Chp 5 Regression

Vocab/Defns

- **regression line**: a straight line that describes how a response variable y changes as an explanatory variable x changes; we use regression line to predict y, given x.
- **line eqn** (slope-intercept): $y = mx+b$ where m = slope, & $(0,b)$ is y-intercept.
  (in the book, they use $y = a+bx$)

  ☆ intercept may or may not be meaningful, depending on data and whether $x=0$ makes sense

  ☆ cannot determine how important a relationship is by looking at slope value; if we change units of measure, the slope can change drastically.

  ☆ sign of slope is useful information

- **least-Squares Regression line**: Rather than "eyeballing" the best fit line, this line is the best fit such that the sum of the squares of vertical distances of data pts from line is as small as possible.

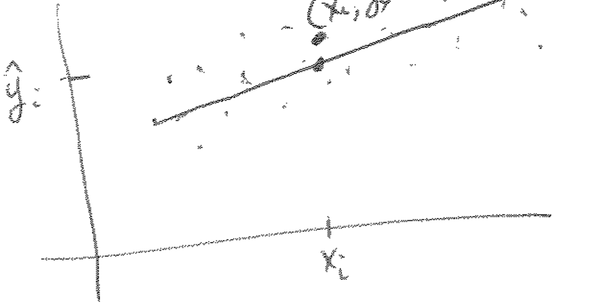
  i.e. line such that $\sum_{i=1}^{n} |\hat{y}_i - y_i|$ is a minimum.

- Eqn: $\boxed{\hat{y} = a+bx \quad w/ \quad b = r\left(\frac{s_y}{s_x}\right) \quad \& \quad a = \bar{y} - b\bar{x}}$

  $\bar{x}$ = mean of all x_i s_y = s.d. for y_i r = correlation

  $\bar{y}$ = mean of all y_i s_x = s.d. for x_i

  ☆ $\hat{y}$ "y hat" is predicted value



Chp 5 (cont)

Ex 1

Do heavier people burn more energy? We have data on the lean body mass and resting metabolic rate for 12 women who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate, in calories burned per 24 hours, is the rate at which the body consumes energy.

Mass	36.1	54.6	48.5	42.0	50.6	42.0	40.3	33.1	42.4	34.5	51.1	41.2
Rate	995	1425	1396	1418	1502	1256	1189	913	1124	1052	1347	1204

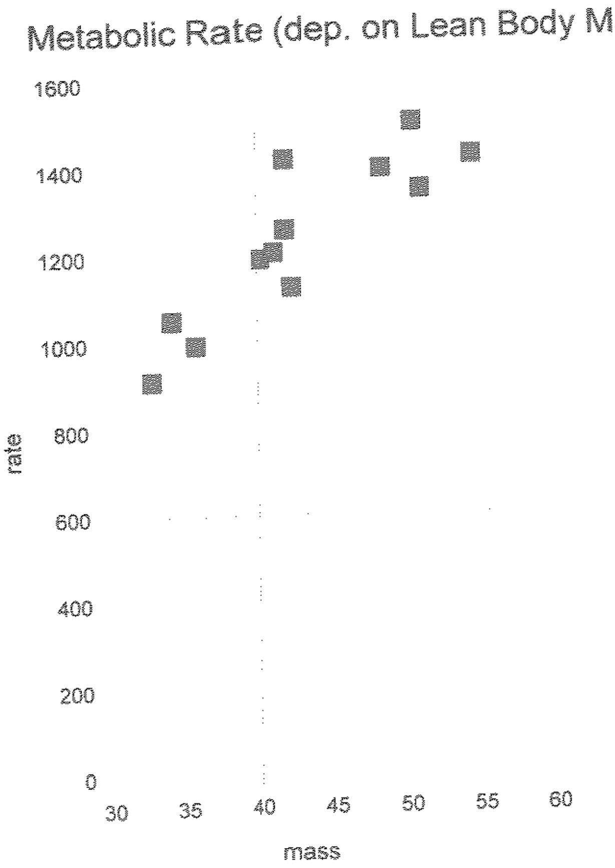
(a) Make a scatterplot that shows how metabolic rate depends on body mass. There is a quite strong linear relationship, with correlation $r = 0.876$.

(b) Find the least-squares regression line for predicting metabolic rate from body mass. Add this line to your scatterplot.

(c) Explain in words what the slope of the regression line tells us.

(d) Another woman has lean body mass 45 kilograms. What is her predicted metabolic rate?

(a)



Chp5 (cont)

Facts about least-squares regression

① Distinction between explanatory + response vars essential! (If we switch them, we get different result (line).)

② slope and correlation have same sign; slope = change of one s.d. in x corresponds to a change of r s.d.s in y.

③ least squares regression line always goes thru $(\bar{x}, \bar{y})$

④ r^2 = fraction of the variation in y-values that is explained by least-squares regression of y on x.

$$\boxed{r^2 = \frac{\text{variation in } \hat{y} \text{ as } x \text{ pulls it along the line}}{\text{total variation in observed values of } y.}}$$

* usefulness of linear regression depends on the strength of linear relationship; $r^2 = 1 \Rightarrow$ all of variation in y is accounted for by line; if $r = \pm 0.7$, $r^2 = 0.49 \Rightarrow$ about half of the variation is accounted for by line.

EX2

How useful is regression? Figure 4.8 (page 114) displays the relationship between golfers' scores on the first and second rounds of the 2007 Masters Tournament. The correlation is $r = 0.192$. Exercise 4.30 gives data on solar radiation (SRD) and concentration of dimethylsulfide (DMS) over a region of the Mediterranean. The correlation is $r = 0.969$. Explain in simple language why knowing only these correlations enables you to say that prediction of DMS from SRD by a regression line will be much more accurate than prediction of a golfer's second-round score from his first-round score.

Chp5 (cont)

Vocab

- residual = observed y-value − predicted y-value

  residual = $y - \hat{y}$.

  (prediction error)

  ★ correlation between residuals and x is always zero.

  ★ mean of least squares residuals = zero

- residual plot: scatterplot w/ y = residual values,

  x = explanatory variable values.

  - helps assess how well a regression line fits data

Ex 3 For data in Ex1, (a) find residuals for the data.

(b) check that the mean is zero.

(a)

mass (x)	rate (y)	Y-hat	residual
36.1	995		
54.6	1425		
48.5	1396		
42	1418		
50.6	1502		
42	1256		
40.3	1189		
33.1	913		
42.4	1124		
34.5	1052		
51.1	1347		
41.2	1204		

Chp5 (cont)

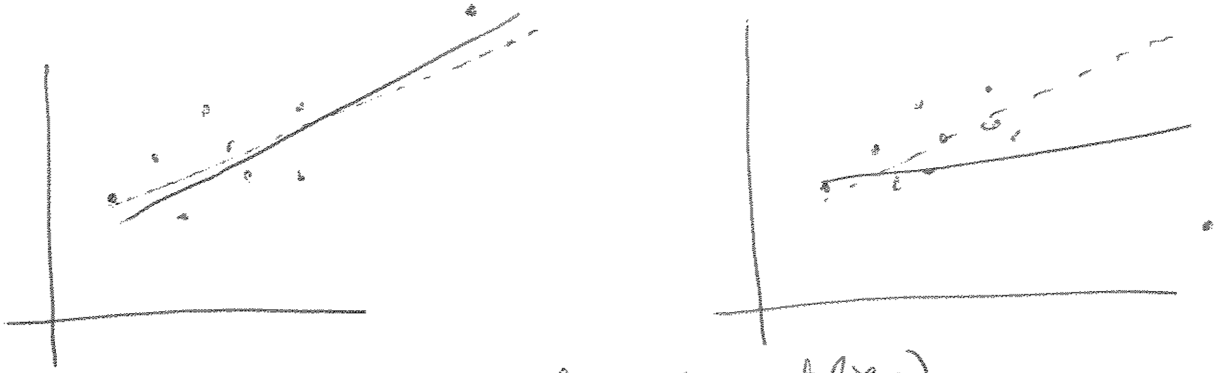
Vocab

- influential observation: an abservation that markedly changes a stat. calculation if it is removed from data set.

★ outliers for scatterplots are often influential (not always).

★ outliers in x-direction are often influential for least-squares regression line (but only if it's not close to the line) This type of outlier pulls least-squares line toward itself. (not all outliers are influential).

Ex 4



(dotted lines are best fit w/o outlier)
(solid lines " " w/ ")

which outlier is influential?

Notes:

- correlation + regression lines only describe linear relationships.
- correlation & regression lines are not resistant.
- few relationships are actually linear.
- beware of "lurking" variables
- extrapolation is not recommended for values far outside the range of collected data.

Chp 5 (cont)

- association (even if strong) is not causation.

Ex 5 (#48)

Regression to the mean. We expect that students who do well on the midterm exam in a course will usually also do well on the final exam. Gary Smith of Pomona College looked at the exam scores of all 346 students who took his statistics class over a 10-year period.[17] The least-squares line for predicting final exam score from midterm-exam score was $\hat{y} = 46.6 + 0.41x$. (Both exams have a 100-point scale.)

Octavio scores 10 points above the class mean on the midterm. How many points above the class mean do you predict that he will score on the final? (*Hint:* Use the fact that the least-squares line passes through the point $(\bar{x}, \bar{y})$ and the fact that Octavio's midterm score is $\bar{x} + 10$.) This is another example of regression to the mean: students who do well on the midterm will on the average do less well, but still above average, on the final.