Correlation & Regression, I

9.07

4/1/2004

Regression and correlation

- Involve bivariate, paired data, X & Y
  - Height & weight measured for the same individual
  - IQ & exam scores for each individual
  - Height of mother paired with height of daughter
- Sometimes more than two variables (W, X, Y, Z, …)

Regression & correlation

- Concerned with the questions:
  - Does a statistical relationship exist between X & Y, which allows some predictability of one of the variables from the other?
  - How strong is the apparent relationship, in the sense of predictive ability?
  - Can a simple linear rule be used to predict one variable from the other, and if so how good is this rule?
    - E.G. $Y = 5X + 6$

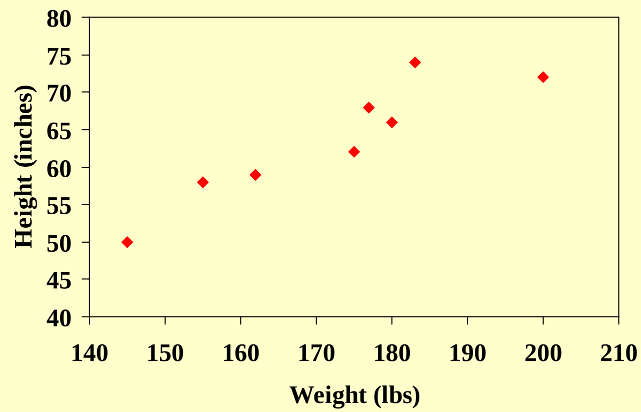
Regression vs. correlation

- Regression:
  - Predicting Y from X (or X from Y) by a linear rule
- Correlation:
  - How good is this relationship?

First tool: scatter plot

- For each pair of points, plot one member of a pair against the corresponding other member of that pair.
- In an experimental study, convention is to plot the independent variable on the x-axis, the dependent on the y-axis.
- Often we are describing the results of observational or "correlational" studies, in which case it doesn't matter which variable is on which axis.

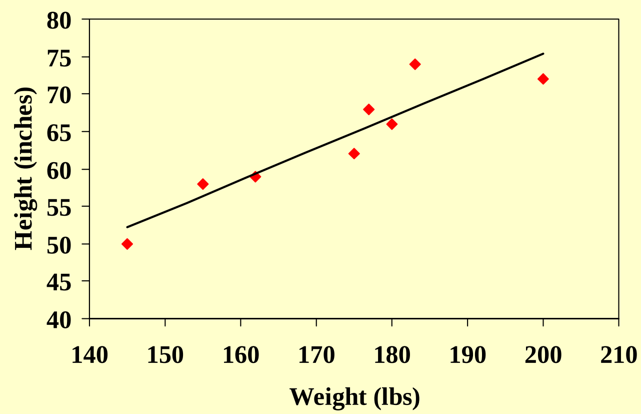
Scatter plot: height vs. weight



2nd tool: find the regression line

- We attempt to predict the values of y from the values of x, by fitting a straight line to the data
- The data probably doesn't fit on a straight line
  - Scatter
  - The relationship between x & y may not quite be linear (or it could be far from linear, in which case this technique isn't appropriate)
- The regression line is like a perfect version of what the linear relationship in the data would look like

Regression line

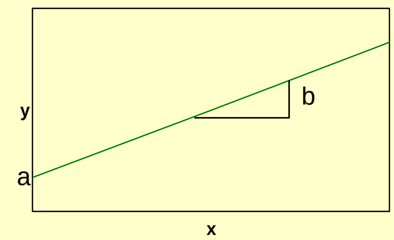


How do we find the regression line that best fits the data?

- We don't just sketch in something that looks good
- First, recall the equation for a line.
- Next, what do we mean by "best fit"?
- Finally, based upon that definition of "best fit," find the equation of the best fit line

Straight Line

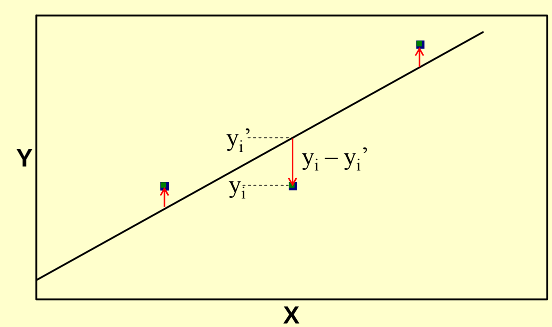
- General formula for any line is $y=bx+a$
- b is the slope of the line
- a is the intercept (i.e., the value of y when $x=0$)



Least-squares regression: What does "best fit" mean?

- If y_i is the true value of y paired with x_i, let y_i' = our prediction of y_i from x_i
- We want to minimize the error in our prediction of y over the full range of x
- We'll do this by minimizing
  $sse = \sum(y_i - y_i')^2$
- Express the formula as $y_i'=a+bx_i$
- We want to find the values of a and b that give us the least squared error, sse, thus this is called *"least-squares" regression*

Minimizing sum of squared errors



For fun, we're going to derive the equations for the best-fit a and b

- But first, some preliminary work:
  - Other forms of the variance
  - And the definition of covariance

A different form of the variance

- Recall:
- $\mathrm{var}(x) = E(x-\mu_x)^2$
  $= E(x^2 - 2x\mu_x + \mu_x^2)$
  $= E(x^2) - 2\mu_x^2 + \mu_x^2$
  $= E(x^2) - \mu_x^2$
  $= \Sigma x_i^2/N - (\Sigma x_i)^2/N^2$
  $= (\Sigma x_i^2 - (\Sigma x_i)^2/N) / N$

N-1 for unbiased estimate

- You may recognize this equation from the practise midterm (where it may have confused you).

The covariance

- We talked briefly about covariance a few lectures ago, when we talked about the variance of the difference of two random variables, when the random variables are not independent
- $\mathrm{var}(m_1 - m_2) =$
  $\sigma_1^2/n_1 + \sigma_2^2/n_2 - 2\,\mathrm{cov}(m_1, m_2)$

The covariance

- The covariance is a measure of how the x varies with y (co-variance = "varies with")
- $\mathrm{cov}(x, y) = E[(x-\mu_x)(y-\mu_y)]$
- $\mathrm{var}(x) = \mathrm{cov}(x, x)$
- Using algebra like that from two slides ago, we get an alternate form:
  $\mathrm{cov}(x, y) = E[(x-\mu_x)(y-\mu_y)]$
  $= E(xy - x\mu_y - y\mu_x + \mu_x \mu_y)$
  $= E(xy) - \mu_x \mu_y - \mu_x \mu_y + \mu_x \mu_y$
  $= E(xy) - \mu_x \mu_y$

OK, deriving the equations for a and b

- $y_i' = a + bx_i$
- We want the a and b that minimize
  $sse = \sum(y_i - y_i')^2 = \sum(y_i - a - bx_i)^2$
- Recall from calculus that to minimize this equation, we need to take derivatives and set them to zero.

Derivative with respect to a

$$\frac{\partial}{\partial a}\left(\sum (y_i - a - bx_i)^2\right) = -2\sum (y_i - a - bx_i) = 0$$

$$\Rightarrow \sum y_i - aN - b\sum x_i = 0$$

$$\Rightarrow a = \frac{\sum y_i}{N} - b\frac{\sum x_i}{N}$$

$$\Rightarrow a = \bar{y} - b\bar{x}$$

This is the equation for a, however it's still in terms of b.

Derivative with respect to b

$$\frac{\partial}{\partial b}\left(\sum (y_i - a - bx_i)^2\right) = -2\sum (y_i - a - bx_i)x_i = 0$$

$$\Rightarrow \sum x_i y_i - \sum (\bar{y} - b\bar{x})x_i - b\sum x_i^2 = 0$$

$$\Rightarrow \frac{1}{N}\sum x_i y_i - \frac{1}{N}\bar{y}\sum x_i + \frac{b}{N}\left(\bar{x}\sum x_i - \sum x_i^2\right) = 0$$

$$\Rightarrow \frac{1}{N}\sum x_i y_i - \bar{x}\bar{y} = b\left(\frac{1}{N}\sum x_i^2 - \bar{x}^2\right)$$

$$\Rightarrow b = \mathrm{cov}(x, y)/s_x^2$$

Least-squares regression equations

- $b = \mathrm{cov}(x, y)/s_x^2$
- $a = m_y - b\, m_x$
  ($\bar{x} = m_x$ Powerpoint doesn't make it easy to create a bar over a letter, so we'll go back to our old notation)
- Alternative notation:
  ss = "sum of squares"
  let $ss_{xx} = \Sigma(x_i - m_x)^2$
  $ss_{yy} = \Sigma(y_i - m_y)^2$
  $ss_{xy} = \Sigma(x_i - m_x)(y_i - m_y)$
  then $b = ss_{xy} / ss_{xx}$

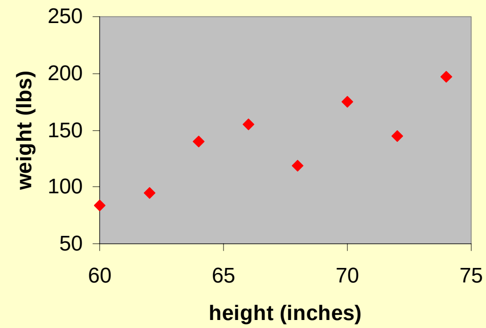
A typical question

- Can we predict the weight of a student if we are given their height?
- We need to create a regression equation relating the *outcome* variable, weight, to the *explanatory* variable, height.
- Start with the obligatory scatterplot

Example: predicting weight from height

x_i	y_i
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150

First, plot a scatter plot, and see if the relationship seems even remotely linear:



Looks ok.

Steps for computing the regression equation

- Compute m_x and m_y
- Compute $(x_i - m_x)$ and $(y_i - m_y)$
- Compute $(x_i - m_x)^2$ and $(x_i - m_x)(y_i - m_y)$
- Compute ss_{xx} and ss_{xy}
- $b=ss_{xy}/ss_{xx}$
- $a=m_y - bm_x$

Example: predicting weight from height

x_i	y_i
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150
Sum=612	1260
m_x=68	m_y=140

$ss_{xx}=240$ $ss_{yy}=10426$ $ss_{xy}=1200$

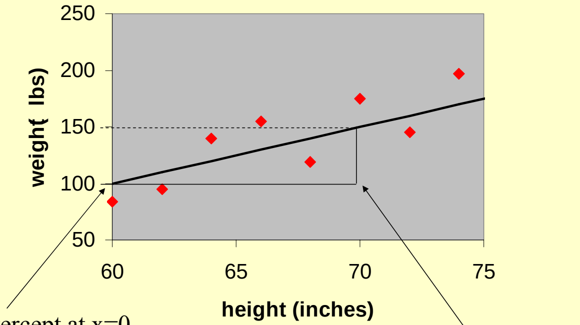
$b = ss_{xy}/ss_{xx} = 1200/240 = 5$; $a = m_y - bm_x = 140-5(68) = -200$

Example: predicting weight from height

x_i	y_i	(x_i-m_x)	(y_i-m_y)	$(x_i-m_x)^2$	$(y_i-m_y)^2$	$(x_i-m_x)(y_i-m_y)$
60	84	-8	-56	64	3136	448
62	95	-6	-45	36	2025	270
64	140	-4	0	16	0	0
66	155	-2	15	4	225	-30
68	119	0	-21	0	441	0
70	175	2	35	4	1225	70
72	145	4	5	16	25	20
74	197	6	57	36	3249	342
76	150	8	10	64	100	80
Sum=612	1260			ss_{xx}=240	ss_{yy}=10426	ss_{xy}=1200
m_x=68	m_y=140					

$b = ss_{xy}/ss_{xx} = 1200/240 = 5$; $a = m_y - bm_x = 140-5(68) = -200$

Plot the regression line

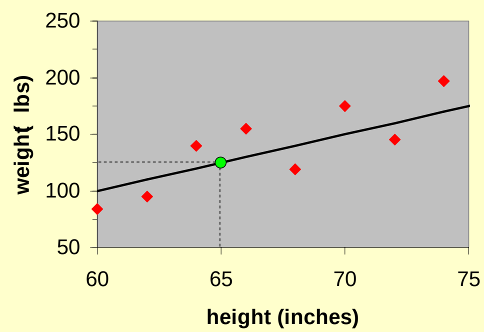


Intercept at x=0 is -200. "Intercept" at x=60 is -200+60*5 = 100

Slope = 5

What weight do we predict for someone who is 65" tall?

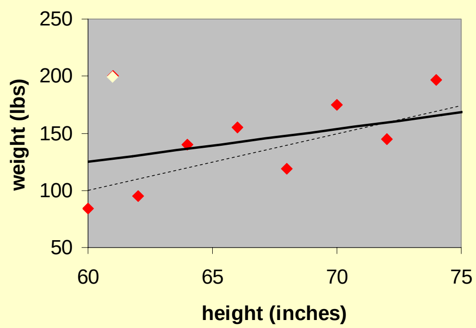
- Weight = -200 + 5*height = 125 lbs



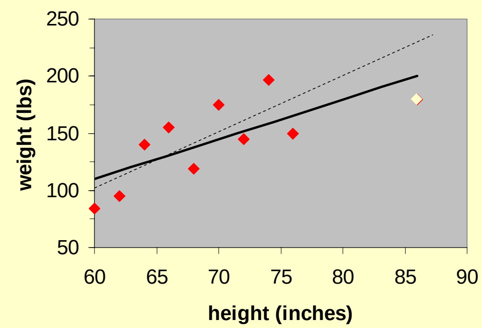
Caveats

- Outliers and influential observations can distort the equation
- Be careful with extrapolations beyond the data
- For every bivariate relationship there are two regression lines

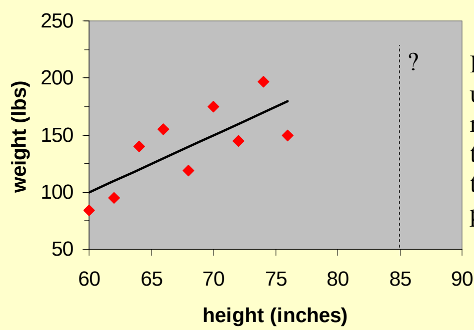
Effect of outliers



Effect of influential observations



Extrapolation

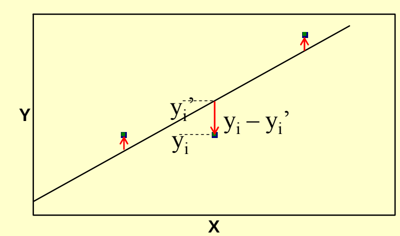


Be careful when using the linear regression eq'n to estimate, e.g., the weight of a person 85" tall.

The equation may only be a good fit within the x-range of your data.

Two regression lines

- Note that the definition of "best fit" that we used for least-squares regression was asymmetric with respect to x and y
  - It cared about error in y, but not error in x.



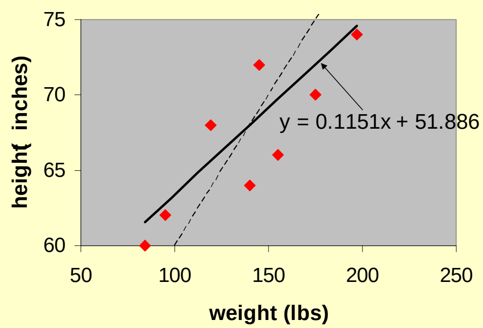
Two regression lines

- Note that the definition of "best fit" that we used for least-squares regression was asymmetric with respect to x and y
  - It cared about error in y, but not error in x.
  - Essentially, we were assuming that x was known (no error), we were trying to estimate y, and our y-values had some noise in them that kept the relationship from being perfectly linear.

Two regression lines

- But, in observational or correlational studies, the assignment of, e.g., weight to the y-axis, and height to the x-axis, is arbitrary.
- We could just as easily have tried to predict height from weight.
- If we do this, in general we will get a different regression line when we predict x from y than when we predict y from x.

Swapping height and weight

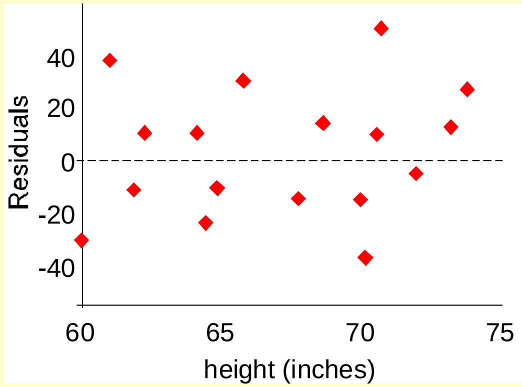


height ≈ 0.11 · weight + 51.89
weight = 5 · height - 200

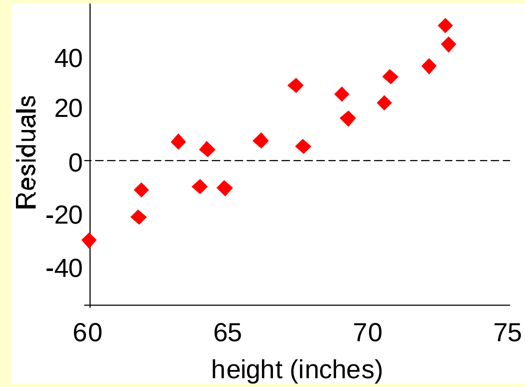
Residual Plots

- Plotting the residuals ($y_i - y_i'$) against x_i can reveal how well the linear equation explains the data
- Can suggest that the relationship is significantly non-linear, or other oddities
- The best structure to see is no structure at all

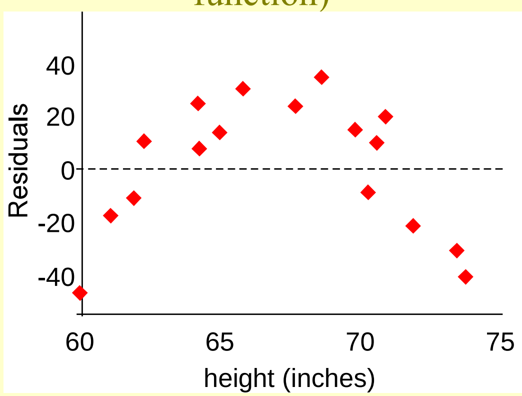
What we like to see: no pattern



If it looks like this, you did something wrong – there's still a linear component!



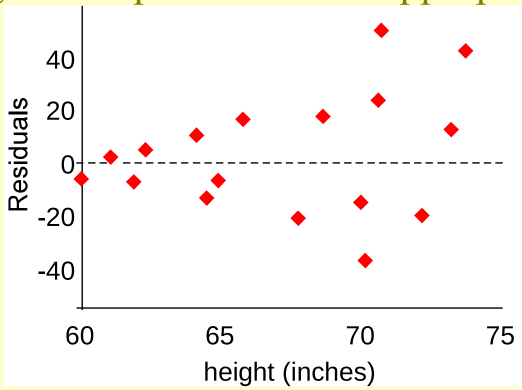
If there's a pattern, it was inappropriate to fit a line (instead of some other function)



What to do if a linear function isn't appropriate

- Often you can transform the data so that it is linear, and then fit the transformed data.
- This is equivalent to fitting the data with a model, y' = M(x), then plotting y vs. y' and fitting that with a linear model.
- This is outside of the scope of this class.

If it looks like this, again the regression procedure is inappropriate



Heteroscedastic data

- Data for which the amount of scatter depends upon the x-value (vs. "homoscedastic", where it doesn't depend on x)
- Leads to residual plots like that on the previous slide
- Happens a lot in behavioral research because of Weber's law.
  - As people how much of an increment in sound volume they can just distinguish from a standard volume
  - How big a difference is required (and how much variability there is in the result) depends upon the standard volume
- Can often deal with this problem by transforming the data, or doing a modified, "weighted" regression
- (Again, outside of the scope of this class.)

Coming up next…

- The regression fallacy
- Assumptions implicit in regression
- Confidence intervals on the parameters of the regression line
- Confidence intervals on the predicted value y', given x
- Correlation