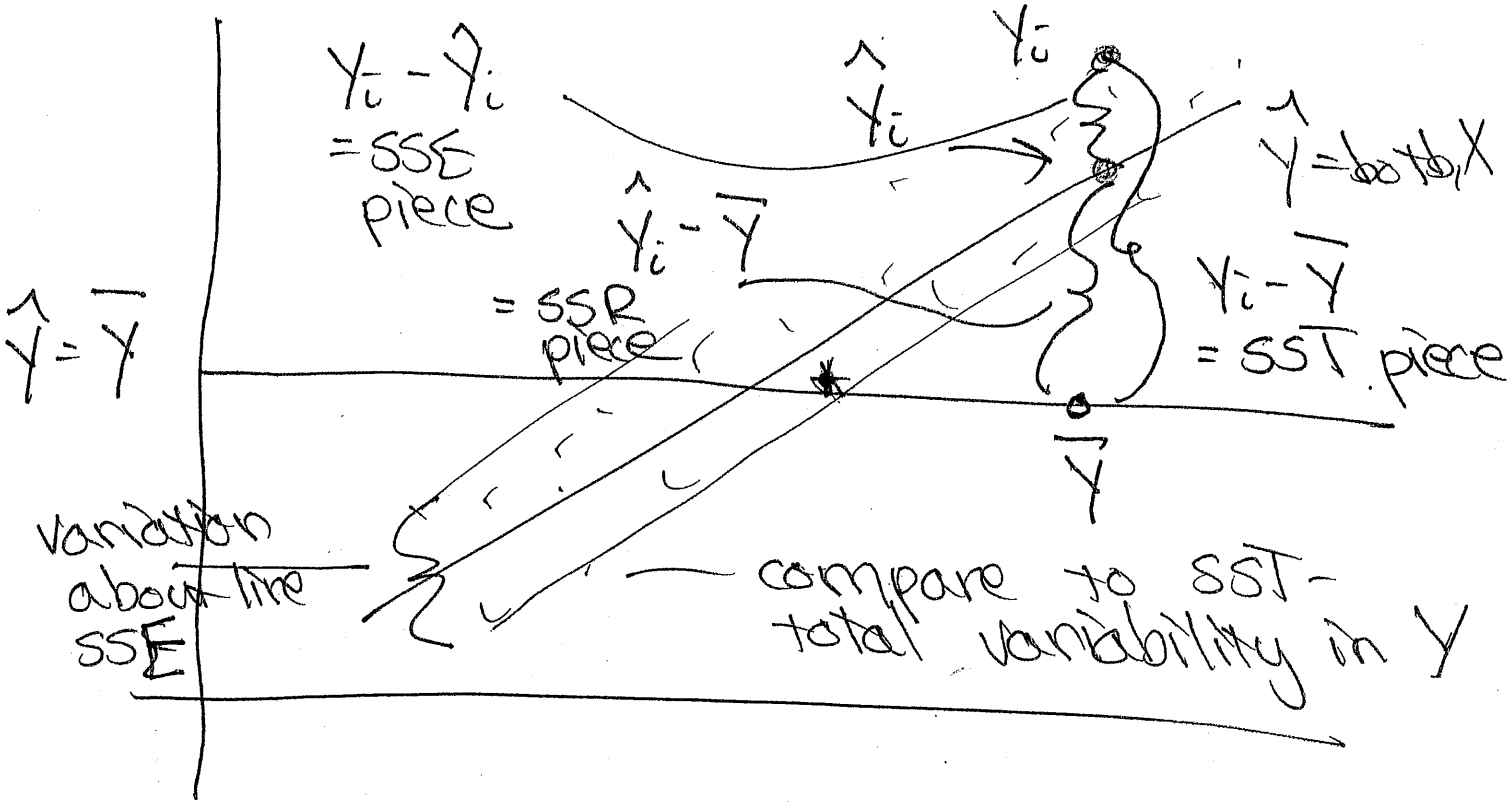
SST = SSE + SSR

(just like SST = SSW + SSB)

SST: total variability in Y

= "sum of squares total"

= "errors" you would make if you always used $\bar{Y}$ as your prediction of Y and ignored X entirely



In general $SS\Sigma \leq SST$ with equality only if X is useless and $\beta_1 = 0$ - i.e. the flat line $\hat{Y} = \bar{Y}$ is the best fitting line

Note that while using X improves the fit overall it doesn't have to give better predictions for each individual point (eg. the one marked *)

MSR = "mean squares for regression"

$= \frac{SSR}{p} \overset{\text{in simple linear reg}}{=} \frac{SSR}{1}$ not very exciting.

= "average variability explained per predictor"

MSE = "mean squared error"

$= \frac{SSE}{n-p-1} \overset{\text{simple linear reg.}}{=} \frac{SSE}{n-2}$

= average squared error per data point (error made using X to predict Y)

$F = \frac{MSR}{MSE} = \frac{\text{explained}}{\text{unexplained}}$ variability

The bigger F is the better we're doing!

Degrees of freedom:

SST: $n-1$ just as before

SSE: $n-2$ - why? Variability about the estimated line

To get the line I need estimate the slope and intercept, b_0 and b_1 - i.e. 2 numbers. So $n-2$ df. left for variability about line

SSR: df = 1 - I add 1 piece of info (b_1) to my model when I use the sloped line instead of just $\bar{Y}$

= # of predictors = p

Evaluating the model

① Is there a significant relationship between X and Y?

Model: $Y = \beta_0 + \beta_1 X + \varepsilon$

No relationship $\Leftrightarrow$ $\beta_1 = 0$ when X drops out of model

H_0: $\beta_1 = 0$ - no (linear) relationship between X and Y; the model is not useful

H_A: $\beta_1 \neq 0$ - there is a relationship; X helps predict Y

Test statistic $F_{obs} = \frac{MSR}{MSE}$ big F $\Rightarrow$ X explains alot

It turns out under H_0 ($\beta_1 = 0$) f_{obs} has an F distribution with $p=1$ and $n-p-1 = n-2$ degrees of freedom

$$p\text{-value} = P(F_{1,n-2} \geq f_{obs})$$

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

(a) Does the model make good predictions?

Root Mean Squared Error:

$$RMSE = \sqrt{MSE} = \sqrt{SSE/n-2}$$

= average distance from the points to the line

To tell how large this is we need a reference point - usually the average value of Y is a good a reference.

e.g. Cancer data: $MSE = 196.27$

$= RMSE = \sqrt{196.27} = 14.01$

on average we'll be off in our predictions of mortality by 14 people per 100,000 use radiation as predictor

My overall mortality rate in sample was 157 deaths / 100,000

$\frac{14}{157} \approx .089 = 8.9\%$ mistake

(3) Is using X a big improvement over not using X? What percentage of the variability in Y is explained by X?

Number for this is called R-squared:

$$R^2 = \frac{SSR}{SST} = \frac{\text{explained}}{\text{total}}$$

Cancer data: $SSR = 8309.55$

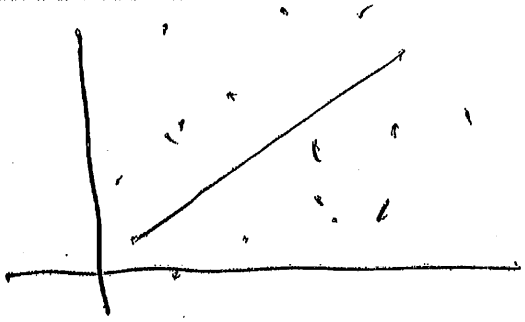
$SST = 9683.5$

$$R^2 = \frac{8309.55}{9683.5} = .8581 = 85.81\%$$

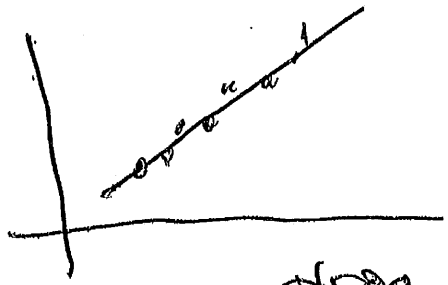
We've explained over 85% of the variation in communities' cancer mortality rates just by knowing radiation exposure. Pretty good

Bonus slides on correlation

(4) How strong is the relationship between Y and X?



not strong



very strong rel- almost perfectly

For this we use a measure called correlation, denoted r.

Note: $R^2 = \frac{SSR}{SST} = \frac{b_1 \cdot SCP}{SST} = \frac{SCP}{SSX} \cdot \frac{SCP}{SST}$

$$= \frac{\frac{1}{n-1} SCP \cdot \frac{1}{n-1} \cdot SCP}{\frac{1}{n-1} SSX \cdot \frac{1}{n-1} \cdot SST}$$

$$= \frac{\left(\frac{1}{n-1} SCP\right)^2}{Var(X)\, Var(Y)}$$

Square root:

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

$\equiv$ correlation of Y and X

Top piece = $\frac{1}{n-1}$ SCP = covariance of Y and X

= measures how Y and X change or vary together.

positive relationship — $\text{Cov}(X,Y) = (x_i - \bar{x})(y_i - \bar{y})$

X high when Y high + · + = +

X low when Y low − · − = +

negative relationship

X ↑ Y ↓ + · − = −

X ↓ Y ↑ − · + = −

no relationship → random mix of +, − terms cancel out → Cov ≈ 0

Problem w/ covariance — units attached so hard to tell what is big.

If we divide Cov(X,Y) by SD(X) and SD(Y) we get rid of the units!

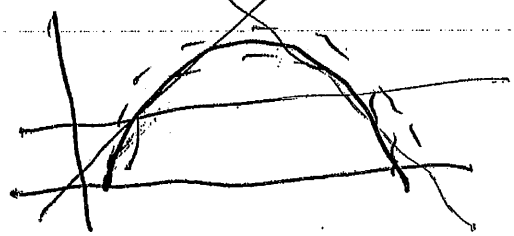
Resulting properties of correlation

① r positive $\Leftrightarrow$ positive linear relation

r negative $\Leftrightarrow$ negative " "

r 0 if no linear relation

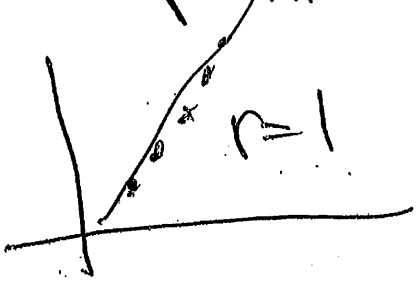
↳ Aside $r=0$ does <u>not</u> mean no relationship - just not a linear one

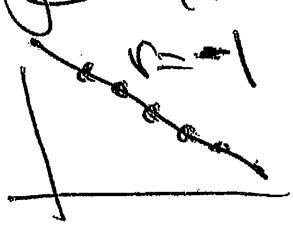


② correlation is unit free so it can be compared across data sets.

③ $-1 \leq r \leq 1$ always

④ Values near ± 1 mean strong relationship; <u>exactly</u> ± 1 it means points lie exactly on a line





sign → direction of relationship but nothing about strength

Rough rule of thumb:

$|r| \geq .7$ usually considered strong

$|r| \approx .5$ medium

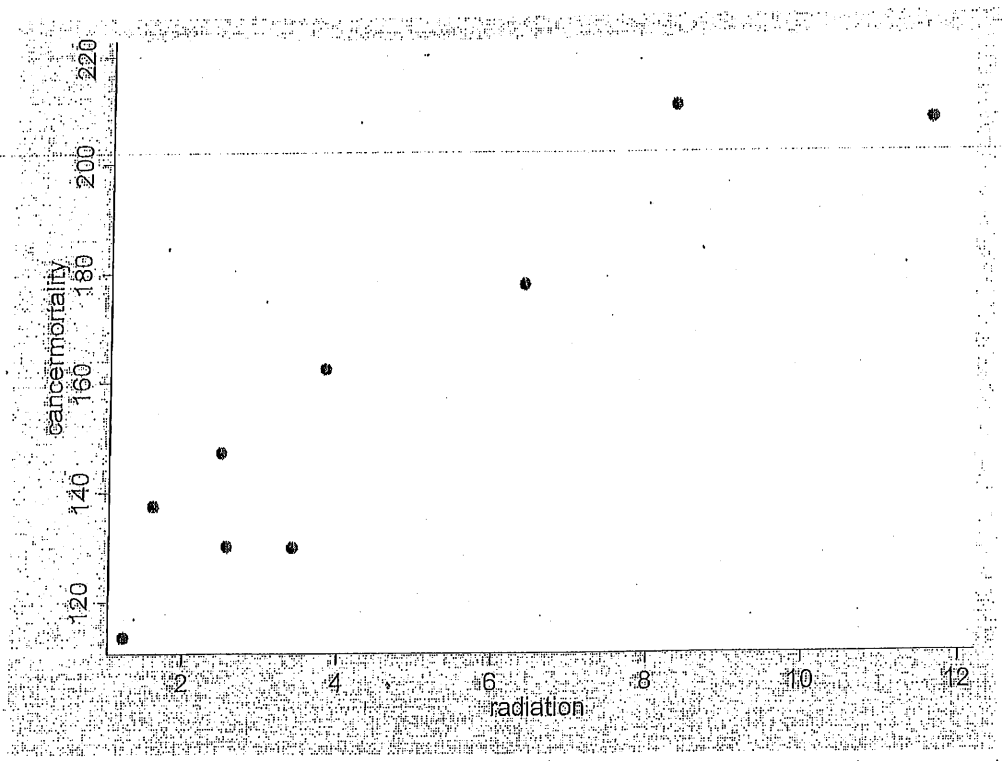
$|r| \leq .3$ fairly weak

But — depends on the context and field.

Cancer mortality data:

$r = \sqrt{R^2} = \sqrt{.85} = .926$

very strong positive relationship



Fobs
p-value

. 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

cancermort~y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
radiation	9.231456	1.418787	6.51	0.000	5.876557	12.58635
_cons	114.7156	8.045664	14.26	0.000	95.69066	133.7406

	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			