Read M&M §10.1 and §10.2 up to page 686. Inference for simple linear regression: normal theory. Fighting your way through computer output.

In Lecture 3 the method of least squares was treated purely as a mathematical method for fitting straight lines to scatterplots. The statistical analysis of the fit makes assumptions about the probabilistic mechanism by which the data in a scatterplot are generated. When the assumptions are satisfied, probability theory can make assertions about the behaviour of various quantities associated with the fit.

In real situations we seldom know when the theoretical model is completely satisfied, but we nevertheless often invoke the statistical theory, based on idealized assumptions, as a guide.

1. An artificial example

The standard theoretical model for a scatterplot with points (x_i, y_i), for $i = 1, 2, \ldots, n$, places no assumptions on how the x_i's are generated, but, conditional on the x_i's it assumes

$$y_i = \alpha + \beta x_i + \epsilon_i \qquad \text{for } i = 1, \ldots, n,$$

where α and β are (unknown) constants, and the ϵ_i's are independent random errors, each $N(0, \sigma)$ distributed for some (unknown) standard deviation σ. Equivalently, conditional on those x_i's, the y_i's are assumed to be independent, with $N(\alpha + \beta x_i, \sigma)$ distributions.

The least squares procedure finds values a and b to minimize

$$\sum_i (y_i - a - bx_i)^2$$

The ***coefficients*** a and b depend on both the x_i's and the y_i's; they are random variables. You could think of a and b as estimates for the unknown true coefficents α and β, and think of the ***fitted values***

$$\widehat{y}_i = a + bx_i$$

as estimates of the mean values $\alpha + \beta x_i$. Conceptually, the data are modelled as a combination SIGNAL + NOISE, where

$$\text{SIGNAL}_i = \alpha + \beta x_i \qquad \text{and} \qquad \text{NOISE}_i = \epsilon_i \qquad \text{for } i = 1, \ldots, n$$

The least squares procedure decomposes the observed data as FIT + RESIDUAL, where

$$\widehat{y}_i = \text{FITTED}_i \qquad \text{and} \qquad e_i = y_i - \widehat{y}_i = \text{RESIDUAL}_i$$

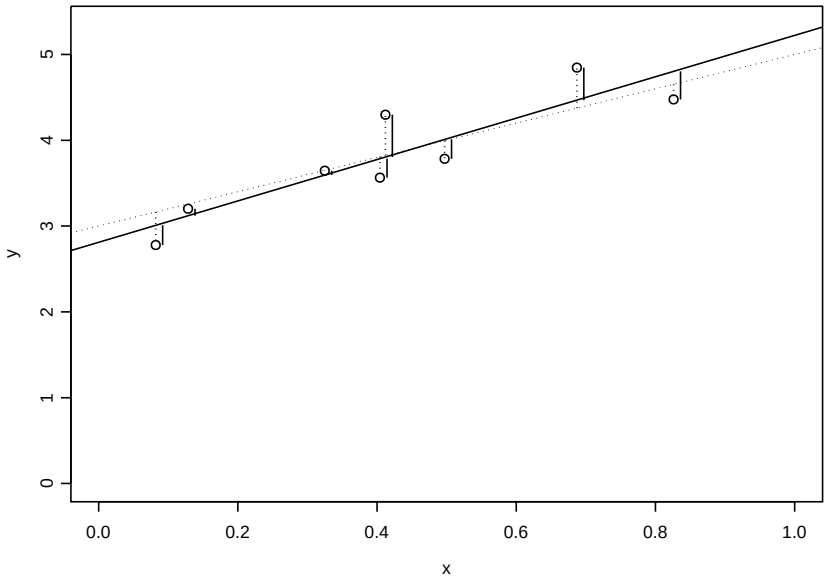
The residuals are the part of the data that remains after removal of a guess at the true signal. We are hoping that the fitted values give a good estimate of the signal, with the residuals capturing some features of the noise.

As an illustration, I artifically generated some data, with $\alpha = 3$, $\beta = 2$, and $\sigma = 0.5$, and $n = 8$. The least squares fit had $a \approx 2.81$ and $b \approx 2.41$. The next table summarizes the data and the least squares fit.

x	$\alpha + \beta x$	ϵ	y	fitted	residual
0.08	3.16	−0.38	2.78	3.01	−0.23
0.13	3.26	−0.05	3.2	3.12	0.08
0.32	3.65	0	3.65	3.59	0.05
0.4	3.81	−0.24	3.56	3.79	−0.22
0.41	3.82	0.47	4.3	3.8	0.49
0.5	3.99	−0.21	3.78	4.01	−0.23
0.69	4.37	0.47	4.85	4.47	0.38
0.83	4.65	−0.18	4.48	4.8	−0.33

In the plot, the dotted sloping line shows the true mean (the line $y = \alpha + \beta x$), and the dotted vertical lines represent the ϵ_i errors. The solid sloping line shows the

least squares fit (the line $y = a + bx$) and the solid vertical lines represent the residuals, slightly offset horizontally, for clarity. Notice how the residuals are similar to the errors, but not quite the same.



Minitab output (slightly edited):

The regression equation is $y = 2.81 + 2.41x$

Predictor	Coef	StDev	T	P
Constant	2.8115	0.2377	11.83	0.000
x	2.4111	0.4926	4.89	0.003

S = 0.3307 R-Sq = 80.0% R-Sq(adj) = 76.6%

ANALYSIS OF VARIANCE

Source	DF	SS	MS	F	P
Regression	1	2.6200	2.6200	23.96	0.003
Residual Error	6	0.6562	0.1094		
Total	7	3.2762			

How much of the probability theory and the detail behind the least squares procedure do you need to understand in order to interpret the output correctly?

In my opinion, there is little point in memorizing formulae for the coefficients a and b, because all calculations are carried out easily by computer. There are a few points worth knowing, because they explain a lot.

(i) For all (linear) least squares problems, the fitted values are expressible as linear combinations of the observed y_i's, with multipliers depending only on the x_i values. For the x_i's in this artificial example, the multipliers are given in the next table.

	y1	y2	y3	y4	y5	y6	y7	y8
fit1	0.38	0.34	0.2	0.14	0.13	0.07	−0.08	−0.18
fit2	0.34	0.31	0.19	0.14	0.13	0.08	−0.05	−0.14
fit3	0.2	0.19	0.15	0.13	0.13	0.11	0.07	0.04
fit4	0.14	0.14	0.13	0.13	0.13	0.12	0.12	0.11
fit5	0.13	0.13	0.13	0.13	0.13	0.12	0.12	0.12
fit6	0.07	0.08	0.11	0.12	0.12	0.14	0.17	0.19
fit7	−0.08	−0.05	0.07	0.12	0.12	0.17	0.28	0.37
fit8	−0.18	−0.14	0.04	0.11	0.12	0.19	0.37	0.49

You should read the table as asserting that

<1> $$\widehat{y}_1 = 0.38y_1 + 0.34y_2 + 0.2y_3 + 0.14y_4 + 0.13y_5 + 0.07y_6 - 0.08y_7 - 0.18y_8$$

and so on. There is a formula that gives the multipliers as a functions of the x_i's, but you don't need to memorize it. For any least squares problem, it would not be hard to produce the values by computer, if you ever wanted to inspect them. For different least squares problems the multipliers are different.

If all the ϵ_i were zero, all the (x_i, y_i) points would lie on the line $y = \alpha + \beta x$. The fitted values $\widehat{y}_i$ would then coincide with the means $\alpha + \beta x_i$. In other words, if we replace the observed y_i values by their means we must recover those same means using least squares:

<2> $$\begin{aligned}\alpha + \beta x_1 = {} & 0.38(\alpha + \beta x_1) + 0.34(\alpha + \beta x_2) + 0.2(\alpha + \beta x_3) + 0.14(\alpha + \beta x_4) \\ & + 0.13(\alpha + \beta x_5) + 0.07(\alpha + \beta x_6) - 0.08(\alpha + \beta x_7) - 0.18(\alpha + \beta x_8),\end{aligned}$$

and similarly for all the other means.

The two formulae, <1> and <2>, have an important consequence when $y_i = \alpha + \beta x_i + \epsilon_i$, namely:

$$\begin{aligned}\widehat{y}_1 = {} & 0.38(\alpha + \beta x_1 + \epsilon_1) + 0.34(\alpha + \beta x_2 + \epsilon_2) + 0.2(\alpha + \beta x_3 + \epsilon_3) \\ & + 0.14(\alpha + \beta x_4 + \epsilon_4) + 0.13(\alpha + \beta x_5 + \epsilon_5) + 0.07(\alpha + \beta x_6 + \epsilon_6) \\ & - 0.08(\alpha + \beta x_7 + \epsilon_7) - 0.18(\alpha + \beta x_8 + \epsilon_8) \\ = {} & \alpha + \beta x_1 + 0.38\epsilon_1 + 0.34\epsilon_2 + 0.2\epsilon_3 + 0.14\epsilon_4 + 0.13\epsilon_5 + 0.07\epsilon_6 - 0.08\epsilon_7 - 0.18\epsilon_8\end{aligned}$$

and so on. That is, $\widehat{y}_1$ differs from the value $\alpha + \beta x_1$, which it is supposed to estimate, by the quantity

$$0.38\epsilon_1 + 0.34\epsilon_2 + 0.2\epsilon_3 + 0.14\epsilon_4 + 0.13\epsilon_5 + 0.07\epsilon_6 - 0.08\epsilon_7 - 0.18\epsilon_8,$$

a linear combination of independent normal errors. The last sum has a normal distribution with mean 0 and variance equal to

$$\begin{aligned}& 0.38^2\text{var}(\epsilon_1) + 0.34^2\text{var}(\epsilon_2) + 0.2^2\text{var}(\epsilon_3) + 0.14^2\text{var}(\epsilon_4) \\ & + 0.13^2\text{var}(\epsilon_5) + 0.07^2\text{var}(\epsilon_6) + 0.08^2\text{var}(\epsilon_7) + 0.18^2\text{var}(\epsilon_8) = \sigma^2 C^2,\end{aligned}$$

with C a constant that you, or the computer, can easily figure out.

The bottom line is that the fitted values are all normally distributed, with means equal to the corresponding $\alpha + \beta x_i$ values, and variances equal to multiples of σ^2 that are easily determined.

(ii) Similarly, the coefficients for the least squares fit are always expressible as linear combinations of the observed y_i's, with multipliers depending only on the x_i values. For the artificial example, the multipliers are given in the next table. In a different least squares problem the multipliers would be different.

	y1	y2	y3	y4	y5	y6	y7	y8
a	0.44	0.4	0.21	0.14	0.13	0.05	−0.12	−0.25
b	−0.75	−0.65	−0.21	−0.04	−0.02	0.17	0.59	0.9

As before, you should interpret the table as meaning

$$a = 0.44y_1 + 0.4y_2 + 0.21y_3 + 0.14y_4 + 0.13y_5 + 0.05y_6 - 0.12y_7 - 0.25y_8$$
$$b = -0.75y_1 - 0.65y_2 - 0.21y_3 - 0.04y_4 - 0.02y_5 + 0.17y_6 + 0.59y_7 + 0.9y_8$$

As with the fitted values $\widehat{y}_i$, the estimator a is normally distributed with mean α, and variance equal to a known multiple of σ^2; and similarly for b.

(iii) How can we estimate the unknown σ^2 in general? Each of the random variables $Z_i = \epsilon_i/\sigma$ has a standard normal distribution. The sum of squares $\sum_i Z_i^2 = \sum_i \epsilon_i^2/\sigma^2$ has a chi-square distribution on n degrees of freedom, with mean value n and standard deviation $\sqrt{2n}$. The random variable $\sum_i \epsilon_i^2$ should be close to $n\sigma^2$. Unfortunately, in general we do not know the ϵ_i, but we can calculate the residuals e_i, which we hope are close to the ϵ_i.

The residuals are slightly more constrained than the ϵ_i's, because of the fitting procedure. We must have $\sum_i e_i = 0$, for otherwise a change in the coefficient a would lead to a smaller sum of squared residuals. [In fact, addition of the mean $\overline{e}$ of the resdiuals to the coefficient a would reduce the sum of squared residuals from $\sum_i e_i^2$ to $\sum_i (e_i - \overline{e})^2$.] Similarly, the sum $\sum_i e_i x_i$ must be zero, for otherwise a small change in the coefficient b would lead to a smaller sum of squared residuals. The two constraints,

$$\sum_i e_i = 0 \qquad \text{and} \qquad \sum_i e_i x_i = 0,$$

imply that all the residuals can be determined once we know $n-2$ of them. As with the chi-square tests from Lecture 8, the constraints reduce the degrees of freedom. The sum of squares $\sum_i e_i^2/\sigma^2$ has chi-square distribution with only $n-2$ degrees of freedom. The random variable $\sum_i e_i^2$ should be close to $(n-2)\sigma^2$. The parameter σ can be estimated by the random variable $s = \sqrt{\sum_i e_i^2/(n-2)}$.

Armed with these facts, let us return to the Minitab output. The first portion was:

Predictor	Coef	StDev	T	P
Constant	2.8115	0.2377	11.83	0.000
x	2.4111	0.4926	4.89	0.003

S = 0.3307 R-Sq = 80.0% R-Sq(adj) = 76.6%

The value $S = 0.3307$ is the estimate of σ. The column headed 'Coef' gives the values for a and b. Under the model, the coefficient b has a $N(\beta, C_1\sigma)$ distribution, where C_1 is a constant that Minitab has calculated, as in (i). The values in the column headed 'StDev' gives the estimated standard deviations. For example, $0.4926 = C_1 s$ is the estimate for $C_1\sigma$. The column headed 'T' is just the ratio 'Coef'/'StDev'. For example, $4.89 = b/(C_1 s)$. If β were zero (just pretend that we don't really know β for the moment), the ratio $b/(C_1\sigma)$ would have a $N(0,1)$ distribution, and the ratio $b/(C_1 s)$ would have a t-distribution on $n-2 = 6$ degrees of freedom. The number 0.003 in the column headed 'P' is the corresponding p-value: it is the probability $\mathbb{P}\{|T| \geq 4.89\}$ for a random variable T with a t-distribution on 6 degrees of freedom.

The p-values can be used to test hypotheses, such as $\beta = 0$, *under the assumption that the model is correct.* That is, the p-value is calculated under the assumption that the y_i are independent $N(\alpha, \sigma)$ random variables.

For a simple, straight line regression, the Analysis of Variance table contains almost the same information as the table just discussed.

Source	DF	SS	MS	F	P
Regression	1	2.6200	2.6200	23.96	0.003
Residual Error	6	0.6562	0.1094		

Total	7	3.2762

I will explain more about the details next week. For the moment just observe that $0.1094 = s^2$ and $23.96 = 4.89^2$, and that the p-value 0.003 is the same as before.

Mintab could also predict the value $\alpha + \beta x_0$, for an x_0 of your choosing. Under the model, the estimator $a + bx_0$ would be normally distributed with mean $\alpha + \beta x_0$ and standard deviation equal to a mutiple $C_0\sigma$ for a constant C_0 that Minitab could calculate. With probability 95%, the interval $a + bx_0 \pm 1.96C_0\sigma$ would contain the true $\alpha + \beta x_0$. If we replaced σ by the estimate s, we would have to increase 1.96 to a value taken from a t-ditribution on 6 degrees of freedom.

If we were trying to predict the behavior of a new value, $y_0 = \alpha + \beta x_0 + \epsilon_0$, with ϵ_0 distributed $N(0, \sigma)$ independently of all the observed data, the width of the interval would have to be increased, because

$$\text{var}(a + bx_0 + \epsilon_0) = \text{var}(a + bx_0) + \text{var}(\epsilon_0) = (C_0^2 + 1)\sigma^2$$

Some of the Section leaders might want to explain this idea further.

2. An ancient example

The method of least squares was invented by Legendre (or maybe by Gauss—it is a matter of dispute) near the start of the nineteenth century, for the solution of problems in astronomy and geodesy.†

Legendre demonstrated his method by a calculation of the "figure of the earth", based on measurement of distances between places of known latitude along a meridean running through France. The calculation was important because it was related to the question of whether the earth bulged slightly more at the equator than a perfect sphere should. The calculation was also closely related to the definition of the kilometer, as one 10,000th of the distance from the equator to the north pole along the meridean through Paris.

A quarter of a century earlier, Roger Boscovitch had developed another method to calculate the ellipticity of the earth (the fraction by which the diameter in the plane of the equator exceeds the diameter through the poles), using the data from the first two columns of the following table. The distances represent the lengths of 1° arcs of latitude, measured in units of toises (1 toise $\approx$ 6.39 feet).

place	latitude	distance	'true' lat	'true' arc
Quito	$0^\circ 0'$	56751	$0^\circ 12'$	56778
Cape of Good Hope	$33^\circ 18'$	57037	$34^\circ 26'$	56958
Rome	$42^\circ 59'$	56979	$41^\circ 54'$	57029
Paris	$49^\circ 23'$	57074	$48^\circ 51'$	57098
Lapland	$66^\circ 19'$	57422	68°	57263

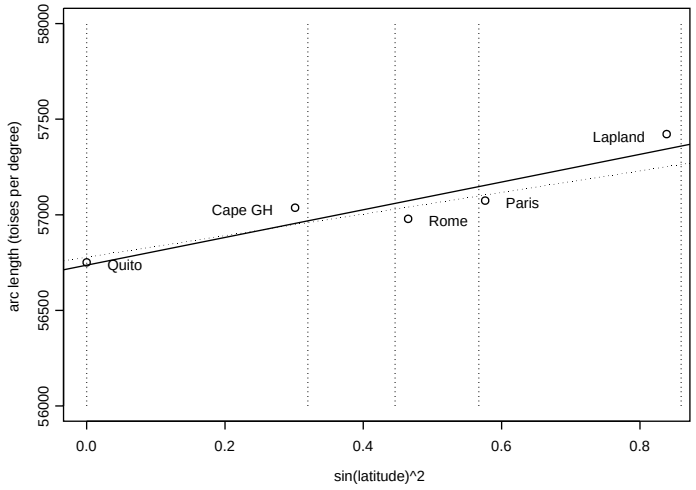
I derived the values in the last two columns by interpolation from a modern atlas (the value for Lapland is of course imprecise, because I don't know which point in Lapland is intended), then application of a theoretical formula using modern values for the diameters of the earth (12,756km in the plane of the equator, 12,714km from pole to pole). The diameter at the equator is about 1.0033 times the diameter at the poles. The earth is (roughly) an oblate spheroid, not a perfect sphere; a cross-section taken through the poles is an ellipse, not a perfect circle.

The geometry of the ellipse gives the relationship,

$$\text{distance} \approx \alpha + \beta \sin^2(\theta) \qquad \text{at latitude } \theta,$$

† The material for this section comes mostly from Chapter 1 of *The History of Statistics: The Measurement of Uncertainty before 1900* by Stephen M. Stigler, Harvard University Press, 1986.

where α denotes the length of a degree of latitude at the equator and $\alpha+\beta$ denotes the length of a degree of latitude at the pole. The constants α and β can be derived from the equatorial and polar diameters. I calculated $\alpha = 56777.6$ and $\beta = 564.5$. The ratio of diameters should be close to $1+\beta/(3\alpha) \approx 1.0033$.



The picture shows the theoretical relationship as a sloping dotted line, with the vertical dotted lines indicating the modern values (as best as I could determine them) for $\sin^2$(latitude). The solid line shows the least squares fit. Of course Boscovitch couldn't have used a computer to fit the least squares line—least squares hadn't even been invented when he studied the problem. If he had been able to, he would have seen some output like:

	Value	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	56736.4856	80.4995	704.8056	0.0000
ss2	724.6624	155.3664	4.6642	0.0186

Residual standard error: 97.13 on 3 degrees of freedom
Multiple R-Squared: 0.8788

He could then have estimated the ratio of diameters as

$$1 + 725/(3 \times 56736) \approx 1.004$$

Notice that α and β, as calculated from modern data, are both within one (estimated) standard deviation of their estimated values:

$$\frac{56736-\alpha}{80} \approx -0.5 \qquad \text{and} \qquad \frac{725-\beta}{155} \approx 1.0$$

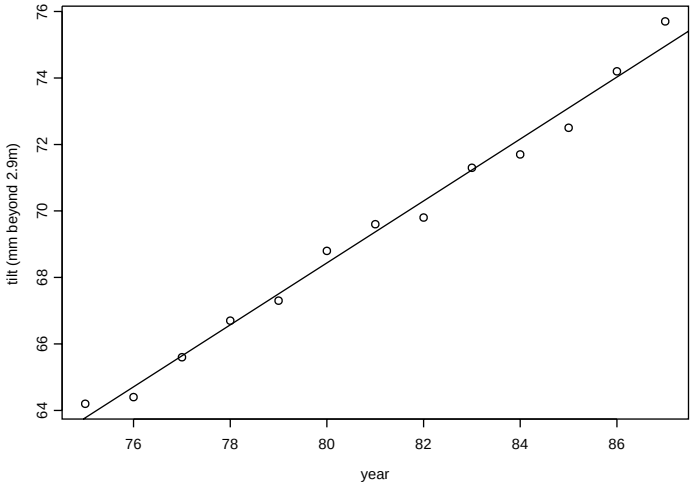
If we believe that the earth is a spheroid, then we know that the plot of arc length per degree versus $\sin^2$(latitude) should be linear, and that departures from the theoretical straight line should be attributed to measurement error. Should we take the estimated standard errors, and the r^2 value, seriously? [They are calculated assuming a specific probabilty model for the data. If the model is a poor approximation, then the calculations have little meaning.] Is it reasonable to treat the errors in measurement as independent $N(0,\sigma)$ random variables, perhaps with some vague appeal to a central limit effect as justification? [It is standard practice to do so, unless examination of diagnostics makes the assumption implausible. With only five observations it is rather difficult to make any serious enquiry into distributional assumptions about errors, but you might try looking

at the residuals.] Should we worry about the errors in the determination of latitudes? [We should, but the details are a bit too complicated for this course.]

3. The leaning tower of Pisa

Exercises 10.8 through 10.10 of M&M (pages 697–698) concern data on the amount on lean in the famous tower over time, during the twentieth century. The lean (or tilt) is given in tenths of a millimeter. The picture shows only the data from 1975 to 1987, with the least squares line for those years superimposed.

year	lean
(18)	(71)
75	642
76	644
77	656
78	667
79	673
80	688
81	696
82	698
83	713
84	717
85	725
86	742
87	757



The output from Mintab seems to indicate a good straight line fit for the data.

Minitab output (slightly edited):

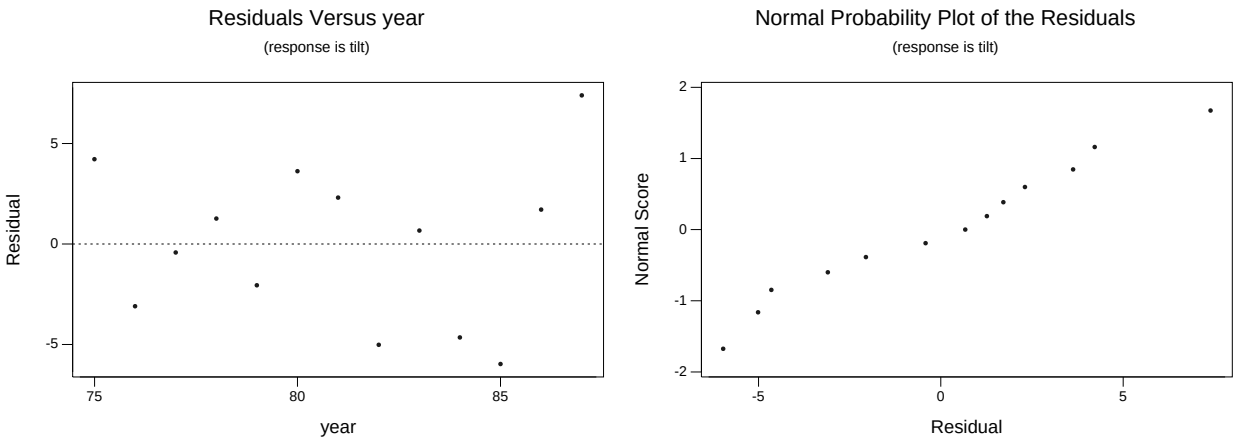
The regression equation is $y = -61.1 + 9.32x$

Predictor	Coef	StDev	T	P
Constant	-61.12	25.13	-2.43	0.033
x	9.3187	0.3099	30.07	0.000

S = 4.181 R-Sq = 98.8% R-Sq(adj) = 98.7%

ANALYSIS OF VARIANCE

Source	DF	SS	MS	F	P
Regression	1	15804	15804	904.12	0.000
Residual Error	11	192	17		
Total	12	15997			



What do you think? Do the residuals look roughly like normal noise? [It might be better to scale the residuals by estimates of their standard errors before we wonder whether they look like normal noise.] Is that a pattern in the residuals that I see, or is my imagination just working too hard at finding patterns? Are you inclined to accept the description of the data for the lean over time as (independent) normal noise superimposed on a linear trend?

I know too little about physics to say whether a linear fit is predicted by some grand theory about the behavior of heavy objects sitting on soft ground. It does not surprise me that, over a short time period, the lean increases at a roughly linear rate—any smooth function is roughly linear over short intervals. [It would also not have surprised me to see a series of jumps following periods of little activity, by analogy with the way earthquakes periodically release pent up strains within the crust.]

It would amaze me greatly if the rate of increase in the lean were constant over a long stretch of time. That is, it would be amazing to see a straight line fitting well to the data over a very long period. For example, I know that various attempts have been made in the past to stop the progressive leaning, with bad effects in some cases.

If the object of the exercise were to predict what might happen in a few more years, it seems reasonable to project the linear relationship out a little beyond 1987. I can have faith that whatever was causing the roughly linear association between tilt and time will keep acting similarly for a little while, without buying into the idea that the linear trend is a “true explanation” for what is going on.