Regression *M&M §2.3 and §10*

Uses

- Curve fitting
- Summarization ('model')
- Description
- Prediction
- Explanation
- Adjustment for 'confounding' variables

Technical Meaning

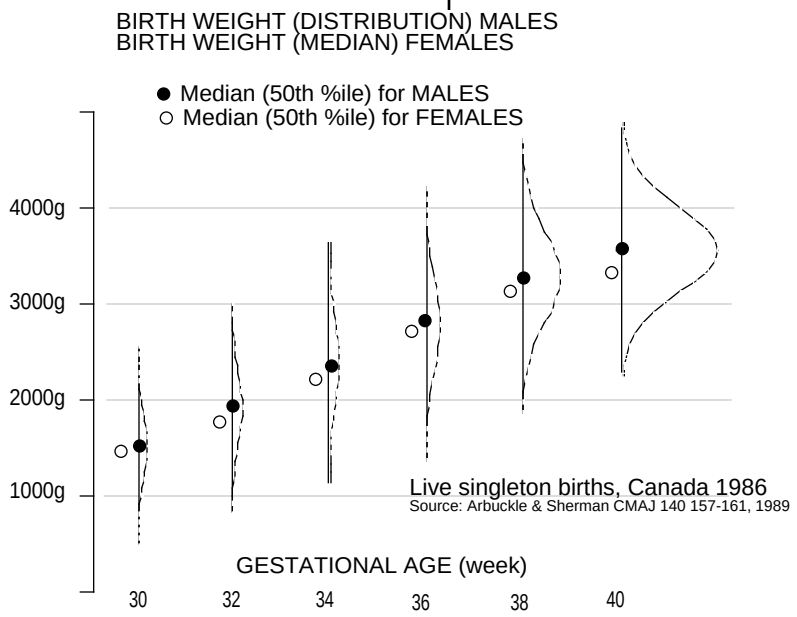
- [originally] simply a line of 'best fit' to data points
- [nowadays] Regression line is the LINE that connects the CENTRES of the distributions of Y's at each X value.
- not necessarily a straight line; could be curved, as with growth charts
- not necessarily $\mu_{Y|X}$'s used as CENTRES ; could use medians etc.
- strictly speaking, haven't completed description unless we characterize the variation around the centres of the Y distributions at each X
- inference not restricted to the distributions of Y's for which we make some observations; it applies to distributions of Y's at all unobserved X values in between.

Examples (with appropriate caveats)

- Birth weight (Y) in relation to gestational age (X)
- Blood pressure (Y) in relation to age (X)
- Cardiovascular mortality (Y) in relation to water hardness (X) ?
- Cancer incidence (Y) in relation to some exposure (X) ?
- Scholastic performance (Y) vis a vis amount of TV watched (X)

Caveat: No guarantee that simple straight line relationship will be adequate. Also, in some instances the relationship might change with the type of X and Y variables used to measure the two phenomena being studied; also the relationship may be more artifact than real - see later for inference.)

	MALES				FEMALES			
Age.	*Tot.*	%-ile; weight,g			*Tot.*	%-ile; weight,g		
wk	*N.*	10th	50th	90th	*No.*	10th	50th	90th
25	*100*	651	810	950	*73*	604	750	924
30	*257*	1 156	1 530	2 214	*216*	1 040	1 485	2 001
31								
32								
33								
34								
35	*1 840*	2 060	2 570	3 140	*1 454*	1 950	2 460	3 040
36								
37								
38								
39								
40	*68 102*	3 020	3 570	4 160	*67 149*	2 900	3 430	4 000
41								
42	*10 309*	3 200	3 770	4 390	*9 636*	3 060	3 610	4 190



Simple Linear† Regression
(one X) (straight line)

Equation

- $\mu_{Y|X} = \alpha + \beta X$ or $\frac{\Delta \mu_{Y|X}}{\Delta X} = \beta = \frac{\text{"rise"}}{\text{"run"}}$

In Practice:
one rarely sees an exact straight line relationship in health science applications;

1 - While physicists are often able to examine the relationship between Y and X in a laboratory with all other things being equal (ie controlled or held constant) medical investigators largely are not. The universe of (X,Y) pairs is very large and any 'true' relationship is disturbed by countless uncontrollable (and sometimes un-measurable factors. In any particular sample of (X,Y) pairs these distortions will surely be operating.
2 - The true relationship (even if we could measure it exactly) may not be a simple straight line.
3 - The measuring instruments may be faulty or inexact (using 'instruments' in the broadest sense).

One always tries to have the investigation sufficiently controlled that the 'real' relationship won't be 'swamped' by factors 1 and 3 and that the background "noise" will be small enough so that alternative models (eg curvilinear relationships) can be distinguished from one another.

---

† Linear here means linear in the parameters. The equation $y = Bx^C$ can be made linear in the parameters by taking logs i.e. $\log[y] = \log[B] + x \log[C]$; $y = a+b\cdot x+c\cdot x^2$ is already linear in the parameters a b and c. The following model cannot be made linear in the parameters α β γ:

$$\text{proportion dying} = \alpha + \frac{1-\alpha}{1+\exp\{\beta - \gamma \log[\text{dose}]\}}$$

Fitting a straight line to data - Least Squares Method

The most common method is that of Least Squares. Note that least squares can be thought of as just a curve fitting method and doesn't have to be thought of in a statistical (or random variation or sampling variation) context. Other more statistically-oriented methods include the method of minimum Chi-Square (matching observed and expected counts according to measure of discrepancy) and the Method of Maximum likelihood (finding the parameters that made the data most likely). Each has a different criterion of "best-fit".

Least Squares Approach:

- Consider a candidate slope (b) and intercept (a) and predict that the Y value accompanying any X=x is $\hat{y} = a + b\cdot x$. The observed y value will deviate from this "predicted" or "fitted" value by an amount $d = y - \hat{y}$
  We wish to keep this deviation as small as possible, but we must try to strike a balance over all the data points. Again just like when calculating variances, it is easier to work with squared deviations[1] :
  $$d^2 = (y - \hat{y})^2$$
  We weight all deviations equally (whether they be the ones in the middle or the extremes of the x range) using $\sum d^2 = \sum (y - \hat{y})^2$ to measure the overall (or average) discrepancy of the points from the line.

[1] there are also several theoretical advantages to least squares estimates over others based for example on least absolute deviations: - they are the most precise of all the possible estimates one could get by taking linear combinations of y's.

- From all the possible candidates for slope (b) and intercept (a) , we choose the particular values a and b which make this sum of squares (sum of squared deviations of 'fitted' from 'observed' Y's) a minimum. ie we search for the a and b that give us the least squares fit.
- Fortunately, we don't have to use trial and error to arrive at the 'best' a and b . Instead, it can be shown by calculus or algebraically that the a and b which minimize $\sum d^2$ are:

$$b = \hat{\beta} = \frac{\sum\{x_i - \bar{x}\}\{y_i - \bar{y}\}}{\sum\{x_i - \bar{x}\}^2} = \frac{r_{xy} \bullet s_y}{s_x}$$

$$a = \hat{\alpha} = \bar{y} - b\ \bar{x}$$

[Note that a least-squares fit of the regression line of X on Y would give a different set of values for the slope and intercept: the slope of the line of x on y is $\frac{r_{xy} \bullet s_x}{s_y}$] . one needs to be careful when using a calculator or computer program to specify which is the explanatory variable (X) and which is the predicted variable (Y)].

Meaning of intercept parameter (a):

Unlike the slope parameter (which represents the increase/decrease in $\mu_{Y|X}$ for every unit increase in x), the intercept does not always have a 'natural' interpretation. It depends on where the x-values lie in relation to x=0, and may represent part of what is really the mean Y. For example, the regression line for fuel economy of cars (Y) in relation to their weight (x) might be

$$\mu_{Y|weight} = 60 \text{ mpg} - 0.01 \bullet \text{weight in lbs} \quad [0.01 \text{ mpg/lb}]$$

but there are no cars weighing 0 lbs. It would be better to write the equation in relation to some 'central' value for weight e.g. 3500 lbs; then the same equation can be cast as

$$\mu_{Y|weight} - 25 = 0.01 \bullet (\text{weight} - 3500)$$

It is helpful for testing whether there is evidence of a non-zero slope to think of the simplest of all regression models, namely that which is a horizontal straight line

$$\mu_{Y|X} = \alpha + 0 \bullet X = \text{the constant } \alpha .$$

This is a re-statement of the fact that the sum of squared deviances around a constant horizontal line at height 'a' is smallest when 'α' = the mean .

[We don't always use the mean as the best 'centre' of a set of numbers. Imagine waiting for one of several elevators with doors in a row along one wall; you do not know which one will arrive next, and so want to stand in the 'best' place no matter which one comes next. Where to stand depends on the criterion being optimized: if you want to minimize the maximum distance, stand in the middle between the one on the extreme left and the extreme right; if you wish to minimize the average distance , where do you stand?, If, for some reason, you want to minimize the average squared distance, where to stand? If the elevator doors are not equally spaced from each other, what then?]

The anatomy of a slope: some re-expressions

Consider the formula: slope = b = $\frac{\sum\{x - xbar\}\{y - ybar\}}{\sum\{x - xbar\}^2}$

Without loss of generality & for simplicity, assume ybar=0.

If we have 3 x's, 1 unit apart (e.g. x_1=1; x_2 =2; x_3 =3),

then... $x_1 - xbar = -1$; $x_2 - xbar = 0$; $x_3 - xbar = +1$

so slope = b = $\frac{\{-1\}y_1 + \{0\}y_2 + \{+1\}y_3}{\{-1\}^2 + \{0\}^2 + \{+1\}^2}$

i.e. slope = $\frac{y_3 - y_1}{x_3 - x_1}$

Note that y_2 contributes to ybar and thus to an estimate of the average y (i.e. level) but not to the slope.

If 4 x's 1 unit apart (e.g. x_1=1; x_2 =2; x_3 =3; x_4 =4), then,...

$x_1 - xbar = -1.5$ $\quad$ $x_2 - xbar = -0.5$
$x_3 - xbar = +0.5$ $\quad$ $x_4 - xbar = +1.5$

and so

slope = b = $\frac{\{-1.5\}y_1 + \{-0.5\}y_2 + \{+0.5\}y_3 + \{+1.5\}y_4}{\{-1.5\}^2 + \{-0.5\}^2 + \{0.5\}^2 + \{+1.5\}^2}$

i.e. slope = $\frac{1.5\{y_4 - y_1\}}{5} + \frac{0.5\{y_3 - y_2\}}{5}$

i.e. slope = $\frac{\frac{3}{2}\{y_4 - y_1\}}{\frac{5}{3}\{x_4 - x_1\}} + \frac{\frac{1}{2}\{y_3 - y_2\}}{\frac{5}{1}\{x_3 - x_2\}}$

i.e. slope = $\frac{9}{10}\frac{\{y_4 - y_1\}}{\{x_4 - x_1\}} + \frac{1}{10}\frac{\{y_3 - y_2\}}{\{x_3 - x_2\}}$

i.e. a weighted average of the slope from datapoints 1 and 4 and that from datapoints 2 and 3, with weights proportional to the squares of their distances on x axis $\{x_4 - x_1\}^2$ and $\{x_3 - x_2\}^2$

Another way to think of the slope:

Rewrite b = $\frac{\sum\{x - xbar\}\{y - ybar\}}{\sum\{x - xbar\}^2}$ as

$$b = \frac{\sum\{x - xbar\}^2 \frac{\{y - ybar\}}{\{x - xbar\}}}{\sum\{x - xbar\}^2} = \frac{\sum weight \frac{\{y - ybar\}}{\{x - xbar\}}}{\sum weight}$$

weight $\{x - xbar\}^2$ for estimate $\frac{\{y - ybar\}}{\{x - xbar\}}$ of slope

Yet another way to think of the slope:

b is a weighted average of all the pairwise slopes $\frac{y_i - y_j}{x_i - x_j}$ with weights proportional to $\{x_i - x_j\}^2$.

e.g. **If 4 x's 1 unit apart**

denote by $b_{1\&2}$ the slope obtained from $\{x_2,y_2\}$ & $\{x_1,y_1\}$, etc...

$$b = \frac{1.b_{1\&2} + 4.b_{1\&3} + 9.b_{1\&4} + 1.b_{1\&3} + 4.b_{2\&4} + 1.b_{3\&4}}{1 + 4 + 9 + 1 + 4 + 1 = 20}$$

jh 6/94

Inferences regarding Simple Linear Regression

How reliable are

(i) the (estimated) slope
(ii) the (estimated) intercept
(ii) the predicted mean Y at a given X
(iv) the predicted y for a (future) individual with a given X

when they are based on data from a sample? i.e. how much would these estimated quantities change if they were based on a different random sample [with the same x values]?

We can use the concept of sampling variation to (i) describe the 'uncertainty' in our estimates via CONFIDENCE INTERVALS or (ii) carry out TESTS of significance on the parameters (slope, intercept, predicted mean).

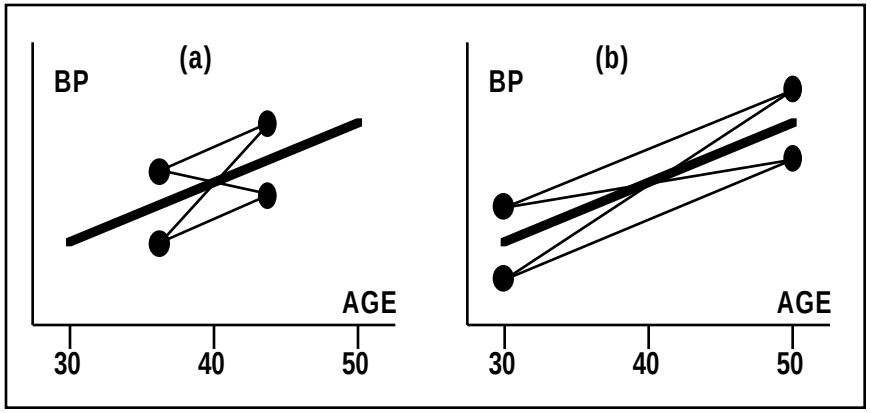
We can describe the degree of reliability of (or, conversely, the degree of uncertainty in) an estimated quantity by the standard deviation of the possible estimates produced by different random samples of the same size from the same x's. We call this (obviously conceptual) S.D. the standard error of the estimated quantity (just like the standard error of the mean when estimating μ). helpful to think of slope as an average difference in means for 2 groups that are 1 x-unit apart.

The size of the standard error will depend on

1. how 'spread apart' the x's are
2. How good a fit the regression line really is (i.e. how small is the unexplained variation about the line)
3. How large the sample size, n, is.

Factors affecting reliability (in more detail)

1. The spread of the X's: The best way to get a reliable estimate of the slope is to take Y readings at X's that are quite a distance from each other. E.g. in estimating the "per year increase in BP over the 30-50 yr. age range", it would be better to take X=30,35, 40, 45, 50 than to take X =38, 39, 49, 41, 42. Any individual fluctuations will 'throw off' the slope much less if the X's are far apart.



thick line: real (true) relation between average BP at age X and X : *thin lines:* possible apparent relationships because of individual variation when we study 1 individual at each of two ages (a) spaced closer together (b) spaced further apart.

Notes

Regression line refers to the relationship between the average Y at a given X to the X , and not to individual Y's vs X. Obviously of course if the individual Y's are close to the average Y, so much the better!

The above argument would suggest studying individuals at the extremes of the X values of interest. We do this if we are sure that the relationship is a linear one. If we are not sure, it is wiser -- if we have a choice in the matter -- to take a 3-point distribution.

There is a common misapprehension that a Gaussian distribution of X values is desirable for estimating a regression slope of Y on X. In fact, the 'inverted U' shape of the Gaussian is the least desirable!

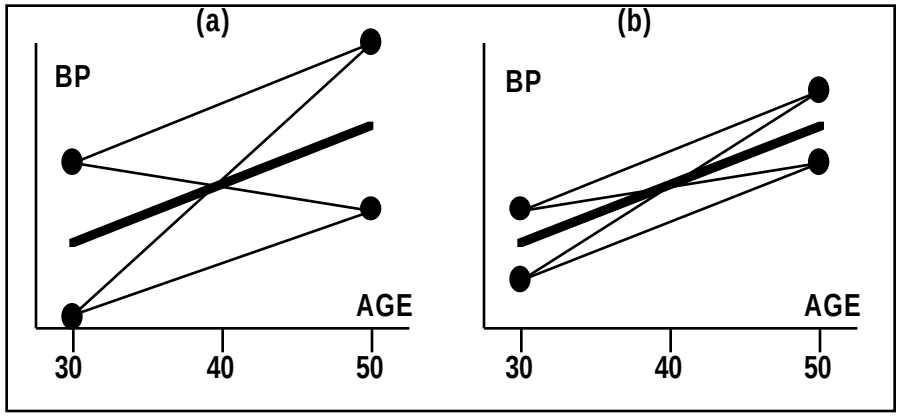
Factors affecting reliability (continued)

2. The (vertical) variation about the regression line: Again, consider BP and age, and suppose that indeed the average BP of all persons aged X + 1 is β units higher than the average BP of all persons aged X, and that this linear relationship

   average BP of persons aged x = $\alpha + \beta \cdot X$
   (average of Y's at a given x = intercept + slope • X)

   holds over the age span 30-50.

   Obviously, everybody aged x=32 won't have the exact same BP, some will be above the average of 32 yr olds, some below. Likewise for the different ages x=30,...50. In other words, at any x there will be a distribution of y's about the average for age X. Obviously, how wide this distribution is about $\alpha + \beta \cdot X$ will have an effect on what slopes one could find in different samples (measure vertical spread around the line by σ)



thick line: real (true) relation between average BP at age X and X :
thin lines: possible apparent relationships because of individual variation when we study 1 individual at each of two ages when the within-age distributions have (a) a narrow spread (b) a wider spread

NOTE: For unweighted regression, should have roughly same spread of Y's at each X.

Factors affecting reliability (continued)

3. Sample Size (n) Larger n will make it more difficult for the types of extremes and misleading estimates caused by 1) poor X spread and 2) large variation in Y about $\mu_{Y|X}$, to occur. Clearly, it may be possible to spread the x's out so as to maximize their variance (and thus reduce the n required) but it may not be possible to change the magnitude of the variation about $\mu_{Y|X}$ (unless there are other known factors influencing BP). Thus the need for reasonably stable estimated $\hat{y}$ [i.e.estimate of $\mu_{Y|X}$]

Standard Errors

$$SE(b) = SE(\hat{\beta}) = \frac{\sigma}{\sqrt{\sum\{x_i - \bar{x}\}^2}} \;;$$

$$SE(a) = SE(\hat{\alpha}) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum\{x_i - \bar{x}\}^2}}$$

(Note: there is a negative correlation between a and b).

We don't usually know σ so we estimate it from the data, using scatter of the y's from the fitted line i.e. SD of the residuals)

If examine the structure of SE(b), see that it reflects the 3 factors discussed above: (i) a large spread of the x's makes contribution of each observation to $\sum\{x_i - \bar{x}\}^2$ large, and since this is in the denominator, it reduces the SE (ii) a small vertical scatter is reflected in a small σ and since this is in the numerator, it also reduces the SE of the estimated slope (iii) a large sample size means that $\sum\{x_i - \bar{x}\}^2$ is larger, and like (i) this reduces the SE.

The formula, as written, tends to hide this last factor; note that $\sum\{x_i - \bar{x}\}^2$ is what we use to compute the spread of a set of x's -- we simply divide it by n–1 to get a variance and then take the square root to get the sd. To make the point here, simplify n-1 to n and write

$$\sum\{x_i - \bar{x}\}^2 \approx n \cdot var(x), \text{ so that } \sqrt{\sum\{x_i - \bar{x}\}^2} \approx \sqrt{n} \cdot sd(x)$$

and the equation for the SE simplifies to

$$SE(b) = \frac{\sigma}{\sqrt{n} \cdot sd(x)} = \frac{SD_{y|x} \,/\, SD_x}{\sqrt{n}}$$

with $\sqrt{n}$ in its familiar place in the denominator of the SE (even in more complex SE's, this is where $\sqrt{n}$ is usually found !)

The structure of SE(a) : In addition to the factors mentioned above, all of which come in again in the expected way, there is the additional factor of $\bar{x}^2$; since this is in the denominator, it increases the SE . This is natural in that if the data, and thus $\bar{x}$, are far from x=0, then any imprecision in the estimate of the slope will project backwards to a large imprecision in the estimated intercept. Also, if one uses 'centered' x's, so that $\bar{x} = 0$, the formula for the SE reduces to

$$SE(a) = \sigma\sqrt{\frac{1}{n}} = \frac{\sigma}{\sqrt{n}}$$

and we recognize this as $SE(\bar{y})$ -- not surprisingly, since $\bar{y}$ is the 'intercept' for centered data.

CI's & Tests of Significance for $\hat{\alpha}$ and $\hat{\beta}$ are based on t–distribution (or Gaussian Z's if n large)

α : $\quad \hat{\alpha} \pm t_{n-2} \bullet SE(\hat{\alpha})$

$$\mathbf{H_0:}\ \alpha = \alpha_0 \qquad t_{n-2} = \frac{\hat{\alpha} - \alpha_0}{SE(\hat{\alpha})}$$

β : $\quad \hat{\beta} \pm t_{n-2} \bullet SE(\hat{\beta})$

$$\mathbf{H_0:}\ \beta = \beta_0 \qquad t_{n-2} = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})}$$

Standard Error for Estimated $\mu_{Y|X}$ or 'average Y at X'

We estimate '**average Y at X' or** $\mu_{Y|X}$ by $\hat{\alpha} + \hat{\beta} \cdot X$. Since the estimate is based on two estimated quantities, each of which is subject to sampling variation, it contains the uncertainty of both:

$$\text{SE(estimated average Y at X)} = \sigma \sqrt{\frac{1}{n} + \frac{\{X - \bar{x}\}^2}{\sum\{x_i - \bar{x}\}^2}}$$

Again, we must use an estimate $\hat{\sigma}$ of σ .

First-time users of this formula suspect that it has a missing $\sum$ or an x instead of an xbar or something. There is no typographical error, and indeed if one examines it closely, it makes sense. **X** refers to the x-value at which one is estimating the mean -- it has nothing to do with the actual x's in the study which generated the estimated coefficients, except that the closer **X** is to the center of the data, the smaller the quantity $\{X - \bar{x}\}$ and thus the quantity $\{X - \bar{x}\}^2$, and thus the SE, will be. Indeed, if we estimate the **average Y** *right at* $X = \bar{x}$, the estimate is simply $\bar{y}$ (since the fitted line goes through $[\bar{x}, \bar{y}]$) and its SE will be

$$\sigma \sqrt{\frac{1}{n} + \frac{\{\bar{x} - \bar{x}\}^2}{\sum\{x_i - \bar{x}\}^2}} \quad \text{or} \quad \sigma \sqrt{\frac{1}{n}} = \frac{\sigma}{\sqrt{n}} = \text{SE}(\bar{y}).$$

Confidence Interval for individual Y at X

A certain percentage P% of individuals are within $t_P \cdot \sigma$ of the mean $\mu_{Y|X} = \alpha + \beta \cdot X$, where t_P is a multiple, depending on P, from the t or, if n is large, the Z table. However, we are not quite certain where exactly the mean $\alpha + \beta \cdot X$ is -- the best we can do is estimate, with a certain P% confidence, that it is within $t_P \cdot \text{SE}(\hat{\alpha} + \hat{\beta} \cdot X)$ of the point estimate $\hat{\alpha} + \hat{\beta} \cdot X$. The uncertainty concerning the mean and the natural variation of individuals around the mean -- wherever it is -- combine in the expression for the estimated P% range of individual variation, which is as follows:

$$\hat{\alpha} + \hat{\beta} \cdot X \pm t \cdot \sigma \sqrt{1 + \frac{1}{n} + \frac{\{X - \bar{x}\}^2}{\sum\{x_i - \bar{x}\}^2}} \ .$$

Both the CI for the estimated mean and the CI for individuals (ie the estimated percentiles of the distribution) are bow-shaped when drawn as a function of **X**. They are narrowest at $X = \bar{x}$, and fan out from there. One needs to be careful not to confuse the much narrower CI for the mean with the much wider CI for individuals. If one can see the raw data, it is usually obvious which is which -- the CI for individuals is almost as wide as the raw data themselves.

cf. data on sleeping through the night; alcohol levels and eye speed.