THE FIRST STEP: UNDERSTANDING SIMPLE LINEAR REGRESSION

MORE ON MARS

The "universe" is only 200 (Fig. 2-1)

Nothing in this chapter rests on the fact that there are only a finite number of Martians (200) in the "entire population". Each circle in the figure might just as well represent 10 Martians or 10000 Martians. The "population" in statistics is often conceptual, referring only to "patients like mine" or "future patients". Science isn't about local, particularistic situations. If one published an article, based on Canadian data, in a U.S.A. journal, then presumably the findings are sufficiently general to apply to U.S.A. patients as well. Otherwise, would U.S.A. readers be interested? Just as we think of science as borderless, we also think of it as timeless. The only reason G&S used the finite number 200 here is because they wanted to be able to show all of the data! So think of each circle as representing say 100 Martians.

Equal variability of Y at all X values? (Fig 2-1)

You may get the visual impression that the (conditional) distributions of Y [=weight] in the vertical "slices" (i.e., the height-specific distribution of weight) at the left and right end of the X-axis are narrower (i.e., have less vertical variation) than the slices near the middle. The authors list "equal variability of all Y | X distributions" as assumption # 3 at the top of page 16. You may think that this "requirement" is not satisfied by the data in Fig 2-1. Notice however that the requirement is in terms of equal STANDARD DEVIATIONS. The reason the variability *seems* larger in the vertical distributors at the middle of the X-axis is that your eye is observing the RANGE. And yes, if there are more observations, the range is likely to be larger, even if the S.D. is the same across the board!

Note that the (relative) numbers of individuals at the different "X" values have nothing to do with the assumptions being (or not being) satisfied. And in any case, if one has any choice in which individuals to study, one would gain by "quota-sampling" from the various X "slices", and lose by taking a purely random sample (i.e., by ignoring the individuals' X values).

Indeed, many textbooks (e.g., NKNW page 33) depict the "population" of y values as a *series* of Gaussian distributions with their means connected by a straight line. KKMN p 45 Chapter 5 show general distributions. Each sub-population or slice is thus infinite in size, and so the question of how many are available for study of each x-value becomes irrelevant. What affects the quality of the parameter estimate derived from the

sample is how well the sample observations are spread out (by nature or by investigator's design) along the X-axis.

• ***"The MEAN weight of Martians of each height increases linearly as height increases"*** (p.12)

This is the first assumption in simple linear regression. Another way to put it is to say that "if we represented the conditional *means* corresponding to the different X's as *dots* -- and then joined the dots --these "center-dots" would form a straight line".

Note that there do not have to be slices (Y values) at every possible value of X -- only at the biologically possible X values. For example, if Y = birthweight, and X = parity, then it makes no sense to think of the "Y" distribution at X = 1.5!

Note that the statement that the 'mean Y at X' INCREASES AS X INCREASES" gives less of a "causal" message.

As I have written in my Notes on Chapter 5 of KKMN, there is nothing in regression *per se* that demands that (a) we work with MEANS [we might be more interested in how MEDIANS behave in relation to X] or (b) that they fall along a single straight line. It just "so happens" that the statistical properties of MEANS and of other least squares estimators (e.g., slopes) are more mathematically "tractable" than those of medians or modes (and differences thereof).

• ***"This line does not make it possible to predict the weight of an INDIVIDUAL Martian"*** (P12)

So, do we necessarily care? If we are not so much interested in individuals as in aggregates of individuals, the variability is less of an issue.

THE POPULATION PARAMETERS (p.12)

• ***"The EFFECTS of one or more independent variables combine to determine the value of the dependent variable"*** (3rd sentence of 1st paragraph)

This is somewhat loose. First, they don't *completely* determine the value of the dependent variable. Even in the perfectly linear Fig. 2-1, the height doesn't completely determine weight. Second, the word "effect" is overstating it. It may be that the X value completely determines the *mean* Y at each X, but to call this the EFFECT of X on Y is putting it too strongly. For all we know, it may be that it is Y that is driving X, or maybe it is a third variable Z that is driving both of them. The authors recognize this a few sentences later, but they should -- at least the first time they use it -- explain that they are using the term "effect" more in a mathematical sense.

"The slope of this so-called LINE OF MEANS" (bottom-p 12)

This is a very useful way to think of a regression equation.

Equation (2.1) $\mu_{y|x} = \beta_0 + \beta_1 x$

This " y | x " terminology is crucial to convey that the means are CONDITIONAL MEANS.

Moreover, the **interpretation of** β_1 is VERY CLEAR AND ACCURATE

$$\beta_1 = \frac{d\,\mu_{y|x}}{d\,x} \qquad \beta_1 \neq \frac{d\,y \mid x}{d\,x} \qquad \beta_1 \neq \frac{d\,y}{d\,x}$$

β_1 is a difference of AVERAGES!

Indeed it makes no sense in a cross-sectional study to think of dy/dx: when we move from persons with a particular X value to those with another X value, which "y difference" are we speaking of? There will be a y difference for every *pair* of persons, one of whom has X=x and the other who has X=x +dx.

• ***"Intercept β_0 = - 8 g"***

Many texts give a meaning to the intercept (e.g., the mean value of Y when X=0). Not all emphasize that this is meaningless if the observations are taken at X values that are a long ways from X = 0! G&S are wise to say that β_0 is the intercept of the LINE and to leave it at that!

• ***"For any given x value, the values of y are NORMALLY DISTRIBUTED"*** (beginning of last paragraph, page 13)

The authors should make clear that this is an ASSUMPTION, which (since they made up the data!) they have had an easy time fulfilling in their Fig 2-1 example.

This assumption, restated in their summary at bottom/top of pp 15/16, is not always critical. It depends on how one uses the regression. [See notes on KKMN Chapter 5].

Likewise, one can sometimes take the *"REQUIREMENT that the standard deviation BE THE SAME FOR ALL VALUES OF X"* (last line p 13) with a grain of salt. Again, see my notes on KKMN Chapter 5 for more on this point.

Equation (2.2)

$$y = \mu_{y|x} + \varepsilon$$

$$= \beta_0 + \beta_1 x + \varepsilon$$

(p.14)

This alternative term is not just "more convenient" (line 15, page 14) but also MORE COMPREHENSIVE than equation (2.1). Equation (2.1) only specifies the SYSTEMATIC part (i.e., the LINE of MEANS). Equation (2.2) specifies both the SYSTEMATIC and the "RANDOM" [or "ERROR" or "VARIATIONS ABOUT THE MEANS"] component.

Indeed, in the unification of different regression models brought about by the notion of generalized linear models, the SYSTEMATIC and the ERROR components of the regression are given equal status. This is evident in the unified INSIGHT and GENMOD procedures in SAS, where one can "toggle" between models by specifying different functions for the systematic components and different distributions for the random component. [By default, the FIT command in INSIGHT will fit model 2.2 with the ε's having Gaussian variation]

For my rants on why (a) **the word derivation is a better term than error** and (b) **"Gaussian" is a better term than "Normal"**, see my notes on KKMN Chapters 1 and 3.

• ***Assumption: "normality"***

One of the most common mistakes new or even experienced data analysts make is to test the "normality" or "Gaussian-ness" of the unconditional (overall) distribution of the Y's, rather than the conditional (X-specific) distribution of the DEVIATIONS (RESIDUALS) from the fitted systematic portion of the regression.

• ***Assumption: "deviations are statistically independent"***
(fourth bullet on p. 14)

This "independence" is usually the case, unless

a the observations are ordered in time/space, and there is some correlation between the deviations of "adjacent" observations from their $\mu_{y|x}$'s [i.e., over and above the fact that adjacent Y's will be similar because of the "adjacent" values of their $\mu_{y|x}$'s]

b the observations are from the same unit (e.g., person or household or medical practice or school...] and because of this have ε's that tend to be on the same side of $\mu_{y|x}$, because of shared influential factors (e.g., personal, genetic, environmental, caregiver, "cultural") that are not captured in the "X" variable(s).

HOW TO ESTIMATE THE LINE OF MEANS (P.14)

The presentation in parts A and B is a nice touch, reminiscent of similar pedagogic strategies in Glantz's Primer of Biostatistics text.

• ***"Before we can test the hypothesis that the apparent trend in the data is due to chance ..."*** (1st line, p16)

Unfortunately, this emphasis on testing null hypotheses, quite prevalent when this book was written, is still with us ten years later. One can imagine that in some situations, the investigator might be interested to establish if there is any NON-ZERO trend in the population. But in most applications the question is not WHETHER there is a trend, but THE MAGNITUDE of the TREND. For these reasons, after estimating the trend from the sample (as the authors must do for their "test") it is wiser to supplement this point estimate with an INTERVAL ESTIMATE (i.e., CONFIDENCE INTERVAL). The same standard errors used in testing are also used to construct the CI's.

• ***"The hypothesis that the apparent trend is due to chance"***
(same sentence, p. 16)

Strictly speaking, this is not an accurate way to say it. *Hypotheses are statements about PARAMETERS (i.e., β_0 and β_1).* They are not about "apparent" or "empirical" trends (i.e., trends seen in the *sample*). Moreover, if one is going use the phrase "due to chance" when dealing with the behaviour of data (i.e., b_0's and b_1's) CONDITIONAL on hypotheses about β_0 and β_1, then one should speak about variations that could occur with a predictable frequency by chance ALONE. The authors are more technically accurate when they come to actually test hypotheses on pp. 26-27.

• ***"Best Estimates from the Data"*** (pp. 16,17)

I like the authors' use of the Arabic b_0 as "an estimate of the Greek β_0" and similarly "b_1 as an estimate of β_1". This is less ostentatious and easier to do on a word processor, than the "hat" notation $\hat{\beta}_0$ and $\hat{\beta}_1$. [See notes on section 5.3 of KKMN]

• ***"The 'best' line"*** (pp. 17-18)

The "least squares" is just one criterion. Others are "least absolute deviations" [in a simple "no X" situation, this leads to the median, whereas the "least squares" criterion leads to the mean], or "maximum likelihood". Simply balancing the residuals -- so that the sum of the negative residuals cancels out the sum of the positive residuals -- is not enough, since any line that goes through the point $\{\bar{x}, \bar{y}\}$ would do this [e.g., line II in Figure 2-4].

The Maximum likelihood criterion requires that one specifies the pattern (shape) of the "error" or residual variation, whereas the least squares criterion *per se* doesn't actually invoke any particular error distribution. The "least absolute deviation" criterion would take more computing time, and (since computing alone is no longer a serious obstacle) the reliability of the resulting estimates is more difficult to quantify.

• **Table 2-1** (computations, followed by points on estimated regression line, and computed residuals)

The last two columns on the left half of Table 2-1 are typically called "***predicted values***" and "***residuals***".

Don't fuss about the manual calculations of b_0 and b_1. But do get in the habit of (a) first plotting the data and (b) making "eye-fit" estimates of β_0 and β_1.

Also, do think of b_0 as the extension (**extrapolation**) back towards X=0 of the point ($Y = \bar{y}$, $X=\bar{x}$) using the fitted slope b_1. Thinking of it this way will give some intuition later as to why -- in some datasets -- b_0's have large standard errors!

Variability about the regression line

• ***"Standard Error of Estimate"*** (p. 20)

This is an unfortunate terminology. A better term, used in some texts and software (including SAS!) is "root mean squared error" or "RMSE". Note from pp. 40-41 of version 1 of text that BMDP calls it Standard Error of Estimate, while SPSS simply calls it Standard Error. In version 2 of text, p 44, both SPSS and SYSTAT say Standard Error of the Estimate".

Given that Standard Error ("SE") is universally used to denote the precision of a *statistic* or parameter estimate, and that this SE is inversely related to the square root of the number (n) of sample observations, it is unfortunate that it has also come to be used for the average "Error" in a *single* observation. ***Note that the RMSE does not vary systematically with n.*** I think the term "Standard Error of Estimate" came to us from psychometrics and reliability testing of individuals; for example, the "blurb" in the GRE handbook published by the Educational Testing Service used to say that GRE scores (which vary from 200 to 800 points *across* individuals) have -- for a certain test -- a "Standard Error of Estimate" of 10 points. The 10 is a ***conceptual standard deviation*** . It refers to the (hypothetical) variation (measured by standard deviation) in a person's test scores if that person took several different versions ("3-hour" samples of questions) of the test. It is akin to the standard deviations that laboratories report

when they measure the same quantity a number of times to judge reproducibility. This "average measurement error" is an attribute of the measurement process used for individuals. On the other hand, in a random sample of n=100 epidemiology students, it might be that the mean of the 100 observed scores was 625, and the *inter*-student standard deviation was 80. Then if we computed the standard error of this point estimate to be the standard error of the mean, i.e., SEM= $80/\sqrt{100}$ =8, then we would use the 8 and the 625 to construct a confidence interval for the *mean* score (μ) of all epidemiologists. If we had n = 400 students, the SEM would drop to 4. However the *within-student* noise in the measurements (i.e., the [conceptual!] variation of an individual student's scores over different versions of the test) would stay at 10.

• ***Why do we divide the sum of the 10 squared residuals by n-2 = 8 (rather than n-1 = 9) to get their variance?*** (eqn. 2.7, p. 20)

For answers, see either section 5.6 of KKMN, or my notes on this section. My notes contain exercises which make it clearer that the divisor is the number of INDEPENDENT residuals one has for assessing variability of the ε's [for example, if one fits a (seemingly perfect) line through just n = 2 datapoints, one has no way to internally judge the variability in the population these 2 datapoints come from; with n = 3, one has one independent residual, and so on ...].

Standard Errors of the regression coefficients (p. 21-24)

• ***Footnote about "normal (Gaussian)" distribution of all possible values of*** $\mathbf{b_0}$ ***and*** $\mathbf{b_1}$

If the ε's are Gaussian , then all the possible b_1's, being a linear combination of n ε's also have Gaussian (sampling) variability.

The authors now seem to be dispensing with (relaxing) this condition or assumption of Gaussian-ness of the ε's , invoking the Central Limit Theorem INSTEAD.

This is appropriate, since ε's are seldom truly Gaussian anyway, and since we often have sample sizes in the 20's, 30's or well beyond that make the (sampling) distribution of the (possible) b_i values close to Gaussian, EVEN IF the underlying INDIVIDUAL ε's are NOT.

In course 513-607, the Central Limit Theorem was used to explain why the sampling behaviour of $\bar{y}$ is closer to Gaussian then the distribution of individual Y's; even if the universe of individual Y values have a decidedly non-Gaussian distribution, the distribution of the possible values of the STATISTIC (or AGGREGATE value) $\bar{y}$ -- calculated from a sample of n of these -- will be effectively indistinguishable from Gaussian when the number (n) of elements (individual y-values) in the average ($\bar{y}$) is sufficiently large.

*"**How large is sufficiently large?**"* depends on the shape of the Y (or ε) distribution. For example, if Y's have a *uniform* distribution on (-0.5, +0.5) then averages of 12 of them will have a "visually indistinguishable from Gaussian" distribution, with mean 0 and standard deviation 0.0833. (Indeed this is how many random number generators generate a stream of random values having a Gaussian distribution: they add (or average) together 12 values taken from a uniform distribution. The "wilder" (the more non-Gaussian and non-symmetric ...) the ε's, the greater the sample size n required for linear combinations of the individual Y's (used to compute means or slopes or whatever...) to have a Gaussian distribution. Most good elementary statistics texts show the "effect" of the Central Limit Theorem as a function of (a) sample size and (b) the shape of the Y (or ε) distribution. See for example M&M p402-405), Colton (p 101-108), Armitage and Berry (p 81-82) or my course 607 notes on Chapter 5.2 of M&M, . See also my comments on this last year when we used KKMN (Chapter 3.3).

• ***Footnote re Neter text ,*** page 23 (*derivation*)

[G&S1: ref was to ver 3] This is the newest edition (4th) of that text. Edition 4 has 4 authors, and I use it in course 513-697, referring to it as NKNW (first N is still Neter!)

• **Equations 2.8 and 2.9 : SE[b_1] and SE[b_0]** (pp 23-24)

I much prefer the rightmost expression

$$SE[b_1)] = \sqrt{\frac{s^2_{y|x}}{\Sigma(x - \bar{x})^2}}$$

I prefer even more my own rearrangements of this, which I describe in pages 9-11 of my 607 notes on Chapters 2 & 9 of M&M under the heading "Factors affecting the reliability (of the estimated slope and intercept)"

The reason I prefer the equivalent formula

$$SE[b_1] = \frac{SD_{y|x}}{\sqrt{n}\ \ SD[X]}$$

is that it explicitly highlights the ***3 influences on the stability or reliability of the estimated slope***.

- The variability (amplitude) of the residuals: $SD_{y|x}$, *(the narrower the better)*
- The spread of the X's: SD[X] *(the wider the better)*
- The sample size: n *(the bigger the better)*

It is also instructive to examine the structure of $SE[b_0]$ in equation 2.9 p 25.

If the x's were centered over X=0, so that $\bar{x} = 0$, then $SE[b_0]$ reduces to

$$SE[b_0] = \frac{SD_{y|x}}{\sqrt{n}}$$

reminiscent of the formula for $SE[\bar{y}]$ or "SEM". But in fact, if the data are centered over $\bar{x} = 0$, then b_0 is in fact $\bar{y}$ [check formula 2.6]

When $\bar{x}$ is a distance from X=0, then b_0 is an "artificial" quantity obtained by projecting from the point $(Y = \bar{y}, X = \bar{x})$ towards X = 0 using the slope b_1 (see equation 2.6). The further the origin X = 0 is from the data (i.e., from $\bar{x}$), the less trustworthy is the resulting b_0. This is reflected algebraically in the second term (the $\bar{x}^2$) inside the square root sign in equation 2.9.
Equation 2.6 gives

$$b_0 = \bar{y} - b_0 \bar{x}$$

The "instability" of b_0 is reflected in its SE, or in its square

$$Var[b_0] = Var[\bar{y}] + \bar{x}^2 Var[b_1]$$

[$\bar{y}$ and b_1 are statistically independent, $\bar{x}$ is considered a 'constant', so the variance of b_0 is the sum of the variances of its two components]

The rightmost component of $Var[b_0]$ is usually rewritten using equation 2.8. The further $\bar{x}$ is from zero, the bigger its contribution to $Var[b_0]$.

The leftmost component of $Var[b_0]$ is none other than $Var[\bar{y}]$, dèja vu.

Centering is a valuable tool for avoiding instability in parameter estimates. For a good example, see my notes on section 5.7 of KKMN, and especially the model fitted to the numbers of hurricanes that hit the USA each decade.

Inferences concerning a "fitted" $\mu_{y|x}$ value

This important topic is not covered in G& S. It is covered in

KKMN Chapter 5.9 pages 57-59 under the heading **"Inferences About the Regression Line"**

M&M 3rd edition, Chapter 10 pages 673-675, under the heading **"Confidence intervals for mean response"**

and in p 12 of my notes on chapters 2/9 of M&M under the heading **"SE for Estimated $\mu_{y|x}$ or '<u>average</u> Y at X' "**

Prediction of a new Y value at $X = X_0$

Again, this topic is not covered in G&S. It is covered in

> ***KKMN Chapter 5.10 pages 59+ under the heading*** **"Prediction of a New Value of Y at X_0"**
>
> ***M&M 3rd edition, Chapter 10 pages 676-678, under the heading*** **"Prediction intervals"**
>
> ***and in p 12 of my notes on chapters 2/9 of M&M under the heading*** **"Confidence Interval for Individual Y at X"**

Obtaining point estimates and intervals from SAS

The bands for $\mu_{y|x}$ ***and for*** **y|x** ***can be obtained***

> ***via*** **INSIGHT** ***(Curves->confidence curves)***

and

> ***via*** **PROC REG in SAS**
>
> **-** ***as*** **(L95M, U95M) and (L95, U95).**

HOW CONVINCING IS THE TREND ? (p25)

• ***Testing the slope of the Regression line*** (p. 26)

The test of H_0: $\beta_1 = 0$ is carried out automatically in practically all statistical packages. See pages 40-41 for the formats used by four packages. By the way, since 1990 one important newcomer to the market is **Stata**. See www.stata.com for more on this excellent multi-platform package.

• Does "p < 0.001" from a test of $\beta_1 = 0$ mean that "it is <u>unlikely</u> that $\beta_1 = 0$"?

Not necessarily. It also depends on the weight of other evidence! Many of us use this "shorthand" to interpret p-values. However, technically speaking this interpretation of a "frequentist" p-value is inaccurate: it treats β_1 as having a posterior (Bayesian) distribution, thereby implying -- without making it explicit -- where one was "coming from" before the data were analyzed i.e., what the analyst's prior distribution was for β_1.and how likely one believes β_1 to be near 0.

• ***"As height increases, weight increases ..."***

See my earlier comments on this over-interpretation.

If in a cross-sectional study, such as the one done in Busselton in 1972, we saw a negative (and statistically significant) slope

linking Y = height and X = age, would we conclude that "as we age, we get shorter" ?

• ***"Of course, ... this small p-value does not guarantee ... It does however ... if β_1 were zero ... "*** (2nd last para, p. 26)

This is a more technically correct way of paraphrasing the p-value.

• ***"Testing hypotheses about, or computing CI's for the intercept β_0 ... "***

One could, but there is often no interest in β_0. Moreover, the stability/reliability of the estimate b_0 of β_0 critically depends on how the X's are represented. For example if we used

X = height minus average height,

then

$b_0 = \bar{y}$.

If our X values are a long ways from X=0, then b_0 will also contain the " $- b_0 \bar{x}$ " component.

Comparing Slopes and Intercepts

I prefer how G&S approach this over the complex way KKMN approach this same question. For practice with this, on a *real* (as opposed to a Martian or science fiction) dataset, see the excerpts from the article on bone density in the 20th and earlier centuries.[on web page for this course]

Notation:

" $s_{\text{parameter estimate}}$" **vs** SE[parameter estimate]

G&S and other authors use the notation " $s_{\text{parameter estimate}}$" to denote the Standard Error of a parameter estimate. I prefer to use SE. The vast majority of computer packages use SE. A few (the ones aimed more at statisticians) use " SD[parameter estimate] ".

• **"Testing the Regression as a whole"**

This is covered as a whole extra chapter (6) in KKMN! I don't find the use of the F test (equation 2.18) as helpful as testing the $\beta_1 = 0$ directly. Taking squares destroys the sign of b_1, and in any case $F = t^2$ when there is only 1 "X" term in the regression. But do see the last paragraph on this topic (last para of p. 32).

CORRELATION COEFFICIENTS AND REGRESSION COEFFICIENTS (p.39)

See my notes on correlation from course 513-607 (Chapter 2 and Chapter 9 from M&M).

• The relation between r and b_1

This is given at the bottom of page 39 as

$$r = b_1 \times \frac{SD[X]}{SD[Y]}$$

I prefer to write it the *other way around*

$$b_1 = r \times \frac{SD[Y]}{SD[X]}$$

The coefficient r is unitless -- the XY dimension in the numerator cancels with the $\sqrt{X^2}\sqrt{Y^2}$ dimensions in the denominator.

Writing $b_1 = r \times SD[Y] / SD[X]$ shows that b_1 is in the correct units Y/X. It also shows that if the Y and X value happen to each have SD's of unity, then the slope b_1 and the correlation coefficient r have the same value. So, r can be thought of as the slope when the Y values are transformed to Z scores and plotted against the X values transformed into Z scores, i.e.,

$$R_{xy} = \text{slope of } Z_Y \text{ on } Z_X$$

last (previous) update: 2001.06.03 (2000.06.04)