Biostatistics 201a - Lecture 9 - 10/12/11

Contents:

- SLR Model
- Interpretation of Parameters
- Least Squares estimates, prediction
- Intro to measures of model fit (F, RMSE, R^2)
- Cancer Mortality Example

\# pages = 10

10/12/11

Linear Regression Basics

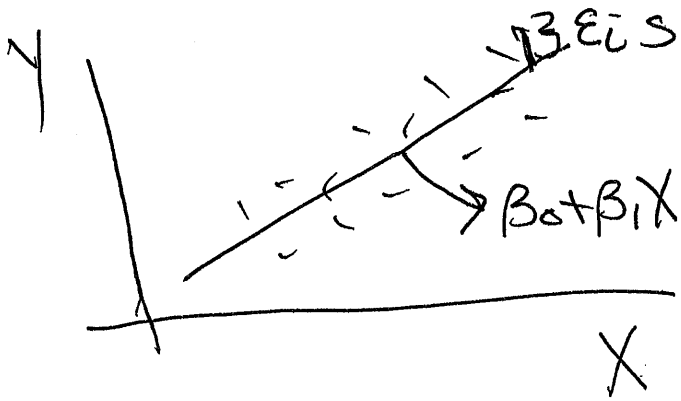
Y = outcome or response variable - thing we're interested in studying

X = predictor variable which we believe is related to and will help us explain Y.

Examples:

- Y = cancer mortality (deaths/100,000 people) in a community near a power plant
  X = radiation exposure level in com.
  expect as X goes up, Y goes up
- Y = miles driven by a car
  X = amount of gas in tank
  expect as X goes up, Y goes down or vice versa

Start by assuming a linear relationship between Y and X - i.e. if we plot Y vs X we get, on average, a straight line



Fits not going to be perfect. Account for uncertainty in model

Linear Regression Model:

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

- Y_i: outcome for i^{th} subject, $i = 1, 2, \ldots, n$
- β_0: intercept
- β_1: slope
- X_i: value of predictor for subject i
- ε_i: individual variation or "error" (not due to X)

$\beta_0 + \beta_1 X_i$: average value of Y at a given value of X - this is the line

This model has 3 parameters - β_0, β_1, and σ^2 which describes on average the variability of points about the line (as in ANOVA we assume $\varepsilon_i \sim N(0, \sigma^2)$)

- ε_i: errors
- N: normal
- 0: pts centered on line
- σ^2: variability

Strength of the relationship depends on how big σ^2 is relative to the variation overall in Y. If $\sigma^2 = 0$, the points lie exactly on the line.

Interpreting the Parameters

① β_0 = intercept = average value of Y when X=0 - may or may not be interesting depending on whether X=0 is an important value.

e.g. cancer / radiation

β_0 = average cancer rate for people with 0 radiation exposure. (i.e. population base rate)

e.g. car example

β_0 = average # of miles driven when tank hits empty - i.e. how far can you drive on a tank of gas

② β_1 = slope = change in Y associated with a 1 unit change in X. i.e. on average if X is 1 pt higher Y is β_1 points higher

e.g. cancer / radiation

for every extra point of radiation exposure, how much does cancer mortality change (prob. go up) on average.

e.g. car

for every extra gallon of gas used from the tank how far have you driven on average - MPG

(3) σ^2 = average (squared) distance from points to the regression line - if you take the square root you get ≈ the average error you make when using X to predict Y.

β_0 + β_1, σ^2 are population values (unknown) that we'll estimate using sample data. We write

$$\hat{Y} = b_0 + b_1 X$$

$\hat{Y}$: estimated/predicted value of Y

b_0 + b_1: intercept and slope of the line that "best fits" the sample data

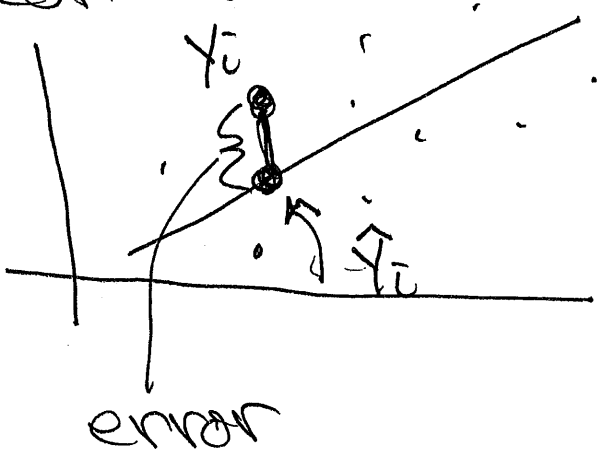
What do I mean by "best fit"?

→ unbiased (right on average)

→ Line to go "close" to all the points

→ Maximum likelihood - what is the population line most likely to have produced the observed data?

All these criteria (if you assume the errors are normally distributed) lead to the "Least Squares" estimate



Want values of b_0 and b_1 that minimize the ~~sum~~ sum of squared distances from the points to the line, i.e.

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

Y_i: actual; $\hat{Y}_i$: predicted value

Using calculus: differentiate with respect to b_0 and b_1, set the results equal to 0 and solve the system of equations.

If you do this you get

$$b_0 = \bar{Y} - b_1\bar{X} \rightarrow \bar{Y} = b_0 + b_1\bar{X}$$

($\bar{Y}$: mean value of Y; $\bar{X}$: mean of X; so best line goes through mean of data)

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{SCP}{SSX}$$

SCP = "sum of cross products"

SSX = "sum of squares for X"

We let STATA "fit" this model we get

$b_0 = 114.7$ = average mortality rate per 100,000 people when radiation exposure is 0.

This implies, since it's non-zero, that there are risk factors for cancer besides radiation

$b_1 = 9.23$ - on average for every extra point of radiation exposure we get on average 9 more deaths from cancer per 100,000 people.

Note the units! b_0 is always in the same units as Y; b_1 is in units of Y / units of X.

Making Predictions: Simply plug in the value of X of interest to the estimated equation $\hat{Y} = b_0 + b_1 X$

e.g. What is the average mortality rate if the radiation exposure is $X = 10$?

$$\hat{Y} = 114.7 + 9.23(10) = 207 \text{ deaths per } 100{,}000$$

Question: How good a job is our model doing?

(1) Is a linear model even reasonable? Plot Y vs X

(2) Even if a line looks good for your data, be wary about using it to make predictions

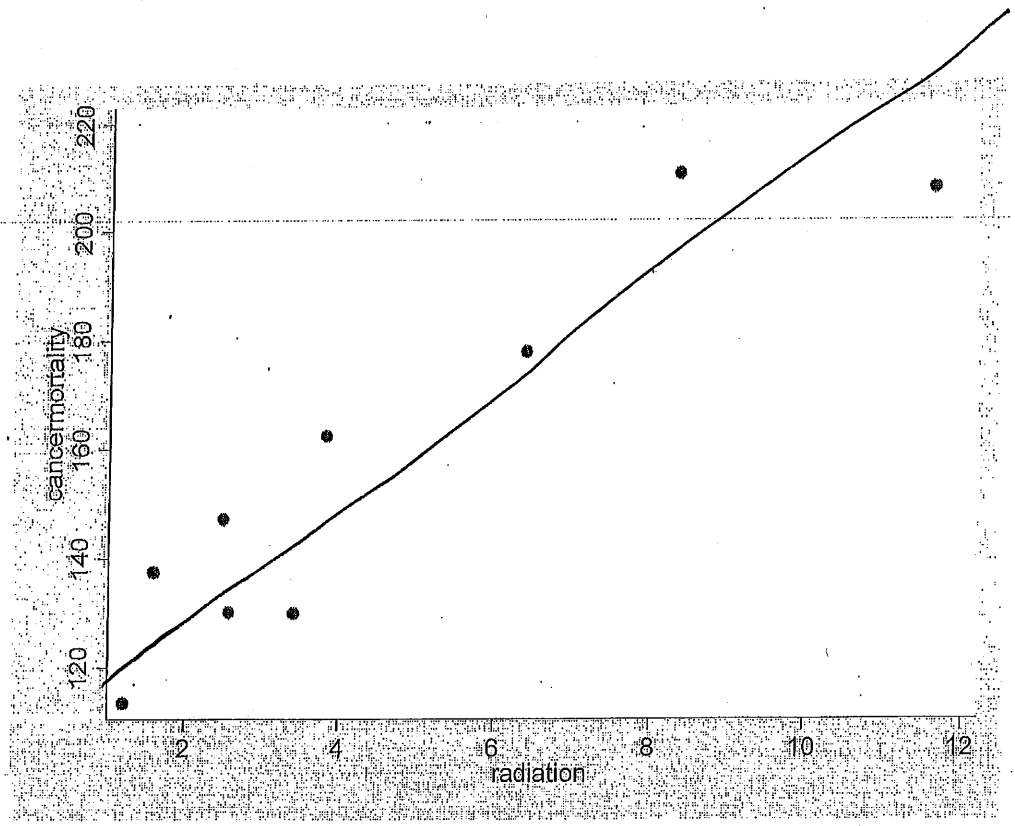
outside range of X values in your sample (extrapolation)

Measures of how well we're doing

(1) F-test - Is there a statistically significant relationship?

(2) How good are the predictions? Compare average error (RMSE) to Y values

(3) Does X explain a large fraction (R^2) of the variability in Y? Is X strongly related to Y (correlation)?



ANOVA table

How well model fits

```
. 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 (p-val)
R-squared	= 0.8581
Adj R-squared	= 0.8378
Root MSE	= 14.01

→ estimate of σ

cancermort~y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
radiation (b_1)	9.231456	1.418787	6.51	0.000	5.876557	12.58635
_cons (b_0)	114.7156	8.045664	14.26	0.000	95.69066	133.7406

↑ tests and CIs

	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			