Are these residuals just noise?

Malcolm Hooper, retired from Normanhurst Boys High School

Background. I am not a statistician. I needed a statistical test to answer the title question. Having searched widely and unsuccessfully, I attempted the maths myself. Is this work valid?

Statistical fitting minimises the sum of squares of the residuals; I like using $R^2=1-S_{N-P}/S_N$. I recently obtained Wolberg's 2006 textbook, after waiting several months for it. I had derived a result two different ways before finding and then understanding the test in Wolberg, section 3.3; the F distribution on [0, ∞] transforms into mine using R^2 on [0, 1].

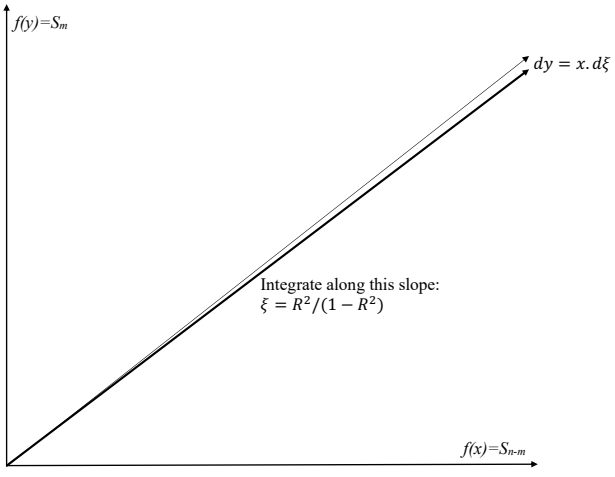
1. Adding just one parameter; my first approach.

I visualized the residuals as defining a point on the unit sphere of $N–P$ dimensions where P parameters have been used to fit N points. The $N–P$ dimensional space of the residuals is orthogonal to the fitting function. Adding one more parameter is equivalent to looking at fraction of the surface in the equatorial band from –R to +R on the unit n-sphere. The probability distribution function was $PDF_{1,n}(R^2) = \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{1}{2})\cdot\Gamma(\frac{n-1}{2})}\frac{(1-R^2)^{\frac{n-3}{2}}}{|R|}$. Monte-Carlo simulations supported this result. Multiple integrations could extend this result to multiple parameters, but it would get messy.

2. Re-imagining the problem; my second approach. Is this valid?

Suppose we are fitting m parameters simultaneously and $n\text{-}m=N\text{-}P\text{-}m$ (usually many) other directions remain in the space of the residuals. The observables are the sum of squares of residuals before the extra parameters are added as $S_n = S_m + S_{n-m}$ and after the extra parameters are added as S_{n-m}. The difference is S_m and $R^2 = \frac{S_m}{S_n} = 1 - \frac{S_{n-m}}{S_n}$ hence $S_m = S_{n-m}\frac{R^2}{(1-R^2)} = S_{n-m}\xi$. Correlation is included and the domain of R^2 is [0, 1].

Working in units where $\sigma = 1$, pure noise should give the sums of squares of residuals as χ^2 distributions. Figure 1 imagines the combined probability distribution function PDF as the χ^2_m distribution in the y direction multiplied by the χ^2_{n-m} distribution in the x direction. The PDF for R^2 is found by integrating along a line of slope $\xi = R^2/(1-R^2)$. We need to weight the integral properly for dR^2.



$$\frac{dR^2}{d\xi}\cdot d\xi = \frac{d\xi}{(1-R^2)^2}$$

Figure 1. As the slope changes from 0 to ∞, the line sweeps out the CDF with S_m and S_{n-m} correlated.

As ξ changes by $d\xi$, it sweeps out the y values as $dy = x\cdot d\xi$.

$$PDF_{m,n}(R^2) = \frac{1}{(1-R^2)^2}\int_0^\infty \chi^2_m(\xi x)\cdot\chi^2_{n-m}(x)\cdot x\cdot dx$$

$$PDF_{m,n}(R^2) = \frac{1}{(1-R^2)^2}\cdot\frac{\xi^{\frac{m}{2}-1}}{2^{m/2}\cdot\Gamma(\frac{m}{2})\cdot 2^{(n-m)/2}\cdot\Gamma(\frac{n-m}{2})}\int_0^\infty e^{-\xi x/2}\cdot e^{-x/2}\cdot x^{(n-2)/2}\cdot dx$$

$$PDF_{m,n}(R^2) = \frac{1}{(1-R^2)^2} \cdot \frac{\xi^{\frac{m}{2}-1}}{2^{n/2} \cdot \Gamma\left(\frac{m}{2}\right) \cdot \Gamma\left(\frac{n-m}{2}\right)} \int_0^\infty e^{-(\xi+1)x/2} \cdot x^{(n-2)/2} \cdot dx$$

$$PDF_{m,n}(R^2) = \frac{1}{(1-R^2)^2} \cdot \frac{\xi^{(m-2)/2}}{2^{n/2} \cdot \Gamma\left(\frac{m}{2}\right) \cdot \Gamma\left(\frac{n-m}{2}\right)} \cdot \left(\frac{2}{\xi+1}\right)^{n/2} \cdot \Gamma\left(\frac{n}{2}\right)$$

$$PDF_{m,n}(R^2) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{m}{2}\right) \cdot \Gamma\left(\frac{n-m}{2}\right)} \cdot (R^2)^{(m-2)/2} \cdot (1-R^2)^{(n-m-2)/2}$$

For m=1, as found previously, $PDF_{1,n}(R^2) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}\right) \cdot \Gamma\left(\frac{n-1}{2}\right)} \frac{\left(1-R^2\right)^{\frac{n-3}{2}}}{|R|}$

The cumulative distribution function *CDF* is found in Lide, 2003, A-36, Eqns 268 & 265. For odd values of n, the *CDF* is found by taking the expression below to sufficient terms, where the last denominator is $n-1$. Substitute $z = 1 - R^2$ to simplify.

$$CDF_{1,n}(R^2) = \sqrt{1-z}\left(1 + \frac{1}{2}z\left(1 + \frac{3}{4}z\left(1 + \frac{5}{6}z\left(1 + \frac{7}{8}z(1+\cdots)\right)\right)\right)\right)$$

For even values of n, the *CDF* is evaluated similarly. The last denominator is $n-1$.

$$CDF_{1,n}(R^2) = \frac{\cos^{-1}(2z-1)}{\pi} + \frac{2\sqrt{z(1-z)}}{\pi}\left(1 + \frac{2}{3}z\left(1 + \frac{4}{5}z\left(1 + \frac{6}{7}z(1+\cdots)\right)\right)\right)$$

Figures 2a and 2b show these *CDF*s and *PDF*s for n in the range 2 to 21.

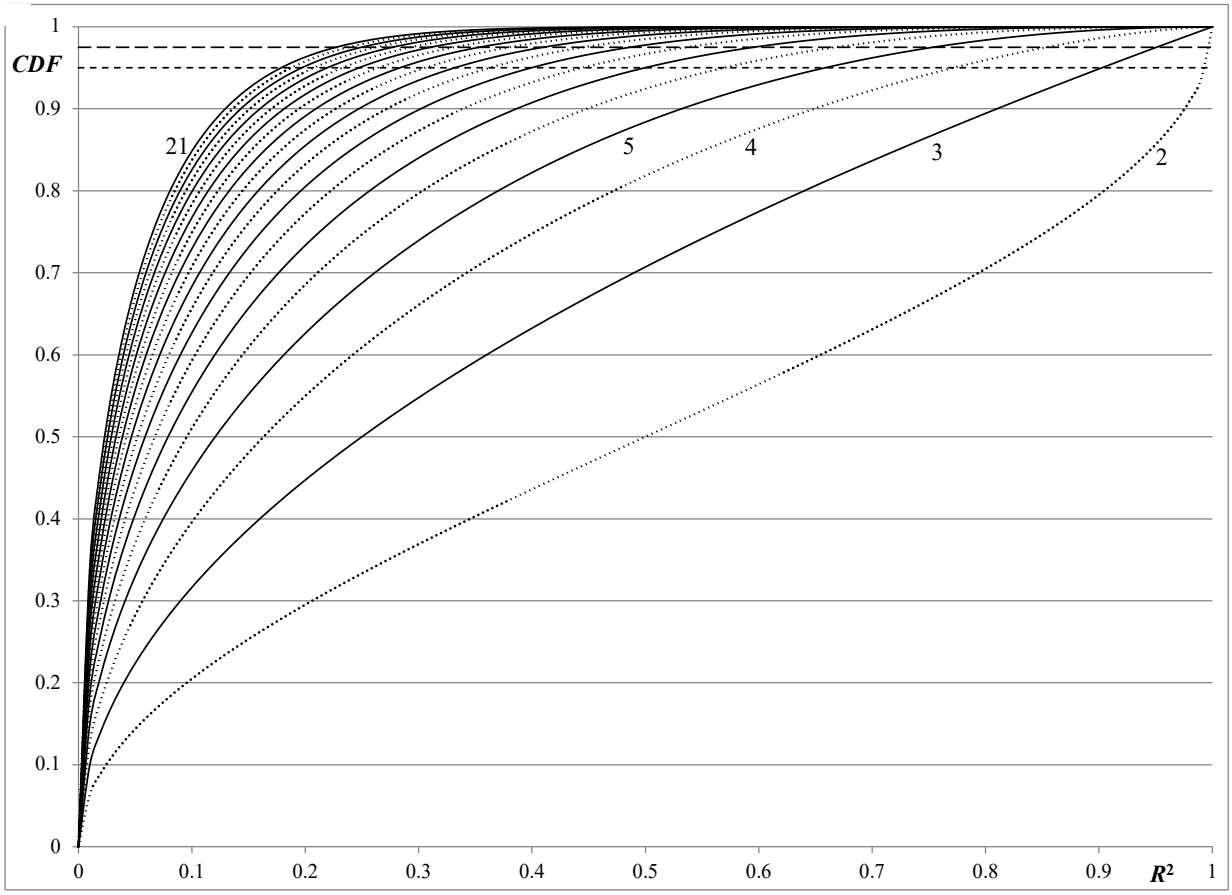


Figure 2a. $CDF_{1,n}$ functions for *m*=1 and *n* from 2 to 21. Even values on n are shown as dotted lines; odd values as solid lines. Cut-off values at *y*=0.95 and *y*=0.975 are shown. These values and more have been collected into Table 3 in the appendix. Low cutoff values are uninformative.

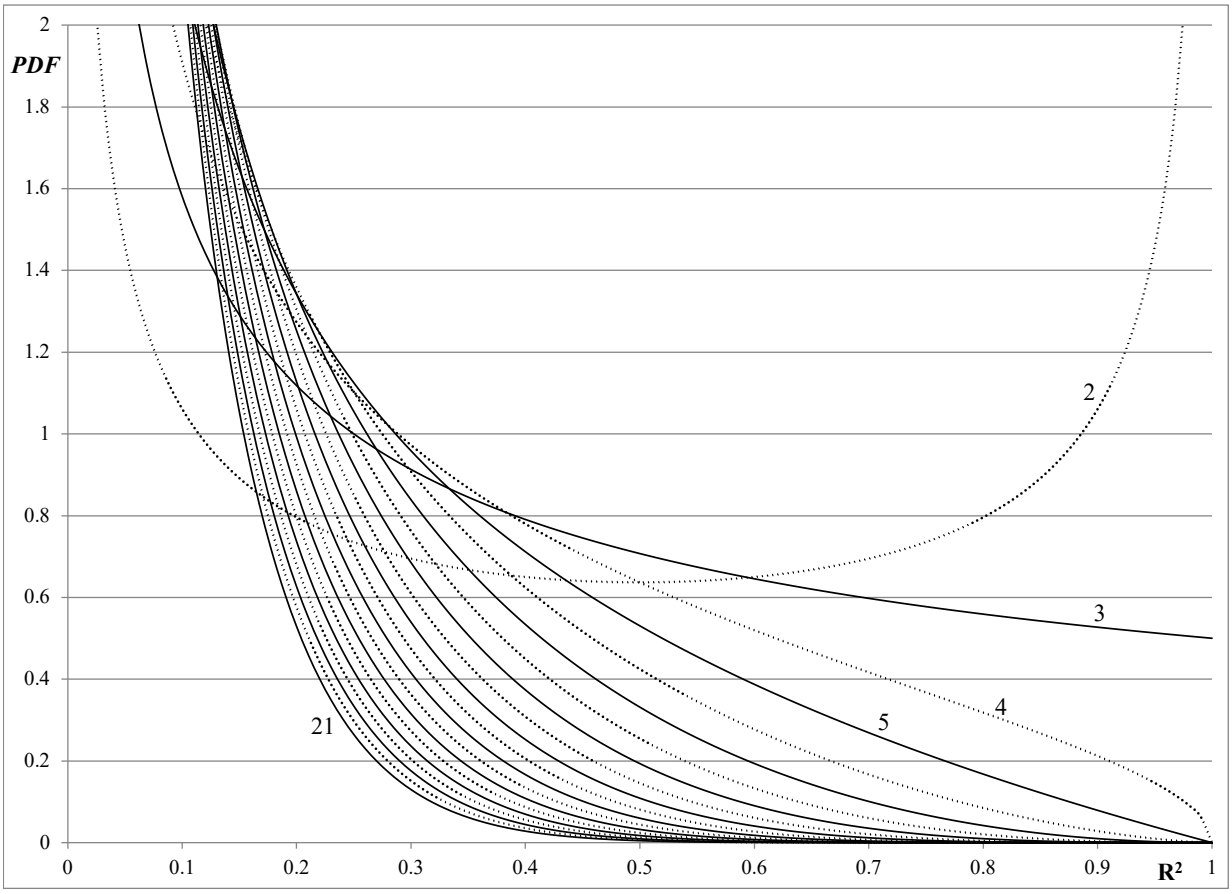


Figure 2b. $PDF_{1,n}$ functions for m=1 with n from 2 to 21. The even values of n are shown as dotted lines; odd values as solid lines.

For m=2, we have $PDF_{2,n}(R^2) = \frac{(n-2)}{2} z^{(n-4)/2}$ and $CDF_{2,n}(R^2) = 1 - z^{(n-2)/2}$

For n = 4, the distribution is uniform on [0, 1]. A chi-squared goodness-of-fit test using 10^5 trials supported this.

The right tail is calculated as $p = z^{(n-2)/2}$ thus $R^2_{crit,2,n} = 1 - p^{2/(n-2)}$. Many situations add two parameters, particularly paired sine and cosine functions. Is this short expression new or useful?

For m=3, $PDF_{3,n}(R^2) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{3}{2}\right)\cdot\Gamma\left(\frac{n-3}{2}\right)} \cdot R \cdot (1-R^2)^{(n-5)/2} = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{3}{2}\right)\cdot\Gamma\left(\frac{n-3}{2}\right)} \cdot \sqrt{1-z} \cdot z^{(n-5)/2}$

The probability for the right tail is a known form (Lide, 2003, A-27, Eqn 127) but I have no urgent need to calculate it exactly.

For m=4, $PDF_{4,n}(R^2) = \frac{(n-2)(n-4)}{4} \cdot \left(z^{(n-6)/2} - z^{(n-4)/2}\right)$

For the right tail, $p = 1 - CDF = \left(\frac{n}{2} - 1\right) z^{(n-4)/2} - \left(\frac{n}{2} - 2\right) z^{(n-2)/2}$

More generally, the mean of the *PDF* is $\int_0^1 x.PDF_{m,n}(x).dx = \frac{m}{n}$, which agrees with symmetry.

Gauss-Jacobi quadrature estimates of the right tail well, especially when p is small and m is even.

3. Transforming from the F-test defined on [1, ∞]

Wolberg's equation 3.3.6, in my notation, is $\frac{S_n}{S_{n-m}} = 1 + \frac{m}{n} F(\alpha, m, n)$ and $R^2 = 1 - \frac{S_{n-m}}{S_n}$.

I have checked the m=1 and m=2 values for *CDF*= 0.90, 0.95, 0.99, and 0.999 for many values of n. The agreement is excellent. In the limit of large n and a given α , the values of R^2/n approach a constant value given by the χ^2 distribution, as expected.

4. Applying these results.

The title question arose while teaching physics to high school students, provoked by two experiments. First, *many* oscillations of a pendulum were timed in a continuous run to determine the period more accurately. A linear plot was expected for elapsed time versus counted oscillations, but the string may have stretched slowly under load, giving quadratic and higher terms. Are the quadratic terms observed in Figure 3 significant? Second, the height of each ring of a freely hanging slinky was measured from bottom to top. The intention was to practise using larger data sets and fit to other than a straight line. Hooke's law predicts a parabolic plot for distance against number of rings. Higher order terms occur; are they significant?

Sets of orthogonal polynomials arise naturally in the stepwise analysis of the residuals, as seen in the parabolas in Figure 3. With N points, a polynomial of order N–1 fits perfectly. After fitting the pendulum data to a line, the low order terms are absent from the residual space.

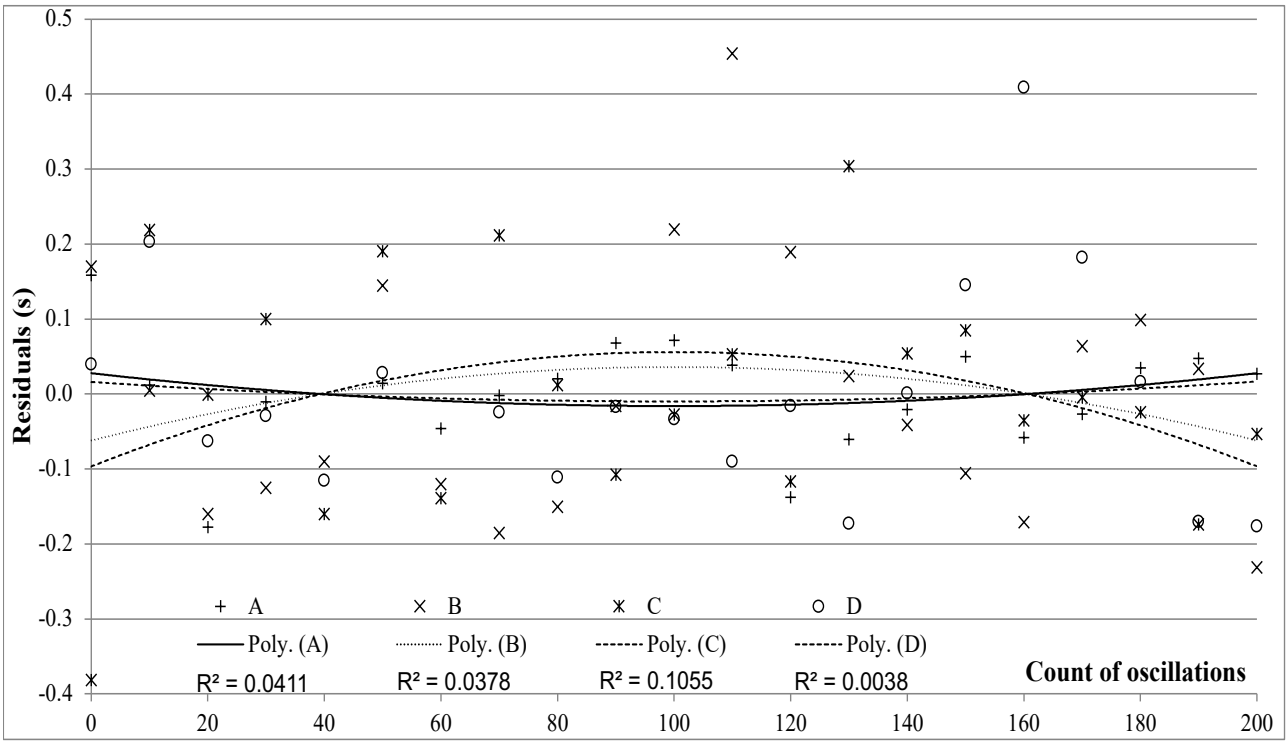


Figure 3. Residuals from timing a pendulum, by four students. Each data set fitted well to a line, with R^2 = 0.99999993, 0.99999964, 0.99999969, & 0.99999974. After projecting out the good fit, we have $\sum_{i=2}^{N-1} R_i^2 = 1$ for the residuals. This equation defines a unit spherical surface in a space of N–2 dimensions. The constant (i = 0) and linear (i = 1) terms are orthogonal to these residuals, so not part of this space.

After linear fitting for 21 points, we have m=1 and n–m=18. A p-value of 0.025 has R^2 = 0.2493. All four student results are below this and support the null hypothesis. These residuals are just noise; the pendulum string is not stretching under load. Many other student data sets agree, as do the parabolas curving both up and down.

The slinky result was more interesting. For the quadratic fit, R^2 = 0.9999925. Figure 4a plots the residuals and shows an obvious pattern, close to a scaled and shifted Legendre cubic. The equation appears to have four parameters, but the constant, linear and quadratic terms are all orthogonal to these residuals; this is a one parameter fit. For n–m=48 and R^2 = 0.8525, $p = 1.4\times10^{-21}$; this signal is real. I think I can estimate or explain this term; it's engineering, the bottom of the slinky rotates.

Figure 4b shows the residuals after removing the cubic; no *visible* pattern is obvious. The raw data were recorded to the nearest half millimeter. The quartic polynomial lies entirely within this limit. This signal might be real; p = 0.012. Three more trials (with more points) produced quartic results of the same sign and similar magnitude. Round-off error, assuming a uniform distribution and thus

calculated as $\Sigma\{(0.5\ \text{mm})^2/12)$, accounts for 74% of the sum of residuals squared. Polynomials of orders five and six were consistent with noise.

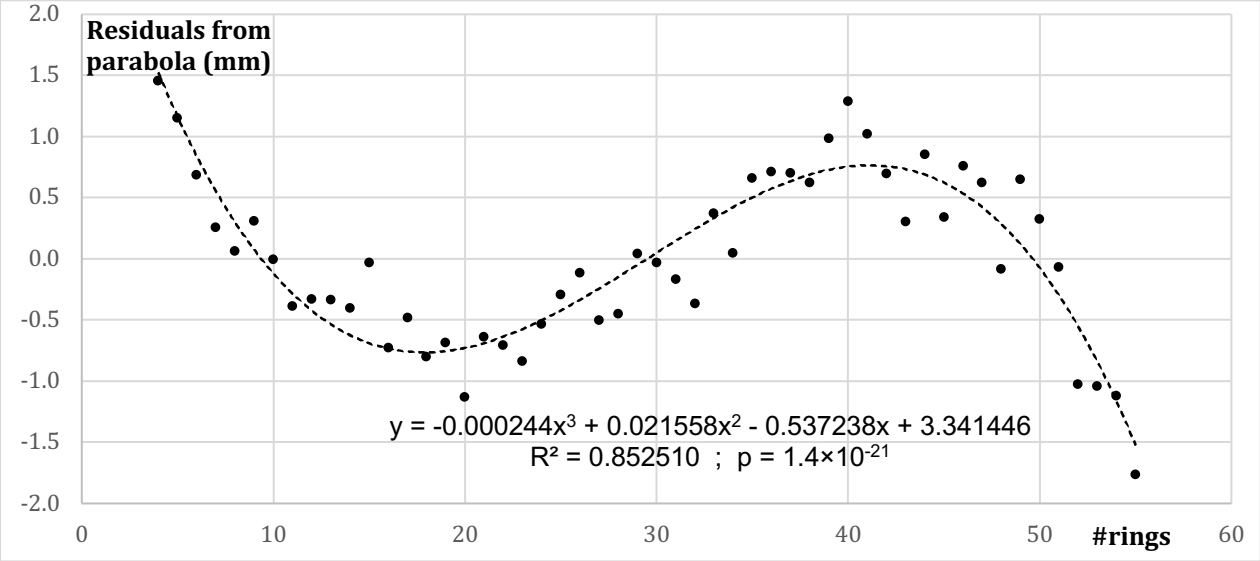


Figure 4a. Residuals from the quadratic fit to the freely hanging slinky. This pattern is obviously "not noise".

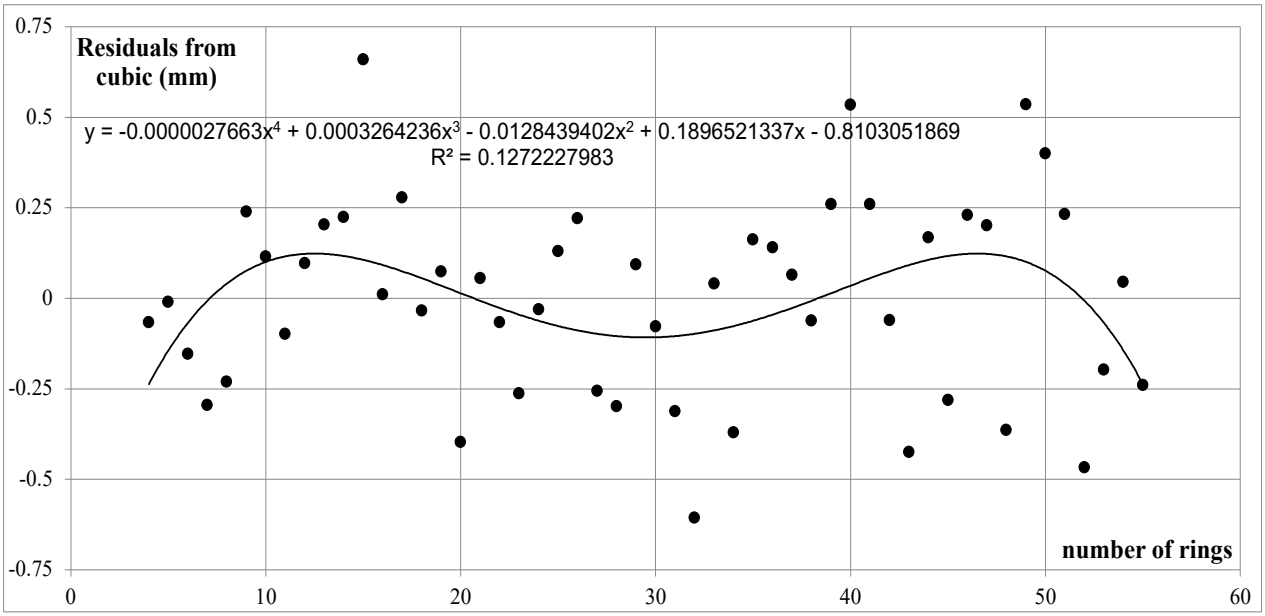


Figure 4b. Residuals from the cubic fit to the freely hanging slinky.

Another example: climate data

The Bureau of Meteorology website gave average annual rainfall and temperature data for several airports near Australian state capitals and some islands. These data were plotted against year and straight lines were fitted. R^2 measures the linear term after removing the average. This test is more sensitive than a t-test comparing the first and second halves of the data, which I had done before.

For all maximum temperatures and most minimum average annual temperatures, the results indicated highly significant warming trends. Their cause may be argued but their existence is clear. Islands were included to counter any "urban heat island" argument.

Most rainfall data are much noisier; significant trends were not found. Perth *is* drying; Macquarie Island near Antarctica is getting wetter.

Further calculations examined linear trends using monthly data; finer patterns were discerned. Perth's rainfall is winter dominant, mainly in June; the decrease of June rainfall was highly

significant ($p = 4.5\times10^{-6}$). Sydney's minimum and maximum temperatures rose significantly for every month while rainfall changes were not significant for any month.

		Ann. Rain (mm)		Min. Temp. (˚C)		Max. Temp. (˚C)	
Location	**BoM #**	R^2	p	R^2	p	R^2	p
Darwin	014015	0.06134	0.0249	0.07252	0.0157	0.39985	3.1E-10
Brisbane (Amberley)	040004	0.01873	noise	0.07116	0.0175	0.26437	1.3E-06
Sydney	066037	0.00177	noise	0.76745	2.2E-27	0.56285	3.2E-16
Melbourne (Laverton)	087031	0.07657	0.0130	0.44193	2.4E-11	0.32549	4.0E-08
Hobart	094008	0.05558	noise	0.49027	1.3E-10	0.37697	7.6E-08
Adelaide	023034	0.02131	noise	0.58595	5.0E-14	0.32231	5.9E-07
Perth	009021	0.22150	1.4E-05	0.28836	4.0E-07	0.44865	2.0E-11
Alice Springs	015590	0.01037	noise	0.00633	noise	0.33634	1.3E-05
Norfolk Island	200288	0.08415	7.4E-04	0.40600	9.2E-11	0.28433	2.1E-07
Macquarie Island	300004	0.35797	4.8E-08	0.21811	3.2E-05	0.28237	1.4E-06

Table 2. Analysis of climate data to 2021 for several sites around Australia, with their Bureau of Meteorology site numbers, where data for 60 to 80 years were available. No such data set near Canberra was found. The long baselines should reduce the effect of decadal cycles such as ENSO, the Indian Ocean dipole, and the 11-year solar oscillation. The R^2 values are for a linear fit.

When quadratic trends of the average annual temperatures were examined, the R^2 values were usually small, indicating just noise. At Sydney Airport, the rate of warming is increasing; for maximum temperatures $p = 0.0009$ and for minimum temperatures $p = 0.0011$. Calculations for Sydney Airport using monthly data for maximum and minimum temperatures had the parabolas concave up, but some terms were negligibly small.

Summary.

Statistical testing supplements rather than replaces plotting the residuals. My approach to the problem appears to be equivalent to using an F-test, but is a different way of seeing the problem.

The noise due to measuring and rounding data is a second constraint to consider when asking "are these residuals just noise?"

References

Lide, D. R. (2003) Editor-in-Chief, *CRC Handbook of Chemistry and Physics*, 84th Edn., CRC Press, Boca Raton

Wolberg, J. (2006) *Data Analysis Using the Method of Least Squares*, Springer-Verlag, Berlin

Monthly and annual temperature and rainfall data were found at:
http://www.bom.gov.au/climate/averages/tables/cw_200440.shtml Accessed 17 Feb 2022

Weights and points for Gauss-Jacobi quadrature were found at:
http://keisan.casio.com/has10/SpecExec. cgi?id=system/2006/1329114617 Accessed 17 Feb 2022