1.1 No. Clerical errors cause the relationship in the observed (i.e. recorded) data to be inexact.

1.2 Functional

1.3 Chemists and other basic scientists often have this attitude! But it is difficult to imagine a perfectly homogeneous source of plastic, and that the items would all have been molded in exactly the same way, and that the measurements would not contain some little amount of error.. I'll bet that if even the same item was measured twice in close succession by two different machines, one might get slightly different hardness values.

1.4 Random variation (the ε portion). I don't like to see these e's referred to as errors; I prefer to speak of (unexplained) variations around a mean.

1.5 No. this is the model for an individual realization of Y. In the mean of all possible Y's the ε's would cancel each other out [that's the meaning of the ε's)

1.6 b) β_1 is the increase in the mean Y for a 1 unit difference in the value of X. Several of you simply said the increase in THE value of Y -- that doesn't make sense if Y contains a random component that will vary for each realization.

1.7 a) clearly no, because there are many possible distributions with $\sigma = 5$.
b) yes, because now the question is about the proportion of a Gaussian(200,5) distribution that falls in the range 195-205, or equivalently what proportion of a standardized Gaussian distribution falls within 1 standard deviation of the mean -- which we know to be about 2/3rds or 68%.

OUT OF INTEREST -- and linking to your courses on mathematical statistics -- If one knows μ and σ, but not the shape of the distribution, can one put any bounds on the fraction of the distribution will be within 1 (or 2? or 3?) standard deviation(s) of the mean? Hint: Remember an inequality associated with someone whose name starts with Tch, like Tchaikovsky (and has about the same number of letters)

1.8 mean or expectation yes, realization no.

1.10 Since the data are crossectional (and not longitudinal) we need to be very careful. The economy and the workforce are in some flux. The over 50's may have come into the company recently, as part of a second career and with less experience than some of the younger employees. Or, maybe, less competent programmers may be kept on, but in low paying jobs, while the more competent ones move up to management! What do you think Bill Gates' salary would be if he stayed as a programmer?

1.12 These are non-experimental data, and those who choose to exercise more are probably different in many health-related ways from those who exercise less. Or, it may be that ill-health has prevented some from exercising as much a they would like! So one would need to know about these differences with respect to other factors and try to form fair comparisons -- the surest way to achieve this

is if the investigator has -- by matching and randomization -- the freedom to form otherwise comparable groups who exercise more or less.

1.16 To derive the LS estimators, one -- like Legendre in 1805 -- needs only calculus. See pp 19-. . To form the ML estimators, one needs to know the mathematical form of the distribution of the ε's i.e. one needs a mathematical statistics course! (see pp 34-) . The LS procedure is *purely mathematical*, whereas the ML procedure is *statistical.*

1.29 The E[Y|X] line goes through the origin.

1.30 The E[Y|X] line is independent of X (i.e. parallel to X-axis .. the slope is 0)

1.33 Solve $\partial\Sigma(y - \beta_0)^2 / \partial\beta_0 = 0$ for β_0. Check that the second derivative is positive (it is 2n)

1.34 $E(\Sigma y_i/n) = (1/n)\Sigma E(y_i) = (1/n)n\beta_0 = \beta_0$

1.37 Agree. If you knew a lot about the Y's at X=29, and a lot about those at X=31, then -- *provided the relationship between E{Y|X} is linear,* and the SD doesn't vary very much, one would -- by interpolation -- know a lot about the Ys at X=30!

Note the assumptions ("the model") in italics.. there is no free lunch!

This ***"borrowing strength from adjacent observations"*** is at the very heart of what statistical models are all about! See an application of this principle in "Borrowing Strength using regression" under resources/materials for session 3 in the material for course 678. (Figs 1 and 2 give the background, Fig 3 the rationale, and Fig 4 the result).

2 See article by Hanley and Lippman.

3 British study: distribution of weight and stature (height) of 4,995 women. The slope of the linear regression of weight on height was 2.7. What are the plausible units for the 2.7?

As per article "Formula for 'Ideal Weight' " -- in material/resources for session 6 for course 678 -- one rule for ideal weight is: [mean of actual weights probably greater than this]

women 100lbs + 5lbs/inch for every inch > 60inches;

men 106 lbs + 6lbs/inch for every inch > 60inches.

For women, that is a slope of

5lbs/inch = 2.27Kg/inch = 2.27Kg/2.54cm = 0.89Kg/cm.

So maybe the measured *weight is in metric* and *height in imperial units*!! That, together with the fact that we are studying actual weights, rather than discussing

someone's idea of ideal weight, would make the slightly higher value of 2.7Kg/inch quite plausible.

To get some individual data on this, I went to two datasets in (`http://www.epi.mcgill.ca/hanley/c678/`)

If I read my output correctly, the summaries *at age 18* from the (small and quite old) *Berkeley* dataset are:

females	ave wt 61 Kg	ave ht 167cm	slope 0.47Kg/cm
males	ave wt 72 Kg	ave ht 180cm	slope 0.33Kg/cm

0.47 Kg/cm = 1.03lb/cm = 1.03/0.4 = 2.6 lb/inch

-- not far from the 2.7 in the British study !!

Q: if studying 48 or 58 year olds, rather than18 year olds, then should add something for weight gain with age.. one rule of thumb I have heard: some physicians say they "expect" (they didn't say "it is good" for!) men to "put on about a pound a year after about age 25!!

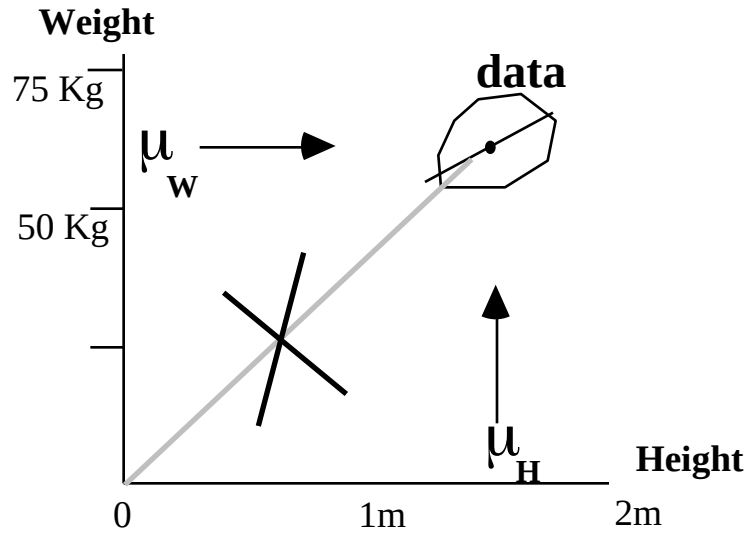
Q: Incidentally, what's wrong with using the ∂ in means to get a slope of

(72-61)/(180-167) = 11/13 = 0.85Kg/cm??

A: It assumes that slope is same for men and women!

Q: what's wrong with using 61/167 = 0.365Kg/cm as a slope

A: It assumes that slope, projected down from (ave height, ave weight) would go through origin!! One should not invent (or assume) data points outside the range under study.



If the model $E[Y|X] = \beta_1 X$ (zero intercept) did indeed hold, then $\bar{y} / \bar{x}$ is an unbiased (but -- depending on the structure for the ε's -- inefficient) estimator of β_1 . See exercise 1.41 and the exercise on models for weights of sheets of paper.

In the *bodyfat* dataset from over 200 males average age 45, age range 22 to 81, the summaries are

ave wt 178lb ave ht 70inches slope 2.47lb/inch

So maybe the imperial 2.7lb/inch in women is not that unreasonable!

I did notice that the height range was **2**9.5 inches to 77.75 inches, and the weightrange 118.5 lbs to 363.15 lls, so Iwonder if there aren't some mistakes in the bodyfat dataset.

Rolf Heinmueller (course 697-Fall 1999) provides us with a nice easy to remember "continental-European" rule of thumb that works well for both males and females:

normal weight(Kg) = height in cm minus 100

Weight[Kg] = **-100**[Kg] + **1**[Kg/cm] × height

1 Kg/cm = 2.2 lb/cm = 2.2/0.4 = 5.5 lb/inch --- half way between the gender-specific slopes for North American ideal weights!!

Q Does the "continental-European" model concern "normal" in sense of *"actual averages in population"* rather than in sense of *"best for health"* or *"ideal"*?
Check for yourself if it is more "generous" than the North American one: calculate what it says is "normal" for someone 60 inches (152.4cm) tall [1Kg=2.2lb].