University of Waterloo W. H. Cherry

REGRESSION: An Introduction – Overview of the Ideas

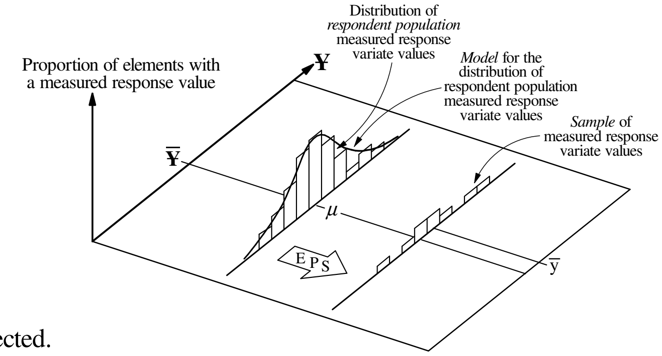
To appreciate the development of statistical methods in this Highlight #52, we first review ideas (for instance, from Part 12 of the STAT 221 Course Materials) about inference for one and two populations.

- Remember that UPPER-case **bold** letters refer to a *population*; the 'cross' (a horizontal line across a letter) enables us to distinguish, in *handwritten* material, a population quantity from a random variable represented by an upper case *italic* letter.
  - A *value* of a random variable is represented by a *lower* case italic letter; a data value (or a specified value like a sample size) is represented by a lower case Roman letter.

These notation conventions assist in maintaining the population/sample and real world/model distinctions.

1. Case 1: *Describing* One Population

As indicated pictorially in the 3-dimensional diagram at the right, we think of a (*respondent*) *population* as having a *distribution* of values for some measured *response* variate (**Y**) of interest; this distribution can be displayed as a *histogram*. Population *attributes* of interest are:



∗ Size: **N** [the number of *elements*];
∗ Location: $\overline{\mathbf{Y}}$ [the *average*];
∗ Variation: **S** [the (*data*) *standard deviation*];

N is usually considered to be:
○ known, and
○ very large compared to the size (n) of any *sample* (to be) selected.

The *study* population is defined by *specifying* values for one or more *explanatory* variate(s); the *respondent* population is those study population elements that *would* provide the data requested under the incentives for response provided in the investigation.

- Usually, *many* explanatory variates affect the measured value of the response variate of an element; dealing with explanatory variates becomes essential in the Plan when *relationship(s)* are being investigated (*e.g.*, when two or more populations are compared to answer a Question with a *causative* aspect) but is usually *less* important when investigating *one* population to answer a Question with a *descriptive* aspect.

STAT 221 Figures 12.1, 12.7, 12.10, 12.13, 12.15 and 12.22 discuss formal inference about *one* population – they use:
⊙ a test of significance, or: ⊙ a confidence interval,
to obtain an *Answer* based on an estimate obtained in the investigation for the value of a respondent population attribute, typically:
∗ an *average* (mainly), or: ∗ a *standard deviation* (in STAT 221 Figure 12.13 only);
[interest in standard deviation for its *own* sake is more common in *industrial* problem solving concerned with reducing variation]. We used statistical methods of data *analysis* to obtain this Answer because we had only *in*complete information – data from a *sample* rather than from a *census* of the respondent population. Statistical methods allowed us to *quantify* the uncertainty arising from:
+ *sample* error and: **+** *measurement* error and to assess: **+** *model* error,
but, as *we* developed them, they took *no* formal account of *other* sources of uncertainty in Answer(s):
– *study* error, **–** *non-response* error (a notable instance of the effect of *missing data*).

Our inference from a *sample* about $\overline{\mathbf{Y}}$ or **S** is based on a *model*, most commonly in these Statistical Highlights an assumed normal distribution for the *population* response **Y**; we write equation (HL52.1), where the Y_j are *probabilistically independent*:

$$Y_j \sim N(\mu,\, \sigma), \quad j = 1, 2, \ldots, \mathrm{n}. \qquad \text{-----(HL52.1)}$$

∗ Y_j is a *random variable* whose distribution represents the possible values of the measured response variate **Y** for the jth unit in the *sample* of n units *selected* from the respondent population, if the selecting and measuring processes were to be repeated over and over;
∗ μ is a model *parameter* (the *mean*) representing $\overline{\mathbf{Y}}$, the *average* of the measured response variate of the elements of the respondent population.

Equation (HL52.1) has a *probability* emphasis; an equivalent, but statistically more convenient, form is equation (HL52.2), where the R_j (and, hence, the Y_j) are *probabilistically independent*:

$$Y_j = \mu + R_j, \quad R_j \sim N(0,\, \sigma), \quad j = 1, 2, \ldots, \mathrm{n}, \quad \text{independent}, \quad \text{EPS}. \qquad \text{-----(HL52.2)}$$

∗ R_j is a random variable (called the *residual*) whose distribution represents the possible *differences*, from the structural component of the model, of the measured value of the response variate **Y** for the jth unit in the sample of n units selected from the respondent population, if the selecting and measuring processes were to be repeated over and over;

(*continued overleaf*)

∗ σ the (*probabilistic*) *standard deviation* of the normal model for the distribution of the residual, is a model *parameter* which represents **Ꞩ**, the (data) *standard deviation* of the measured response variate **Ɏ** of the elements of the respondent population; this standard deviation (and, hence, σ) *quantifies* the *variation* of the measured response variate over the elements of the respondent population – as this variation increases, so does the respondent population standard deviation (and, hence, so does σ).

Note the steps leading to this *normal* probability model for Y_j: the (discrete) population *histogram* for **Ɏ** has a shape that is approximated by a (continuous) normal p.d.f. and, under *equiprobable* selecting (EPS) [for which each of the **Ꞥ** respondent population elements has the *same* inclusion probability], this normality for the *population* carries through into the model for the random variable Y_j in the *sample*. [Elsewhere, EPS may be referred to as *simple random* selecting (SRS).]

- The normal model for **Ɏ** is the basis of t distribution estimating procedures (*e.g.*, in Figure 12.15), where μ representing $\overline{\text{Ɏ}}$ is estimated by $\bar{y}$ (taken as the *sample* average $\bar{y}$) and σ representing **Ꞩ** is estimated by s [taken as the *sample* (data) s.d. s].
  - If a value is *given* in the Question statement for **Ꞩ** (to be used as the value of σ), we can relax the requirement for a normal model for the distribution of **Ɏ**; the *approximate* normal distribution of the relevant estimator of μ representing $\overline{\text{Ɏ}}$ is then a consequence of the Central Limit Theorem. However, this case is of little importance in practice because **Ꞩ** is *seldom* known when estimating μ representing $\overline{\text{Ɏ}}$.
    - ○ The pictorial representation of this case would be like the diagram overleaf at the upper right of page HL52.1 but with*out* the (normal) model for the distribution of respondent population response values (and, hence, without label μ).
  - Because we encounter probabilistic statements mainly in relation to the *sample* and its attributes, it is easy to forget such statements are a *consequence* of our model for the *respondent population* response, coupled with EPS of the sample.

Thus, the most notable features of our models are the *distribution* of **Ɏ** in the respondent population and the *processes* of *selecting the sample* and *measuring variate values* (see Statistical Highlights #6, #18, #21 and #38); we also recognize the importance of:

+ adequate *access* to an appropriate *study* population to manage study error;
+ a high *response rate* to manage non-response error for the *respondent* population];
+ assessing *model error.*

NOTES: 1. The diagram overleaf at the upper right of page HL52.1 reminds us of *sample error* and *measurement error* – the *difference* between the measured value of $\overline{\text{Ɏ}}$ (the respondent population average) and $\bar{y}$ (the measured *sample* average) in a *particular* investigation. In the diagram, the combination of these two errors is *known* because $\overline{\text{Ɏ}}$ is known but, in real investigating, we *only* know $\bar{y}$; statistical methods allow us to quantify the sampling and measuring uncertainty in our *estimate* of $\overline{\text{Ɏ}}$ – see Statistical Highlight #74.

2. An investigation involving one population is usually trying to answer a Question(s) with a *descriptive* aspect – such investigating typically seeks to describe the state of affairs in some population, often in terms of estimates for the values of one or more population attributes; there may also be *comparison* among the Answers from *different* investigations of descriptive Questions – for example, comparison among the Answers from a series of political polls taken over time – but this does not alter the descriptive nature of the Question in each investigation.

   By contrast, most investigations of *two* populations are inherently *comparative*, sometimes *with* explicit concern for *causation*, sometimes with*out* it or with this concern only *im*plicit.

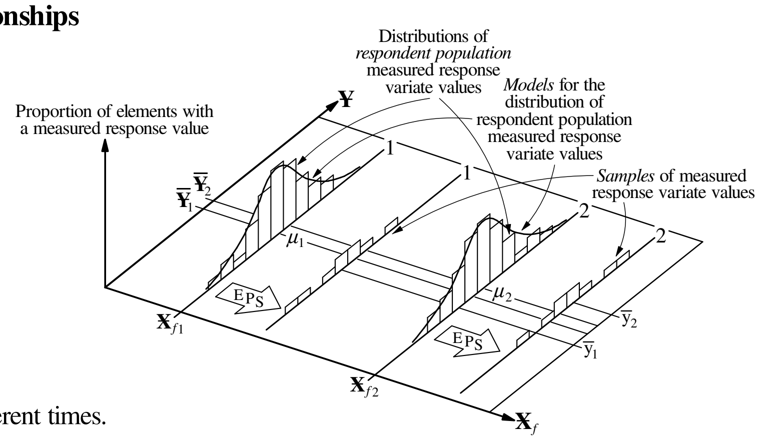
   - When a comparative Plan is concerned with estimating an attribute involving a *difference*, there may be some cancellation of *measuring* inaccuracy.

2. Case 2: *Comparing* Two Populations – Relationships

Our picture of *two* populations is, in essence, a duplication of the case for *one* population, as indicated by the 3-dimensional diagram at the right. Population attributes of interest are again:

∗ Size: **Ꞥ**$_1$ and **Ꞥ**$_2$,
∗ Location: $\overline{\text{Ɏ}}_1$ and $\overline{\text{Ɏ}}_2$,
∗ Variation: **Ꞩ**$_1$ and **Ꞩ**$_2$,

except that a subscript (1 or 2) is added to distinguish the two respondent populations. The *new* idea is the use of the value of a *focal explanatory variate* (here denoted **X**$_f$) to specify what distinguishes one respondent population from the other.



- The two populations may be the *same* elements at different times.

STAT 221 Figures 12.11 and 12.19 discuss formal inference about *two* populations – these Figures use:

⊙ a test of significance, or: ⊙ a confidence interval,

to obtain an *Answer* based on estimates obtained in the investigation for the values of respondent population attributes, typically:

University of Waterloo W. H. Cherry

REGRESSION: An Introduction – Overview of the Ideas (continued 1)

∗ two *averages* (mainly), or: ∗ two *standard deviations* (in Cases 2 and 3 of STAT 221 Figure 12.19 only);
[our concern with standard deviations was only *secondary* to our primary interest in averages]. We used statistical methods of data *analysis* to obtain this Answer because we had only *in*complete information – data from *samples* rather than from *censuses* of the two respondent populations. Statistical methods allowed us to *quantify* the uncertainty arising from:
+ *sample* error and: **+** *measurement* error and to assess: **+** *model* error,
but, as *we* developed them, they again took *no* formal account of *other* sources of uncertainty in Answer(s):
– *study* error, **–** *non-response* error (the most notable instance of the effect of *missing data*).

Our inference from the *samples* about $\bar{\mathbf{Y}}_1 - \bar{\mathbf{Y}}_2$ (or about $\mathbf{S}_1/\mathbf{S}_2$) is based on an extension of the model, written as equations (HL52.1) and (HL52.2) on pageHL52.1, to *two* populations – in the expressions (HL52.3) and (HL52.4) at at the right, subscript i = 1 or 2 indexes the relevant respondent population and the sample from it, where the Y_{ij} are *probabilistically independent*:

$$Y_{ij} \sim N(\mu_i,\ \sigma_i),\quad j = 1, 2, ..., n_i. \qquad \text{-----(HL52.3)}$$

$$Y_{ij} = \mu_i + R_{ij},\quad R_{ij} \sim N(0,\ \sigma_i),\quad i = 1, 2,\quad j = 1, 2, ..., n_i,\quad \text{independent},\quad \text{EPS}. \qquad \text{-----(HL52.4)}$$

∗ Y_{ij} is a random variable whose distribution represents the possible values of the measured response variate $\mathbf{Y}_i$ for the jth unit in the *sample* of n_i units selected from respondent population i, if the selecting and measuring processes were to be repeated over and over;

∗ μ_i is a model *parameter* (the *mean*) representing the *average* of the measured response variate $\mathbf{Y}_i$ for the elements of respondent population i;

∗ R_{ij} is a *random variable* (called the *residual*) whose distribution represents the possible *differences*, from the structural component of the model, of the measured value of the response variate $\mathbf{Y}_i$ for the jth unit in the sample of n_i units selected from respondent population i, if the selecting and measuring processes were to be repeated over and over;

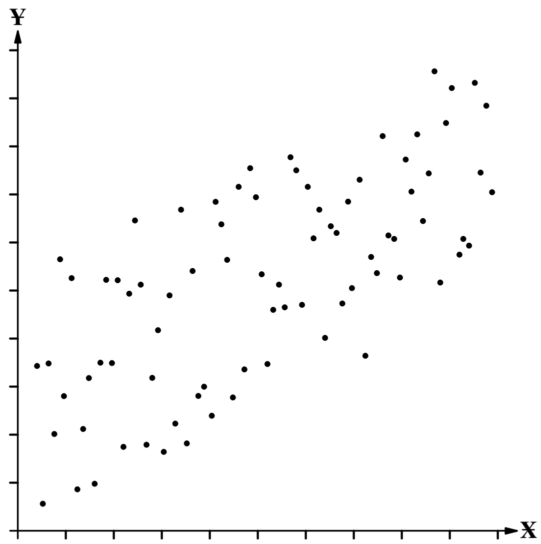
∗ σ_i the (*probabilistic*) *standard deviation* of the normal model for the distribution of the residual, is a model *parameter* representing $\mathbf{S}_i$, the (*data*) *standard deviation* of the measured response variate $\mathbf{Y}_i$ of the elementsts of respondent population i; this standard deviation (and, hence, σ_i) quantifies the *variation* of the measured response variate over the elements of respondent population i – as this variation increases, so does each standard deviation (and, hence, so does σ_i).

Note again the steps leading to this *normal* probability model for Y_{ij}: the (discrete) population *histograms* for $\mathbf{Y}_1$ and $\mathbf{Y}_2$ have a shape that is approximated by a (continuous) normal p.d.f. and, under *equiprobable* selecting, this normality for the *populations* carries through into the models for the random variables Y_{ij} in the *samples.*

3. Case 3: *Relationships* – Many Populations

Comparing two populations in Case 2 is one instance of investigating a *relationship* – whether there is an effect on a measured response variate $\mathbf{Y}$ of (two values of) a focal explanatory variate $\mathbf{X}_f$. In Case 3, our concern is still with relationships, but from a more general perspective.

We think first of *one* population of $\mathbf{N}$ elements, each with a value for a response variate $\mathbf{Y}$ and an explanatory variate $\mathbf{X}$; such a data set $(\mathbf{X}_i, \mathbf{Y}_i)$, i = 1, 2, ..., $\mathbf{N}$, would usually be shown graphically as a scatter diagram like the one at the right. The Question we want to answer is: *What is the* form *of the relationship between* $\mathbf{Y}$ *and* $\mathbf{X}$ *in the respondent population*?



One answer to this question, for the $\mathbf{N}$ $(\mathbf{X}, \mathbf{Y})$ values, is $\mathbf{Y} = f(\mathbf{X})$, where $f(\mathbf{X})$ is a polynomial of degree $\mathbf{N}$ passing through *all* the $\mathbf{N}$ points (which are assumed to have *distinct* $\mathbf{X}$ values); such a model for the form of the relationship contains $\mathbf{N}+1$ parameters. Depending on the positions of the points (*i.e.*, depending on the elements' $\mathbf{X}$ and $\mathbf{Y}$ values), a polynomial of *lower* degree may also pass through all the points but, in general, as the degree *de*creases, more and more of the points will lie *off* the curve representing the polynomial. *Our* concern *here* is with using a *straight line* (a *first*-degree polynomial with *two* parameters) to model the linear component of the *trend* in the points; this straight line *may* pass through one or more of the points or through *none* of them.

There is only *one* polynomial of degree $\mathbf{N}$, specified by a unique set of values of its $\mathbf{N}+1$ parameters, which passes through all $\mathbf{N}$ points, but there is *more* than one line which could be thought of as representing the straight-line component of their

(*continued overleaf*)

trend. The criterion we use to select *one* of these lines is called *least squares*: the *line* of *best fit* to the **N** points is the one which *minimizes the sum of the squares of the vertical distances of the points from the line* (*i.e.*, the one which minimizes the sum of the squared *residuals*). Such a line is often referred to, for historical reasons, as *the regression of* **Y** *on* **X**; its two parameters $\mathbf{B}_0$ (the **Y**-intercept) and $\mathbf{B}_1$ (the slope) are *attributes* of the population.

An illustration of the matters described in the previous paragraph is shown at the right. There are five (**X**, **Y**) values:

X	1	3	6	10	15
Y	3	5	5	7	10,

and the fourth-degree polynomial which passes through the five points is:

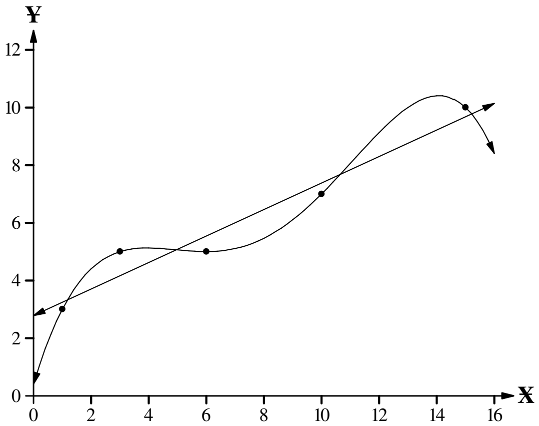
$$7560\mathbf{Y} = 3060 + 25236\mathbf{X} - 6205\mathbf{X}^2 + 608\mathbf{X}^3 - 19\mathbf{X}^4$$

[or: $\mathbf{Y} \simeq 0.4048 + 3.3381\mathbf{X} - 0.8208\mathbf{X}^2 + 0.08042\mathbf{X}^3 - 0.002513\mathbf{X}^4$];

also, as shown in STAT 221 Figure 13.3, the regression of **Y** on **X** is:

$${}_{\text{reg}}\overline{\mathbf{Y}} = 2.7778 + 0.4603\mathbf{X}.$$

The diagram at the right shows a scatter diagram of the five (**X**, **Y**) values with the polynomial and the straight line of best fit superimposed on it.



The discussion so far in Case 3 is in terms of a *population* of **N** (**X**, **Y**) values but the more usual data set in practice is a *sample* of n (x, y) values, where the sample size n is generally much *smaller* than the population size **N**. As explained in more detail in STAT 221 Figure 13.3, the modification to the discussion above is that, for *sample* data, the method of least squares yields only *estimates*, b_0 and b_1, of the *model parameters* β_0 and β_1 representing the population attributes $\mathbf{B}_0$ and $\mathbf{B}_1$; *we* also refer to the line as the *estimated* regression of **Y** on **X** (although 'estimated' is generally omitted elsewhere). As always, the *method* of selecting the sample is *central* to assessing the likely size of the error associated with inference(s) from the sample about the relationship in the respondent population; one statistically acceptable method is, of course, *equiprobable* selecting.

- Formal statistical inference (interval estimating, significance testing) from a sample about the straight-line component of the relationship in the respondent population usually requires *distributional* assumptions about the population; for simple linear regression, the common *normal* model is the expressions (HL52.5) and (HL52.6) at the right, where the Y_j (and the R_j) are *probabilistically independent*:

$$Y_j \sim N(\beta_0 + \beta_1 x_j, \sigma), \quad j = 1, 2, \ldots.., n. \qquad \text{-----(HL52.5)}$$

$$Y_j = \beta_0 + \beta_1 x_j + R_j, \quad R_j \sim N(0, \sigma), \quad j = 1, 2, \ldots.., n. \quad \text{independent, EPS.} \qquad \text{-----(HL52.6)}$$

  - ∗ Y_j is a random variable whose distribution represents the possible values of the measured response variate **Y** for the jth unit in the sample of n units selected from the respondent population, if the selecting and measuring processes were to be repeated over and over;
  - ∗ β_0 is a model parameter which represents the **Y**-*intercept* of the straight-line model for the relationship between **X** and the average of **Y** measured in the respondent population; *i.e.*, the ordinate of this straight line when **X** = 0;
  - ∗ β_1 is a model parameter which represents the *slope* of the straight-line model for the relationship between **X** and the average of **Y** measured in the respondent population; *i.e.*, the change in the measured average of **Y** for unit change in **X** over the elements of the respondent population;
  - ∗ x_j is the value of the explanatory variate **X** for the jth unit in the sample of n units selected from the respondent population;
  - ∗ R_j is a random variable (called the *residual*) whose distribution represents the possible *differences*, from the structural component of the model, of the value of the measured response variate **Y** for the jth unit in the sample of n units selected from the respondent population, if the selecting and measuring processes were to be repeated over and over;
  - ∗ σ the (probabilistic) *standard deviation* of the normal model for the distribution of the residual, is a model parameter representing the (data) *standard deviation* of the measured response variate **Y** of the elements of the respondent population with value x_j for explanatory variate **X**; this standard deviation (and, hence, σ) quantifies the *variation* of the measured response variate of the elements of the respondent population with value x_j for explanatory variate **X** – as this variation increases, so does σ.

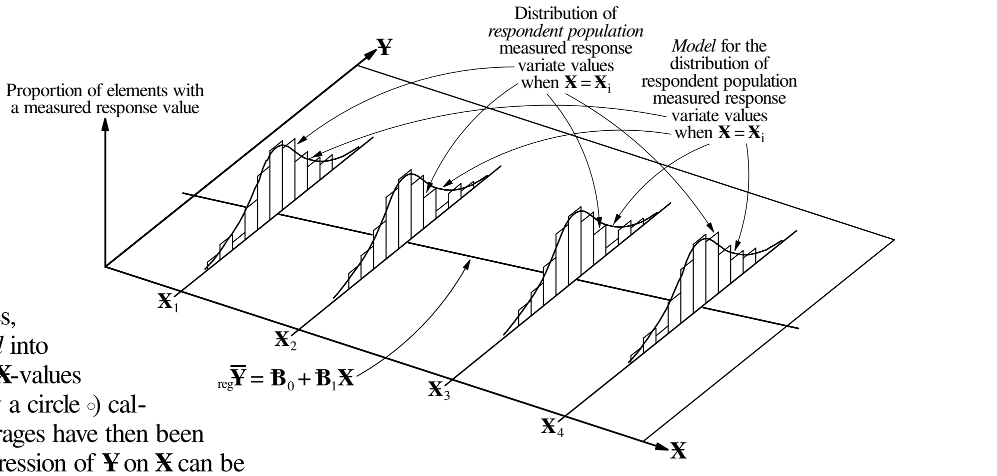
This model somewhat *changes* our original picture of *one* population of **N** (**X**, **Y**) values – there is now *conceptually* a normally distributed 'population' of **Y** values at *any* **X** value $\mathbf{X}_i$. This idea is shown pictorially in the diagram at the top of the facing page HL52.5 which, for convenience, is limited to *four* **X** values; the diagram shows how the *model* for discussing relationships in Case 3 makes this case *appear* to be an extension of Case 2 to *many* populations, one for *each* value of **X**.

- This diagram differs from the two earlier diagrams (on the first and second sides of this Highlight #52, pages HL52.1 and HL52.2) in that it shows *only* the populations and *omits* the samples; if it *were* to include the latter, there would typically be only *four* observations – one from each 'population' at the four **X** values.
- *Un*like Case 1 and Case 2, there may be situations in Case 3 where the population histograms, implied by the model of

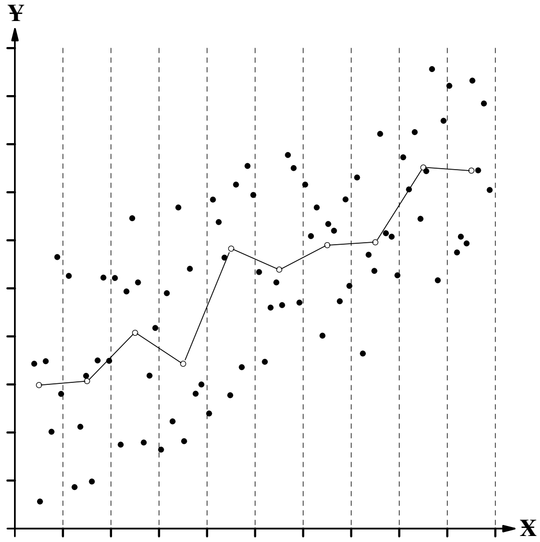
University of Waterloo W. H. Cherry

REGRESSION: An Introduction – Overview of the Ideas (continued 2)

equations (HL52.5) and (HL52.6) on the facing page HL52.4 for Y_j, do not actually exist – there may be only *one* (or a *few*) values of $\mathbf{Y}$ when $\mathbf{X} = \mathbf{X}_i$.



- The model assumption that the trend in the $\overline{\mathbf{Y}}_i$s is *linear* in $\mathbf{X}$ can be illustrated as shown in the diagram at the lower right; the scatter diagram is the one given on page HL52.3 but, as indicated by the dashed vertical lines, the points have now been *grouped* into ten intervals on the basis of their $\mathbf{X}$-values and the *average* of $\mathbf{Y}$ (indicated by a circle ○) calculated for each group; these averages have then been been connected by lines. The regression of $\mathbf{Y}$ on $\mathbf{X}$ can be thought of an *idealization* of this roughly ('untidy') linear trend in these *averages*.



○ A notational difficulty is that $\overline{\mathbf{Y}}_i$ is the natural symbol for the *actual* average of $\mathbf{Y}$ when $\mathbf{X} = \mathbf{X}_i$, but we also need a symbol for the *hypothetical* average of $\mathbf{Y}$ under the idealization of the straight-line regression model. In these Statistical Highlights and Course Materials (but likely nowhere else), *we* use ${}_{\text{reg}}\overline{\mathbf{Y}}_i(\mathbf{X}_i)$ where:

- the **bar** reminds us we are dealing with an *average* response;
- the subscript **reg** reminds us this average is a *model* quantity;
- the $\mathbf{X}_i$ reminds us the model average is a (straight-line) function of the value of the explanatory variate $\mathbf{X}$;
  [When there is no ambiguity, $(\mathbf{X}_i)$ can be omitted and the population model average ordinate shortened to ${}_{\text{reg}}\overline{\mathbf{Y}}_i$].

When *probability* modelling is introduced (as in STAT 221 Figure 13.7) to deal with *inferences* based on *sample* data, the symbol for this average ordinate is $\mu_Y(x_j)$, which follows naturally from ${}_{\text{reg}}\overline{\mathbf{Y}}_i(\mathbf{X}_i)$; also, *we* write the *equation* of the straight line fitted by the method of least squares as ${}_{\text{reg}}\overline{\mathbf{Y}} = \mathbf{B}_0 + \mathbf{B}_1\mathbf{X}$ or, for our illustrative population data set of 5 observations, as ${}_{\text{reg}}\overline{\mathbf{Y}} = 2.7778 + 0.4603\mathbf{X}$. We need to remember from the start that a regression model deals with an *average* response, although it may take time for us to get used to this change to our long-standing *mathematical* description of a straight line as $y = b + mx$.

It is the concern with *averges* that distinguishes most sharply regression modelling from the idea of a polynomial *passing through all the observations*; we are no longer concerned with this latter idea in this Highlight #52. The two approaches to fitting equations to points have different intents and outcomes.

- As in Cases 1 and 2, note the steps leading to the *normal* probability model for Y_j: the (discrete) population *histogram* for $\mathbf{Y}$ at each value of $\mathbf{X}$ has a shape that is approximated by a (continuous) normal p.d.f. and, under *equiprobable* selecting, this normality for the *populations* carries through into the models for the random variables Y_j for each value x_j of $\mathbf{X}$ in the *sample*.

NOTES: 3. The four diagrams on this and the previous two pages HL52.3 and HL52.4 show the case of a *direct* relationship beship between $\mathbf{Y}$ and $\mathbf{X}$ – both variables tend to increase or decrease together (*i.e.*, $\mathbf{B}_1 > 0$); of course, *inverse* relationships, where one variable tends to *in*crease as the other *de*creases (*i.e.*, $\mathbf{B}_1 < 0$) are also important in practice.

4. A straight line fitted to *bi*variate data (or, equivalently, the values of its two parameters) can be regarded as a useful data *summary*, in the same way an *average* is a useful summary of *uni*variate data.

5. One purpose of fitting a straight line or other function to a set of bivariate data may be to *extrapolate* beyond the interval of the observed $\mathbf{X}$ values to 'predict' a value for (the average of) $\mathbf{Y}$. The diagram at the upper right of the facing page HL52.4 reminds us of the importance of choosing a 'sensible' model and of the dangers of extrapolating;

- the points in this diagram show an increasing roughly linear trend, but we have no *empirical* information

NOTES: (cont.) 5. about whether this form of relationship between $\mathbf{Y}$ and $\mathbf{X}$ persists *out*side the range of the data (*i.e.*, $\mathbf{X}$ values outside the interval from about 0 to 16);

- the 'predicted' $\mathbf{Y}$ value for $\mathbf{X}$ *out*side the interval 0-16 would differ *substantially* between the *straight line* of best fit (which passes through *none* of the points but follows their *trend*) and the fourth degree polynomial (which passes through *all* the points but departs sharply from their trend ouside the interval of observation) – in the context of this Highlight #52, the polynomial is introduced *only* for illustrative purposes and our *real* concern is with the straight line.

6. If there is interest in whether a change in $\mathbf{X}$ *causes* a change in $\mathbf{Y}$, the accuracy of the inference depends on whether the data come from an *experimental* Plan [in which the *investigator(s)* determine the $\mathbf{X}$ values and then measure $\mathbf{Y}$] or from an *observational* Plan [in which the $\mathbf{Y}$ values arise from the values of $\mathbf{X}$ that occur in the units that happen to be investigated].

7. Note the sometimes confusing terminology used in connection with the topic of 'regression' introduced in Case 3 and the phrase *simple linear regression*.
   - *Regression*, in its statistical useage, means a process for finding a mathematical form for the relationship between (the average of) a measured *response* variate and one or more *explanatory* variates; like all mathematical models, regression *models* are only an *approximate* description of what is being modelled.
   - *Simple* refers to the *straight-line* regression model with *one* explanatory variate: ${}_{\text{reg}}\overline{\mathbf{Y}} = \beta_0 + \beta_1\mathbf{X}$, which is also *the (model) regression of* $\mathbf{Y}$ on $\mathbf{X}$; *simple* is to be contrasted with *multiple*.
   - *Linear* refers to a regression model being linear in the *parameters* – it does *not* mean linear in the explanatory variate(s), so that: ${}_{\text{reg}}\overline{\mathbf{Y}} = \beta_0 + \beta_1\mathbf{X} + \beta_2\mathbf{X}^2$ *is* linear; linear is to be contrasted with *non-linear*.
   - *Multiple* regression models have *two or more* explanatory variates (on their right-hand side) – for example: ${}_{\text{reg}}\overline{\mathbf{Y}} = \beta_0 + \beta_1\mathbf{X}_1 + \beta_2\mathbf{X}_2$.
   - *Non-linear* refers to a regression model that is *not* linear in the *parameters* – for instance:
     ${}_{\text{reg}}\overline{\mathbf{Y}} = \beta_1\mathbf{X}^{\alpha}$ is non-linear (because of α) but can be made linear by a logarithmic transformation;
     ${}_{\text{reg}}\overline{\mathbf{Y}} = \beta_0 + \beta_1\mathbf{X}^{\alpha}$ is (inherently) non-linear.

8. Although we limit our discussion largely to fitting *straight lines* to data in these Statistical Highlights and Course Materials, the method of least squares can also be used to fit a second or higher degree polynomial (or a more complicated model) to a set of bi- (or multi-)variate data.

9. Information about two matters of historical interest in Case 3, taken from page 45 of an established 'regression' text (*Applied Regression Analysis*, Norman R. Draper and Harry Smith, Third Edition, John Wiley & Sons, New York, 1998), is as follows:

> It appears that Sir Francis Galton (1822-1911), a well-known British anthropologist and meteorologist, was responsible for the introduction of the term *regression*. He originally used the word *reversion* in an unpublished address *Typical laws of heredity in man* to the Royal Institution on February 9, 1877. The later word *regression* appears in his Presidential address to Section H of the British Association at Aberdeen, 1885, printed in *Nature*, pages 507-510, September, 1885 [DC Library call number: PER Ql.N2], and also in a paper *Regression towards mediocrity in hereditary stature*, *Journal of the Anthropological Institute*: **15**, 246-263, 1885. In the latter, Galton reports (page 246) on his initial discovery that the offspring of seeds ..... *did* not *tend to resemble their parent seeds in size, but to be always more mediocre* (i.e., *more average) than they – to be smaller than the parents, if the parents were large; to be larger than the parents if the parents were very small. ..... The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it*. Galton then describes how the same features were observed in the records of *heights of 930 adult children and their respective parentages, 205 in number.* In essence, he shows that if variate y represents child's height and variate x represents a weighted average of the mother's and father's heights (see the original paper for details), an estimated regression of $\mathbf{Y}$ on $\mathbf{X}$ something like $y = \overline{y} + \tfrac{2}{3}(x - \overline{x})$ is appropriate, although he does not phrase it like this. Galton's paper makes fascinating reading, as does the account in *The History of Statistics* by S.M. Stigler, pages 294-299, Belknap Press of Harvard University, 1986 [DC Library call number: QA276.15.S75 1986]. Today, Galton's analysis would be called *correlation analysis*, a term for which he is also responsible. The term *regression* soon came to be applied to relationships in situations *other than* the one from which it arose, including situations where the explanatory variate can*not* reasonably be represented by a random variable, and its use has persisted to this day, In most current model-fitting situations, there is no element of 'regression' in the original sense, but the word is so established that it is retained in most statistical communication (including these Course Materials).
>
> There has been a dispute about who first discovered the method of least squares. It appears it was discovered independently by Carl Friedrich Gauss (1777-1855) and Adrien Marie Legendre (1752-1833), that Gauss started using it before 1803 (he claimed in about 1795, but there is no corroboration of this earlier date), and that the first account was published by Legendre in 1805. When Gauss wrote in 1809 that he had used the method earlier than the date of Legendre's publication, controversy concerning the priority began. The facts are carefully sifted and discussed by R.L. Plackett in *Studies in the history of probability and statistics. XXIX. The discovery of the method of least squares; Biometrika*: **59**, 239-251, 1972 [DC Library call number: PER QH301.B5], a paper well worth reading. Also of interest are:

University of Waterloo W. H. Cherry

REGRESSION: An Introduction – Overview of the Ideas (continued 3)

NOTES: (cont.) 9.

- C. Eisenhart: *The meaning of 'least' in least squares*, *Journal of the Washington Academy of Sciences*: **54**, 24-33, 1964 [DC Library call number: PER Q11.W32], which is reprinted in:
  - *Precision, Measurement and Calibration*, edited by H.H. Ku, U.S. National Bureau of Standards Special Publication 300, 1969, Volume 1;
- *Gauss, Carl Friedrich* in: *International Encyclopedia of the Social Sciences*, Macmillan Co., Free Press Division, New York, 1968, Volume 6, pages 74-81 [CG Library call number: H40.A215 1968], and:
  - *Encyclopedia of Statistical Sciences*, John Wiley & Sons, New York, 1983, Volume 3, pages 305-309 [DC Library call number: Ref QA276.14.E5 1982];
- S.M. Stigler: *Gergonne's 1815 paper on the design and analysis of polynomial regression experiments*, *Historia Mathematica*: **1**, 431-447, 1974 (see page 433) [DC Library call number: PER QA21.H58x];
  - see also Stigler's book *The History of Statistics* [DC Library call no.: QA276.15.S75 1986] referenced on the facing page.

The ideas introduced in Case 3 occur in the media and related sources, as illustrated in the following six Examples.

4. Example HL52.1: The shorter article EM9701 from *The Globe and Mail* of January 3, 1997, page B7, involves classifying bivariate observations in terms of whether or not they follow a decreasing linear trend. In the notation of Case 3, **Y** is a country's average test score and **X** is the proportion of its 17-year-olds taking physics classes.

EM9701:

Enrolment and test scores

Business often complains that education standards are slipping. Last month, this column suggested that what's really happening is not a decline in standards as much as an increase in the percentage of the population who are taking part in our education system.

If you look at the results of international tests of student ability, it often seems that Canada – with one of the highest education budgets in the world – is not doing well enough. It certainly seems we are not getting what all that money should buy.

But it turns out that, with few exceptions, the countries that score highest on international tests are the ones with the smallest proportion of students in classes, and the low scorers are the ones with the highest proportion in class.

The chart at the right shows the results of a well-documented set of tests in physics, part of the Second International Science Study of 1988.

The results of all but three of the 13 countries in the study – the United States, Hong Kong and Norway – fall within a narrow band, where the results can be simply described. The lower the proportion of a country's 17-year-old population taking physics classes, the higher the average scores.

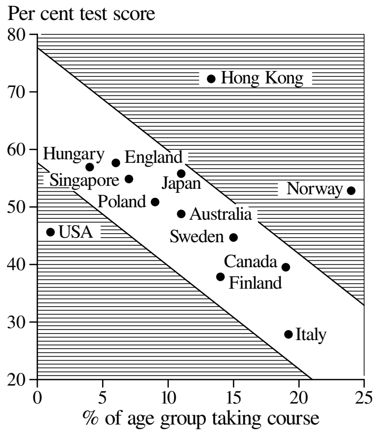
Hungary and England, which ranked second and third in the world-wide tests, had only about 5 per cent of their 17-year-olds in physics classes. In Canada and Italy, which ranked 11th and 13th, about 20 per cent were taking physics classes. Of the exceptions, Hong Kong and Norway did better than they should have, considering the number studying physics, and the United States did worse.

What does this tell us? A small intellectual elite will score higher on average than a large student body with a wider range of intellectual capacities. But what population of 17-year-olds would you rather be recruiting employees from – Hungary's or Canada's?

John Kettle is a consulting futurist based in Oshawa, Ont.

Physics test scores and participation

Second International Science Study, 1988



Sources: Robert Crocker (Memorial University), Second International Science Study and John Kettle

5. Example HL52.2: The article EM9201, from the journal *Nature*: **355**, 25, January 2, 1992 [DC Library call number: PER Q1.N2], is of interest because ts main Answer is based on *extrapolation* of regression lines.

EM9201:

Will women soon outrun men?

The average running velocity ($\overline{v}$) is a crucial determinant of the metabolic demands imposed by competitive running events[1,2]. Its historical progression, therefore, is likely to be important in understanding the physiological determinants of the seemingly inexorable progression of record performances.

We therefore established $\overline{v}$ as a function of historical time (t) for the world records at all the standard Olympic events from the 200 m to the marathon (42,195 m) for men, decade-by-decade, throughout this century[3,4]. We were able to establish this relationship only for events up to 1,500 m since the 1920s for women; data were inadequate for the 5,000 and 10,000 m, although we judged there to be sufficient for the marathon.

In men, the progression of $\overline{v}$ appears to be a linear function of t, with slopes for the different events being remarkably similar (left-hand digram overleaf) – in agreement with the results of Ryder *et al*.[5] These ranged from 5.69 to 7.57 m min^{-1} decade^{-1}, with no systematic variation with increasing race distance. The marathon slope, however, was appreciably greater (9.18 m min^{-1} decade^{-1}).

For women, there were also no significant differences in the slopes among the different events up to the 1,500 m (middle diagram overleaf). The slope, however, was approxi-

(*continued overleaf*)

mately double that for the men, ranging from 14.04 to 17.86 m min^{-1} decade^{-1}. As for men, the rate at which $\overline{v}$ increased in the marathon was appreciably greater (37.75 m min^{-1} decade^{-1}). Despite the potential pitfalls, we could not resist extrapolating these record progressions into the future.

Unless the progression rate of men's records increases relative to that of women, then $\overline{v}$ for these events will be no different for men and women within the first half of the twenty-first century (right-hand diagram below). Beyond that time, current progression rates imply superior performance by women. The projected intersection for the marathon is 1998.

The suggestion that women could, so soon, be running these races as fast as men seems improbable at first appearance. None of the current women's world record holders at these events could even meet the men's qualifying standard to compete in the 1992 Olympic games. However, it is the *rates* of improvement that are so strikingly different – the gap is progressively closing[6]

Although it is difficult to establish a precise metabolic energy equivalent of these rates of improvement, one may estimate from the data of Margaria *et al.*[7] and Wyndham *et al.*[8] that, for the events up to 10,000 m, the progression requires a rate of increase in O_2 consumption of about 10 ml $\text{min}^{-1}\,\text{yr}^{-1}$ for men and more than double that for women.

It is unlikely that we will learn when, and how rapidly, the current high rate of improvement was established, owing to the lack of reliable times over reliable distances in the past. Some world records are available, however, as far back as the 1860s (reference 4); they are consistent with the values 'expected' from the current progression rates. Whether the world record progression rate will begin to slow, either relatively abruptly or more progressively, will only become apparent in the future.

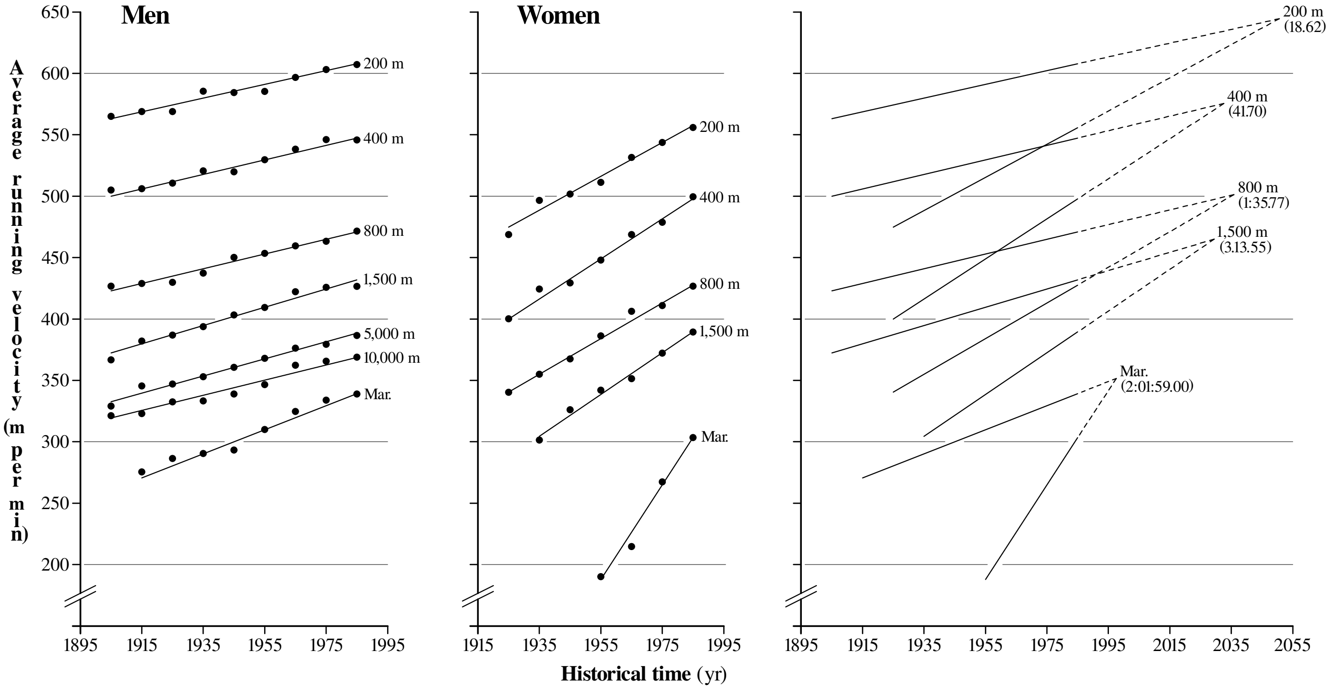
In any event, these results pose four challenging questions to physiologists. Why is:

- the world record progression in the various events so linear over an interval of approximately a century;
- the slope of the record progression so similar from the sprints to the 10,000 m;
- the record progression in the marathon appreciably greater;
- the record progressions for women increasing at such a rapid rate relative to men?

BRIAN J. WHIPP SUSAN A. WARD

Department of Physiology and Anesthesiology,
School of Medicine,
University of California,
Los Angeles, California 90024, USA

1. Lloyd, B.B. *Circ. Res.* **XX** & **XXI** (suppl. 1), 218-226 (1967).
2. Di Prampero, P.E. *Revista di Cultura Sportiva* **3**, 3-7 (1984).
3. Matthews, P. *Track and Field Athletics: The Records* (Guinness, Enfield, 1986).
4. *Progression of World's Best Performances and Official I.A.A.F. Records* (International Athletic Foundation, Monaco, 1987).
5. Ryder, H.W. Carr, H.J. & Herget, P. *Sci. Am.* **234**(#6), 109-119 (1976).
6. Dyer, K.J. *J. Biosocial Sci.* **9**, 325-338 (1977).
7. Margaria, R., Cerretelli, P., Aghemo, P. & Sassi, G. *J. appl. Physiol.* **18**, 367-371 (1963).
8. Wyndham, C.H., Strydom, N.B., Van Rensburg, A.J. & Benade, A.J.S. *S. Afr. Med. J.* **43**, 996-1002 (1969).



World record progression, expressed as average running velocity versus historical time, for men and women, with best-fit linear regressions (solid lines) superimposed. In the right-hand diagram, the regression lines for the common events for men and women (solid lines) are extrapolated (dashed lines) to their points of intersection; the predicted world record times at these intersection points are shown in parentheses (h:min:s).

A December, 2020, Google search yields the world record data given in the table below. The race times at the extrapolated crossing points in the right-hand diagram above are shown in **bold** type in the central column; adjacent to them in smaller type are the ratios of men's records to these extrapolations and the ratios of the women's to the men's records. Three matters seem inhospitable to the implications of the article reprinted above and its title overleaf near the bottom of page HL52.7:

RACE	MEN Time	MEN Name	MEN Year				WOMEN Time	WOMEN Name	WOMEN Year
200 m	19.19	Usain Bolt	2009	1.0306	**18.62**	1.1120	21.34	Florence Griffith-Joyner	1988
400 m	43.03	Wayde van Niekerk	2016	1.0319	**41.70**	1.1062	47.60	Marita Koch	2006
800 m	1:40.91	David Rudisha	2012	1.0537	**1:35.77**	1.1226	1:53.28	Jarmila Kratochvilová	1983
1,500 m	3:26.00	Hicham El Guerrou	1998	1.0643	**3:13.55**	1.1168	3:50.07	Genzebe Dibaba	2015
Marathon	2:01:39	Eliud Kipchoge	2018	0.9973	**2:01:59**	1.1021	2:14:04	Brigid Kosgei	2019

- the relatively long standing of some records;
- differences of still at least 10% between the men's and women's world records;
- except for the marathon, the differences of about 3% to about 6½% that remain between the men's records and the values at the extrapolated crossing points.

(*continued*)

University of Waterloo

W. H. Cherry

REGRESSION: An Introduction – Overview of the Ideas (continued 4)

6. Example HL52.3: The article EM9111 reprinted below from *The Globe and Mail* of August 10, 1991, pages A1 and A2 is of interest because it involves a projection (*i.e.*, an extrapolation) from 1964 of the future requirements for the number of medical doctors in Canada and a claim in the middle column of the article that the projection was much too high.

EM9111:

Cutback in supply of MDs urged

1964 projections 'grossly in error'

BY ROD MICKLEBURGH
Health Policy Reporter

The most comprehensive study yet into the question of physician supply calls for a cut of nearly 10 per cent in the number of students being admitted to Canadian medical schools.

The massive report also urges less reliance on foreign-trained doctors in Canada and stricter controls on their entry into the country.

The report, prepared by medical research professors Morris Barer of the University of British Columbia and Greg Stoddart of McMaster University, further recommends that greater attention be paid to controlling payments to existing physicians, such as by placing more of them on salary or setting income limits for doctors in certain specialities.

Although the report, commissioned by the conference of deputy ministers of health, has not been released, *The Globe and Mail* has obtained a copy of its recommendations and executive summary.

Sources said the entire report is about 700 pages long, the product of 10 months of investigation, more than 70 interviews with experts in the field and eight other reports commissioned separately by the researchers.

"This is the first comprehensive, complete report we have had on this matter," said former Ontario deputy health minister Martin Barkin, who received a copy before leaving his post last week.

"It is certainly going to provoke a lot of discussion."

Health-care critics have long argued that Canada is producing too many new doctors, pointing to this as a major factor in the country's escalating medicare costs.

Between 1975 and 1987, for instance, the number of physicians in Canada rose by 46 per cent, while the population increased by less than 13 per cent.

The Barer-Stoddart report agrees that measures must be taken to curb the ongoing supply of doctors.

"The long-term trend of annual increases in the rate of growth of physician supply in excess of population growth continues without obvious or compelling justification," the report says.

It attributes excessive medical school enrolment to "grossly in error" population projections contained in the 1964 Hall report on health care in Canada.

The report also blames "too little, and fragmented, control exercised over the entry [into Canada] ..... of graduates of foreign medical schools."

Although the report concludes there is no "optimal" number of physicians for Canada, it recommends restricting entry to 1,600 students a year into Canadian medical schools. "This represents a reduction of less than 10 per cent."

This year, just over 1,700 students graduated from medical schools in Canada, down from a peak of 1,835 in 1985.

Dr. Barkin said he feels the proposed cut is unfair. There's no reason why Ontario can't continue to turn out 600 new doctors a year from its own medical schools, he argued. The problem arises from the issuing of 500 additional new medical licences each year to doctors from other areas.

"Why not deal with those 500 instead of depriving some of our own citizens of the right to attend medical school?"

The Barer-Stoddart report did not attempt to set a specific quota for doctors trained outside Canada.

But it did stress the need for "a more concerted effort ..... to monitor and strictly enforce conditions of entry for visa trainees and visa physicians."

The report added that specific policies should be enacted to encourage more Canadian doctors to fill unpopular positions currently staffed by a high proportion of foreign-trained physicians.

Above all, the report emphasized the complexity of the issue, pointing out that every measure taken in one area of health care has an impact in another area, and that therefore it is not enough to implement only a few of its comprehensive set of 53 recommendations, which also include the setting of quality standards, revising the way physicians are paid, and improving the allotment of specialist and general-practice residencies in teaching hospitals.

But the report said governments need not throw up their hands in despair and settle for the status quo. "Most importantly, we need to get over our collective Canadian-style fear of changing directions, even when such redirection is clearly called for," the authors say.

"Letting policy develop by default, disinterest, or insufficient will .... can only be expected to bring about satisfactory results by those who believe in winning lotteries."

Eva Ryten, research director of the Association of Canadian Medical Colleges, would not comment on the report. "The medical schools are going to have to take time to digest it," she said. "This report is so complicated and the issues are so complex that anything I say won't deal seriously enough with these issues."

No date has been set for release of the complete report, which is up to Manitoba deputy health minister Frank Maynard, but copies are being circulated among a small number of interested parties.

Dr. Barkin said he expects the report will be high on the agenda for a meeting of Canadian health ministers next month.

He added there is already a written understanding between Ontario doctors and the province to tackle the issue as part of its newly-established, high-powered joint management committee.

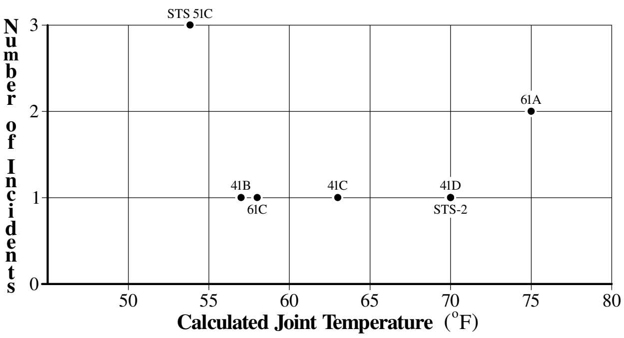
7. Example HL52.4: The excerpt EM8911 given overleaf on page HL52.10 is reprinted from the *Journal of the American Statistical Association* **84**(#408), pages 945-957, December, 1989 [DC Library call number: PER QA276.A599x]; the introduction of the article, by Siddartha R. Dalal, Edward B. Fowlkes and Bruce Hoadley and entitled *Risk Analysis of the Space Shuttle: Pre*-Challenger *Prediction of Failure*, is the basis of the information overleaf. It is of interest as an illustration of the use of scatter diagrams in the Analysis stage of the FDEAC cycle and of the possible consequence of assessing a trend from only *part* of a set of bivariate data. [Author

(*continued overleaf*)

Bruce Hoadley is interviewed in Program 16 of *Against All Odds: Inside Statistics* in the segment on the *Challenger* disaster.]

EM8911: On the night of January 27, 1986, the night before the space shuttle *Challenger* accident, there was a three-hour teleconference among people at Morton Thiokol (manufacturer of the solid rocket motor), Marshall Space Flight Center [NASA (National Aeronautics and Space Administration) center for motor design control], and Kennedy Space Center. The discussion focused on the forecast of a 31°F temperature for launch time the next morning, and the effect of low temperature on O-ring performance. A data set, Figure 1a below at the left, played an important role in the discussion. Each plotted point represents a shuttle flight that experienced thermal distress in the field-joint O-rings; the x axis shows the joint temperature at launch and the y axis shows the number of O-rings that experienced some thermal distress. The O-rings seal the field joints of the solid rocket motors, which boost the shuttle into orbit. Based on the U-shaped configuration of points (identified by the flight number), it was concluded that there was no evidence from the historical data about a temperature effect.

Figure 1a. Frequency of O-Ring Distress vs. Temperature

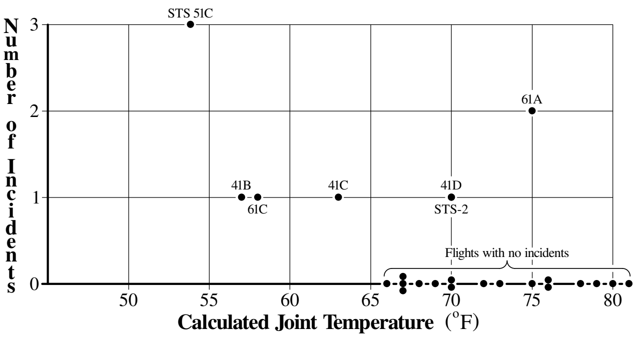


Nevertheless, there was a debate on this issue, and some participants recommended that the launch be postponed until the temperature rose above 53°F – the lowest temperature experienced in previous launches – because the corresponding flight had the highest number of distressed O-rings. Some participants believed, based on the physical evidence, that there *was* a temperature effect on O-ring performance; for example, one of the participants, Roger Boisjoly, stated: *temperature was indeed a discriminator.* In spite of this, the final recommendation of Morton Thiokol was to launch the *Challenger* on schedule. The recommendation transmitted to NASA stated that *Temperature data [are] not conclusive on predicting primary O-ring blowby.* The same telefax stated that *Colder O-rings will have an increased effective durometer ('harder'), and 'Harder' O-rings will take longer to 'seat'* [Presidential Commission Report, Vol. 1 (PC1), p. 97 (Presidential Commission on the Space Shuttle *Challenger* Accident 1986)].

After the accident, a commission was appointed by President R. Reagan to find the cause. The commission was headed by former Secretary of State William Rogers and included some of the most respected names in the scientific and space communities. The commission determined the cause of the accident to be the following: *A combustion gas leak through the right Solid Rocket Motor aft field joint initiated at or shortly after ignition eventually weakened and/or penetrated the External Tank initiating vehicle structural breakup and loss of the Space Shuttle* Challenger *during mission 51-L* (PC1, p. 70). This is the type of failure that was debated the night before the *Challenger* accident.

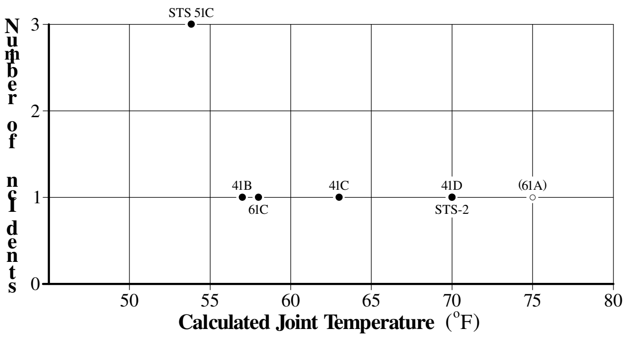
The Rogers Commission (PC1, p. 145) noted that a mistake in the analysis of the thermal distress data – Figure 1a at the left – was that flights with zero incidents were left off the plot because it was felt that these flights did not contribute any information about the temperature effect (see Figure 1b below). The Rogers Commission concluded that *A careful analysis of the flight history of O-ring performance would have revealed the correlation of O-ring damage in low temperature* (PC1, p. 148).

Figure 1b. Frequency of O-Ring Distress vs. Temperature



This article aims to give more substance to this quote and show how statistical science could have provided valuable input to the launch decision process. Clearly, the key question was: *What would have constituted proof that it was unsafe to launch?* Since our model of the phenomeon is stochastic, our answer is necessarily probabilistic. As in the teleconference, a good start would have been an examination of the thermal distress data – Figure 1b – for the presence of a temperature effect. Nevertheless, the most important question was: *What is the probability of catastrophic field-joint failure if we launch tomorrow morning at 31°F?* Both these issues are addressed in the article.

NOTE: 10. It is interesting to speculate, if (hypothetically) the right-most point in Figure 1a above had involved *one* (instead of two) distressed O-rings, whether the modified Figure (shown at the right) would have been interpreted to yield the *correct* Answer (as in Figure 1b at the right above) about the temperature-O-ring distress relationship.



In light of information in the lower paragraph in the left-hand column above, the obituary EM1201 reprinted at the top of the facing page HL52.11 is of interest. [An obituary for Dr. Needleman (see the article EM9204 on page HL52.12) is given in Figure 7.3 of the STAT 220 Course Materials.]

The *Challenger* disaster is also discussed in Statistical Highlight #31 in which, as well as the articles EM8911 and EM1201 reprinted in this Higlight #52, there is more information in article EM8601 on the second side of Highlight #31.

(*continued*)

REGRESSION: An Introduction – Overview of the Ideas (continued 5)

EM1201: The Sydney Morning Herald, February 14, 2012

Space engineer's warnings brushed aside before *Challenger* exploded

ROGER BOISJOLY
US ROCKET ENGINEER
25-4-1938 – 6-1-2012

ROGER Boisjoly, an engineer whose warnings of catastrophy were ignored on the eve of the 1986 *Challenger* space shuttle disaster, which killed seven crew members and plunged the US space program into crisis, has died aged 73.

Boisjoly worked at Morton Thiocol, the company that made shuttle booster rockets. In his time he had watched several successful shuttle launches, including *Discovery* on January 24, 1985.

Examining *Discovery's* discarded boosters, he was horrified that seals in the rockets had been burned through. Only a secondary seal had prevented *Discovery* becoming a fireball.

By the middle of that year, Boisjoly thought he had identified the fault, and wrote a report to alert the company and NASA.

The problem was that in cold temperatures, the seals' rubber stiffened and became more likely to fail. Morton Thiocol formed a study group but, by January 27, 1986, as temperatures dropped to zero on the eve of *Challenger's* blast-off, little progress had been made. In a pre-launch conference, he and four other engineers demanded the launch be postponed.

But their testimony was brushed aside. The next day, unable to watch, his worries seemed unfounded. He had expected the seals to fail on the launch pad. As *Challenger* cleared the launch tower, a colleague whispered to him: "We just dodged a bullet". Seconds later, the spacecraft exploded.

Roger Mark Boisjoly was born in Lowell, Massachusetts, and studied mechanical engineering at the city's university. He then worked for aerospace companies in California on projects including NASA's lunar module and the moon vehicle.

After the *Challenger* disaster, Boisjoly experienced intense feelings of guilt and depression not helped by the fact that many in the business he loved rejected him as an unwelcome whistleblower.

Boisjoly is survived by his wife, the former Roberta Malcolm, and two daughters.

TELEGRAPH

8. Example HL52.5: The article EM9901, from *The Globe and Mail* of June 9, 1999, pages A1 and A8, is of interest because its Answer depends on identifying a *change* in the slope of two trends which had previously been roughly linear over time.

EM9901:

Skin-cancer rate levels off as Canadians cover up

People taking to heart sun-protection measures

JEN ROSS
The Globe and Mail, Toronto

The incidence of the most serious form of skin cancer is finally levelling off after years of steady increases, and researchers are giving protective clothing and sunscreen much of the credit.

A Statistics Canada report released yesterday shows that the number of Canadian men who developed melanoma peaked at about 10 per 100,000 in the late 1980s and has remained roughly the same since. For women, the rates peaked at about nine per 100,000 in the late 1980s and fell slightly during the 1990s.

"People have started taking to heart the sun protection measures that began in the mid-eighties", said the report's author, Leslie Gaudette, who works with Health Canada's cancer bureau.

Alastair Carruthers, president of the Canadian Dermatology Association, is thrilled to see sun-awareness programs he helped create 15 years ago coming to fruition.

"There is always a lag time between our starting a program and seeing its actual effect", he said, adding that the Statistics Canada report helps to refute previous studies that claimed sunscreen is not useful in preventing skin cancer. "This shows our message is right; don't stop using sunscreen."

The report cited data from the 1996 Sun Exposure Survey, which indicated about four in 10 Canadians over the age of 15 wore protective clothing during their leisure hours, covered their heads, used sunscreen on their faces and stayed in the shade as much as possible.

Ms. Gaudette said many people began wearing sunscreen and protective clothing in response to public-awareness campaigns of the 1980s, but changing fashion trends may have also played a role.

"Until recently, fashion trends had been gradually exposing more skin", she said. "There's much more variation in skirt length now than in the sixties and seventies, when the miniskirt really came in."

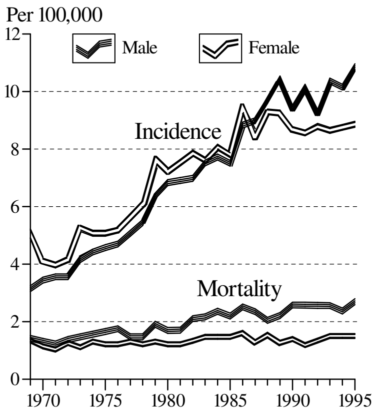
Ms. Gaudette said one of the ways she made the fashion connection was by looking at where people were developing melanoma. She said it was appearing less frequently on women's legs, where it was once common.

The study also revealed a shift in the likelihood of developing melanoma between the sexes. Since the late 1960s, women had been more likely than men to be diagnosed with melanoma, but that began to change in the late 1980s. Incidence rates are now about 10 per cent higher for men than women.

The chance of dying from melanoma, once considered a near-lethal disease, has also leveled off. Survival rates for melanoma five years after diagnosis are 88 per cent for women and 74 per cent for men.

Debbie Altow, a spokeswoman for the Canadian Cancer Society, said mortality rates may also be stabilizing as a result of increased public awareness.

SKIN CANCER
Canadian rates for melanoma.



Source: Statistics Canada The Globe and Mail

(*continued overleaf*)

NOTE: 11. Other matters of statistical interest in the article EM9901 reprinted overleaf on page HL52.11 are:

- the inference, from *observational* data, of a *causal* connection between use of sunscreen and a lower rate of melanoma (see the last paragraph of the left-hand column overleaf on page HL52.11);
- the identification of another explanatory variate – fashion, as reflected in skirt length – from the analysis of *site-specific* melanoma data (see the middle paragraph of the middle column overleaf).

9. Example HL52.6: Of statistical interest in the context of this Highlight #52, in the 1992 article EM9204 from *Newsweek* [March 16, page 62 (Vol. 119, Issue 11)] reprinted below, is a dispute over the form of a statistical model appropriate for analyzing data on lead levels in teeth and IQ; the work was part of data-based investigating over many years by Dr. Needleman and his colleagues of possible effects of lead exposure on children's intellectual development. The article discusses a claim that age should have been included in the model as an explanatory variate.

EM9204:

Lead, Lies and Data Tape

A nasty academic fight becomes a federal case

By Sharon Begley

Scientific misconduct is exceedingly rare and extremely serious. Charges have been brought alleging plagiarism or faking data or falsifying results. The latest case, however, involves the manner in which a researcher strung together a set of equations in order to find a message hidden in a stack of raw data. To reach for a metaphor, this is like bringing a felony indictment for jaywalking.

Two psychologists, both of whom have testified for the lead industry and one of whom has received tens of thousands of dollars in research grants from the industry, have filed misconduct charges against the scientist who first linked "low" levels of lead to cognitive problems in children. They don't suspect that Herbert Needleman of the University of Pittsburgh stole, faked or fabricated data. Rather, they say, he selected the data and the statistical model – the equations for analyzing those data – that show lead in the worst possible light. That's a dispute usually aired in research journals. Now it's become a federal case – and even those scientists most diligent about pursuing misconduct are uneasy. "If it gets to which statistical model is appropriate," a leading government fraud-buster told *Newsweek*, "it gets real hard to believe a misconduct charge."

The case began last year. Psychologists Sandra Scarr of the University of Virginia and Claire Ernhart of Case Western Reserve University filed charges of scientific misconduct against Needleman with the National Institutes of Health. The allegations centre on a 1979 paper. It describes how Needleman and colleagues measured the lead in baby teeth, looking for a link between lead and intelli- gence. NIH told Pittsburgh to convene a panel of inquiry. The panel's report, submitted in December and obtained by *Newsweek*, found that Needleman didn't "fabricate, falsify or plagiarize." It did have problems with how he decided whether or not to include particular children in his analysis, but called this "a result of lack of scientific rigour rather than the presence of scientific misconduct." The panel found Needleman's statistical model "questionable," though. On that basis, the university launched an investigation.

Scarr, Ernhart and the Pittsburgh panel all condemn Needleman for not using a different model – one that, say, factored in the age of each child. If he had, they say, lead would not have had an impact on IQ. But last year Environmental Protection Agency scientist (and recipient of a MacArthur Foundation "genius" award) Joel Schwartz re-analyzed Needleman's data. He factored in age explicitly. "I found essentially identical results," he says.

Flawed sampling? Another criticism addresses whether Needleman ignored data he didn't like. Scarr alleges that he looked at the children's lead levels and IQ score, and only then "decided in or out for each child." In fact, "the reasons for exclusion can be found in the protocol," says econometrician Hugh Pitcher of Battelle Memorial Institute, who analyzed the 1979 data when he was at EPA. They include such things as the child's having a head injury. The selection, says Pitcher, was done before the researchers knew the kids' IQs.

This is not to say Needleman's work was perfect, just that any lapses did not change the outcome. Ernhart insists this is not good enough. "He doesn't feel it's necessary to do things the way you're supposed to," she says. "You have the sense that he was going to demonstrate the effects of lead no matter what."

How does this case affect lead policy? "We don't even use Needleman's study any more," says EPA's Schwartz: it has been superseded by research showing effects of lead at even lower levels (*Newsweek*, July 15, 1991). The politicization of misconduct may be just starting, though. Crying fraud, says an NIH scientist, "can be used to railroad people you don't like."

10. Postscript

Regression [*i.e.*, mathematical modelling of the *trend(s)* displayed by a set of multivariate data] is a widely used statistical method of data analysis; users, particularly those with limited statistical background, routinely give little or no recognition to some matters raised in this introductory Highlight #52, such as:

∗ the idea of regression as the end of a progresson of statistical methods for describing one, then two, then many populations;
∗ the difference in intent between modelling the *exact* behaviour of individuals and the apparent *overall* behaviour of a group;
∗ use of the notation $_{\text{reg}}\overline{\mathbf{Y}}$ as a reminder of 'idealizing' (or 'mathematizing') an (often untidy) *trend* in an *average* response variate.

Raising such matters here involves two competing disadvantages:

− they may distract the beginning student from the primary goal of this Highlight #52 (and of Part 13 of STAT 221) – becoming comfortable with formulating regression models and competent in using them;
+ an Answer based on a regression model may be compromised if such matters are ignored.

In a similar vein, distinctions introduced in STAT 221 Figure 13.3 between the respondent population regression, the *model* regression, and the *estimated* (model) regression are seldom made elsewhere and involve the same two competing disadvantages.
The reader should keep these caveats in mind when studying the relevant Materials.