COMPUTING PRESCRIPTIONS

Editor: Isabel Beichl, isabel.beichl@nist.gov

FITTING NATURE'S BASIC FUNCTIONS PART II: ESTIMATING UNCERTAINTIES AND TESTING HYPOTHESES

By Bert W. Rust

IN THE LAST ISSUE, WE CONSIDERED THE LINEAR STATISTICAL MODEL

$$\mathbf{y} = \Phi\alpha^* + \in, \qquad E(\in) = \mathbf{0}, \qquad E(\in\in^T) = \Sigma^2. \tag{1}$$

Here, $\mathbf{y}$ is an m-vector of measurements, α^* is an n-vector of unknown parameters (with $n \le m$), Φ is the $m \times n$ least-squares matrix (with linearly independent columns that do not depend on α^*), and $\in$ is an m-vector of random errors with expected value $\mathbf{0}$ and $m \times m$ positive definite variance matrix Σ^2. When Σ^2 is not known, we usually assume that

$$\Sigma^2 = \sigma^2 \mathbf{I}_m, \tag{2}$$

where $\mathbf{I}_m$ is the mth-order identity matrix, and σ^2 is an unknown constant variance. In this case, the best linear unbiased estimate (BLUE) for α^* is

$$\hat{\alpha} = \left[\Phi^T\Phi\right]^{-1}\Phi^T\mathbf{y}, \tag{3}$$

which we can compute without knowing σ^2.

The development in Part I was motivated by global annual average temperature data (see http://cdiac.esd.ornl.gov/trends/temp/jonescru),[3] which Figure 1 plots as discrete circles. The two curves are the best-fitting first- and fifth-degree polynomials,

$$\phi(t,\alpha) = \sum_{\nu=1}^{n} \alpha_\nu (t - t_0)^{\nu-1}, \quad n = 2, 6, \tag{4}$$

where $t_0 = 1856.0$. The fifth-degree polynomial tracks the data better, but we need more statistical analysis to determine whether the improvement obtained justifies the addition of four new free parameters. This is one of the questions that we address in this installment. Several texts cover all the statistical material here, including the two classics[1,2] cited in Part I (see last issue).

Simple diagnostics for the fit

The (minimal) sum of squared residuals for the BLUE,

$$\mathrm{SSR} = \sum_{i=1}^{m} \hat{r}_i^2 = [\mathbf{y} - \Phi\hat{\alpha}]^T[\mathbf{y} - \Phi\hat{\alpha}], \tag{5}$$

is the most fundamental diagnostic. For the two fits in Figure 1, $(\mathrm{SSR})_2 = 2.783993$ and $(\mathrm{SSR})_6 = 1.678972$. The latter is smaller than the former, but this does not necessarily mean that the fifth-degree polynomial is a better model than the straight line. It explains more of the total variance in the record, but some of that total properly belongs to the errors $\in$, and a model with too many free parameters might capture variance that should be relegated to the residuals.

A measure of the variance assigned to the model is the *coefficient of determination*,

$$R^2 = 1 - \frac{\mathrm{SSR}}{\mathrm{CTSS}}, \tag{6}$$

where

$$\mathrm{CTSS} = \sum_{i=1}^{m} (y_i - \bar{y})^2, \quad \text{with} \quad \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i. \tag{7}$$

CTSS measures the total variation of the measurements about their mean. Because SSR measures their variation around the fit, it follows that R^2 is the fraction of the total variance explained by the model. For the fits in Figure 1, $R_2{}^2 = 0.6179$ and $R_6{}^2 = 0.7696$, so the straight line and the fifth-degree polynomial respectively explain 61.79 percent and 76.96 percent of the total variance.

Uncertainties in the estimates

The uncertainties in the $\hat{\alpha}_j$ depend on the distribution of the random errors $\in_i$. Given $\Sigma^2 = \sigma^2\mathbf{I}_m$, we must estimate a variance matrix for $\hat{\alpha}$. From Equations 1 and 3, we see that $E(\hat{\alpha}) = \alpha^*$, so the matrix we want is

$$\mathbf{V}(\hat{\alpha}) = E\left[(\hat{\alpha} - \alpha^*)(\hat{\alpha} - \alpha^*)^T\right]. \tag{8}$$

It is straightforward to substitute Equation 3 into Equation 8 and use the assumptions in Equations 1 and 2 to show that

$$\mathbf{V}(\hat{\alpha}) = \sigma^2[\mathbf{\Phi}^T\mathbf{\Phi}]^{-1}. \tag{9}$$

In Part I, we noted that the old method of computing $\hat{\alpha}$ produced this inverse matrix as a byproduct. Newer subroutines, based on the QR factorization, often do not return this important result, even though doing so would require little extra effort. Because **Q** is an orthogonal matrix,

$$\mathbf{\Phi} = \mathbf{Q}\begin{bmatrix}\mathbf{R}\\ \mathbf{O}\end{bmatrix} \Rightarrow [\mathbf{\Phi}^T\mathbf{\Phi}]^{-1} = \mathbf{R}^{-1}(\mathbf{R}^{-1})^T, \tag{10}$$

and inverting the upper triangular matrix **R** is easy. Unfortunately, none of the least-squares subroutines in Linpack, Lapack, or Matlab takes this extra step.

Equation 9 depends also on σ^2. When that value is not known, the expression

$$\hat{\sigma}^2 = \frac{1}{m-n}\sum_{i=1}^{m}\hat{r}_i^2 = \frac{\text{SSR}}{m-n} \tag{11}$$

gives an unbiased estimate for it. Using that estimate in Equation 9 gives the estimated variance matrix

$$\hat{\mathbf{V}}(\hat{\alpha}) = \frac{\text{SSR}}{m-n}[\mathbf{\Phi}^T\mathbf{\Phi}]^{-1}. \tag{12}$$

The diagonal elements $\hat{V}_{j,j}$ are variances for the corresponding $\hat{\alpha}_j$, so $\pm 1\sigma$ intervals are defined by

$$\left[\hat{\alpha}_j - \sqrt{\hat{V}_{j,j}}\ ,\ \ \hat{\alpha}_j + \sqrt{\hat{V}_{j,j}}\right],\quad i = 1, 2, \ldots, n. \tag{13}$$

I used this formula to compute the uncertainties given in Equations 12 and 29 in Part I.

For the straight-line fit, the ratios

$$\frac{|\hat{\alpha}_1|}{\sqrt{\hat{V}_{1,1}}} = 19.6 \quad \text{and} \quad \frac{|\hat{\alpha}_2|}{\sqrt{\hat{V}_{2,2}}} = 15.2 \tag{14}$$

indicate a high improbability that $\hat{\alpha}_1^* = 0$ or that $\hat{\alpha}_2^* = 0$. Most scientists would agree that you should regard with suspicion any measurement whose magnitude is not more than three standard deviations greater than zero. For the fifth-degree polynomial fit, the ratios

$$\frac{|\hat{\alpha}_j|}{\sqrt{\hat{V}_{j,j}}} = 8.8,\quad 4.3,\quad 5.2,\quad 5.7,\quad 6.0,\quad 6.3 \tag{15}$$

suggest that all the coefficients are statistically significant, a fact that we will confirm later.

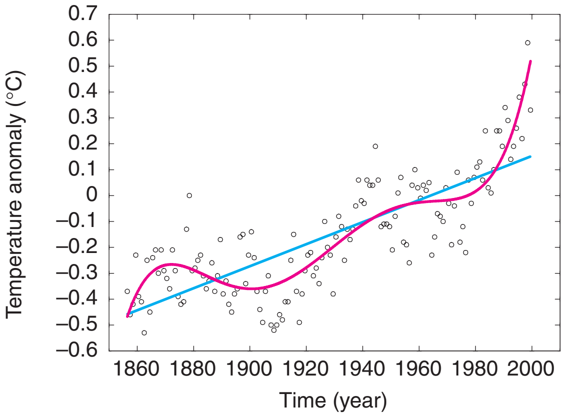

Figure 1. The straight line and fifth-degree polynomial fits to global yearly average temperature data. The plotted circles are temperature anomalies obtained by subtracting the mean temperature (14° C) for the years 1961 to 1990 from the actual measured averages.

Estimate correlations

The off-diagonal elements of $\hat{\mathbf{V}}(\hat{\alpha})$ also contain important information. Specifically, $\hat{V}_{i,j} = \hat{V}_{j,i}$ estimates the covariance between $\hat{\alpha}_i$ and $\hat{\alpha}_j$. Covariance relationships are clarified by computing the *correlation matrix*

$$\hat{C}_{i,j} = \frac{\hat{V}_{i,j}}{\sqrt{\hat{V}_{i,i}}\sqrt{\hat{V}_{j,j}}},\quad i,j = 1, 2, \ldots, n. \tag{16}$$

Each $\hat{\alpha}_i$ is perfectly correlated with itself, so $\hat{C}_{i,i} = 1.0$, and for the off-diagonal elements, $-1.0 \le \hat{C}_{i,j} \le 1.0$, with $\hat{C}_{i,j} = 0.0$ representing complete absence of correlation. Cross-correlations with $|\hat{C}_{i,j}| \to 1.0$ indicate that the columns of Φ are almost linearly dependent, which suggests that the model has more parameters than the data can support. It could mean that the model is wrong, or it could mean only that we need more data to reliably determine the parameters. And sometimes we can reduce high correlations by transforming the independent variable.

For the straight-line fit, $\hat{C}_{1,2} = -0.866$, which is totally unremarkable. The largest $|\hat{C}_{i,j}|$ for the fifth-degree polynomial was at $\hat{C}_{5,6} = -0.995$, which is remarkable, but the ratios given in Equation 15 for $\hat{\alpha}_5$ and $\hat{\alpha}_6$ indicate that the high correlation is not a problem. In fact, for these data, the cross-correlations are sensitive to the choice of t_0. Choosing t_0 = 1928.0 rather than t_0 = 1856.0 gives $\hat{C}_{5,6} = 1.97 \times 10^{-16}$. Choosing t_0 = 0000.0 gives $\hat{C}_{5,6} = -0.999993$. All three values give fitted curves and residuals that are graphically indistinguishable.

Assigning confidence levels

So far we have made no assumptions about the probability distribution for ∈. Henceforth, we will use the most common assumption, which is that the errors are independently, identically normal: $\in \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_m)$. Therefore, it is easy to show that

$$\hat{\alpha}_j \sim n\left(\alpha_j^*, V_{j,j}\right) \;\Rightarrow\; \frac{\hat{\alpha}_j - \alpha_j^*}{\sqrt{V_{j,j}}} \sim n(0,1), \tag{17}$$

where $V_{j,j}$ is the jth diagonal element of $\mathbf{V}(\hat{\alpha})$. If we know σ^2, we can use the second expression in Equation 17 with a table of the standard normal distribution to construct confidence intervals for the α_j^*.

Let $(1 - p)$ be the desired confidence level (p = 0.05 for 95 percent confidence), and let $f_n(u)$ be the $n(0, 1)$ probability density function, which is symmetric about $u = 0$. Then, for any positive κ,

$$\Pr\left\{-\kappa < \frac{\hat{\alpha}_j - \alpha_j^*}{\sqrt{V_{j,j}}} < \kappa\right\} = \int_{-\kappa}^{\kappa} f_n(u)du. \tag{18}$$

Standard normal tables usually tabulate the quantity

$$F(\kappa) = \int_{-\infty}^{\kappa} f_n(u)du \tag{19}$$

as a function of κ, so

$$\int_{-\kappa}^{\kappa} f_n(u)du = F(\kappa) - F(-\kappa) = 2F(\kappa) - 1. \tag{20}$$

We want to choose κ so that

$$2F(\kappa) - 1 = 1 - p \;\;\Rightarrow\;\; F(\kappa) = 1 - p/2. \tag{21}$$

If κ_p is the corresponding number from the table, then

$$\Pr\left\{-\kappa_p < \frac{\hat{\alpha}_j - \alpha_j^*}{\sqrt{V_{j,j}}} < \kappa_p\right\} = 1 - p, \tag{22}$$

from which it follows that

$$\Pr\left\{\hat{\alpha}_j - \kappa_p\sqrt{V_{j,j}} \; < \; \alpha^* \; < \; \hat{\alpha}_j + \kappa_p\sqrt{V_{j,j}}\right\} = 1 - p. \tag{23}$$

The most commonly used values for p and κ_p are

$$\begin{array}{c|ccc} p & 0.3333 & 0.05 & 0.01 \\ \hline \kappa_p & 0.967 & 1.960 & 2.576 \end{array}. \tag{24}$$

When σ^2 is not known, we must estimate the $V_{j,j}$ by using Equation 12 and replace the normal distribution with the student's t-distribution with $m - n$ degrees of freedom. These probability density functions are also symmetric about the origin, so the construction of confidence intervals is almost the same as with the normal distribution. The only difference is the table from which we get the value of κ. For small values of $m - n$, the t-distributions are flatter than the normal, so the confidence intervals will be wider. As $m - n \to \infty$, the t-distributions approach the normal distribution, and the widening becomes insignificant for $m - n \geq 120$. The temperature record has m = 144, so we can use the normal distribution to construct confidence intervals. In fact, the $\pm 1\sigma$ intervals in Equations 12 and 29 of Part I are 68.3 percent confidence intervals.

For the fifth-degree polynomial, the estimate with the greatest relative uncertainty was $\hat{\alpha}_2 = 0.0336 \pm .0078$. Let's construct a 99 percent confidence interval for this estimate. Taking p = 0.01, we have κ_p = 2.576, so Equation 23 gives

$$\Pr\{0.0135 < \alpha_2^* < 0.0537\} = 0.99, \tag{25}$$

so, it is highly unlikely that $\alpha_2^* = 0$.

Testing hypotheses

Formally testing a hypothesis to see if we should omit some terms in a linear model requires a comparison of the sums of squared residuals obtained by fitting both the full and reduced models. Write the full model in the partitioned form

$$\mathbf{y} = (\Phi_1, \Phi_2)\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \in, \qquad \in \sim N(\mathbf{0}, \sigma^2\mathbf{I}_m), \tag{26}$$

where α_2 is a k-vector of parameters to be omitted, and Φ_2 is the $m \times k$ submatrix of corresponding columns. We can always write the full model in this form by interchanging the matrix columns and the corresponding parameters. We want to test the *null hypothesis*

$$H_0: \; \alpha_2^* = \mathbf{0}, \tag{27}$$

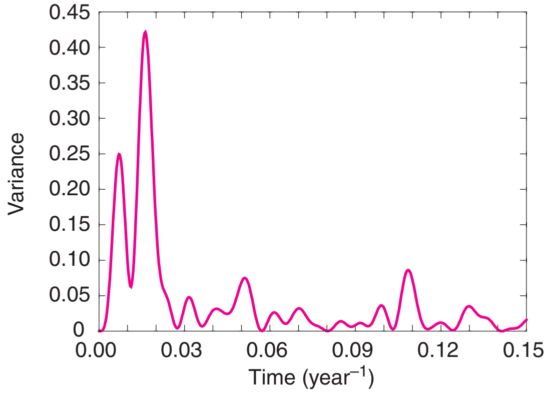

Figure 2. The truncated Fourier power spectrum of the straight-line residuals. We discarded frequencies between 0.15 and 0.5 yr^{-1} to expand the low frequencies.

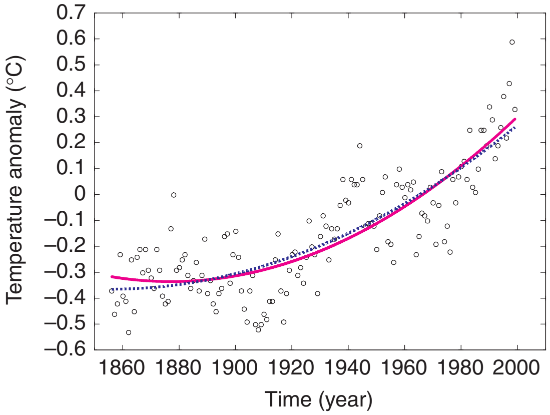

Figure 3. Quadratic and reduced quadratic fits to the global yearly average temperature anomalies. The solid curve is the full quadratic, and the dashed curve is the reduced quadratic (Equation 33).

so we fit the full model to get $(\mathrm{SSR})_F$ and the reduced model,

$$\mathbf{y} = \Phi_1 \alpha_1 + \in, \quad \in \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_m) \tag{28}$$

to get $(\mathrm{SSR})_H$, which will be larger than $(\mathrm{SSR})_F$. The *F-test* is based on the ratio

$$u = \frac{(\mathrm{SSR})_H - (\mathrm{SSR})_F}{(\mathrm{SSR})_F} \cdot \frac{m-n}{k}, \tag{29}$$

which is compared with a percentage point from a table of the $F(k, m-n)$ distribution—the *F-distribution with* k *and* m – n *degrees of freedom*. The tables usually give the percentage points for a given probability level on a single page. So the page for probability p will tabulate the percentile F_p as a function of k and $m - n$. The values of p and F_p are related by

$$\int_0^{F_p} f_{k,m-n}(\xi) d\xi = p, \tag{30}$$

where $f_{k,m-n}(\xi)$ is the probability density function for the $F(k, m-n)$ distribution. If the computed u exceeds the value F_p from the table, then the probability of obtaining the reduction $(\mathrm{SSR})_H - (\mathrm{SSR})_F$ by chance is less than $1 - p$, so we must reject the null hypothesis. Put another way, the SSR reduction obtained by including the α_2 terms is statistically significant at the $100p$ percent level. Conversely, if $u < F_p$, then we should accept null hypothesis.

For the fifth-order polynomial, $m - n = 138$ and $(\mathrm{SSR})_F = 1.678972$. The null hypothesis is

$$H_0 : \alpha_3^* = \alpha_4^* = \alpha_5^* = \alpha_6^* = 0, \tag{31}$$

so $k = 4$, and $(\mathrm{SSR})_H = 2.783993$, which gives $u = 22.70629$. To test at the 95 percent level, we get the percentile $F_{0.95}(4,138) = 2.4373$, which is much smaller than u, so we reject the null hypothesis.

A time series diagnostic

We selected the fifth-order polynomial because it had just enough "wiggle" to accommodate the quasicycle in the straight-line residuals. Figure 2 is an unwindowed periodogram estimate of the power (variance) spectrum for those residuals. The dominant peak is at a frequency of 0.0163 yr^{-1}, which corresponds to a period of 61.4 years. The unanticipated secondary peak is at frequency 0.007 yr^{-1}, which corresponds to a period of 143 years. Because $m = 144$, this peak does not represent a real cycle but rather indicates a nonlinear baseline.

Figure 3 shows a quadratic polynomial fit to the data as a solid curve. The estimated coefficients were

$$\begin{aligned} \hat{\alpha}_1 &= -0.314 \pm 0.031 \text{ °C}, \\ \hat{\alpha}_2 &= -0.0018 \pm 0.0010 \text{ °C / yr}, \\ \hat{\alpha}_3 &= (4.21 \pm 0.67) \times 10^{-5} \text{ °C / yr}^2, \end{aligned} \tag{32}$$

with $\mathrm{SSR}_3 = 2.173416$. Clearly, $\hat{\alpha}_1$ and $\hat{\alpha}_3$ are beyond suspicion, but what about $\hat{\alpha}_2$? Fitting the reduced quadratic model

$$\phi(t,\alpha) = \alpha_1 + \alpha_3 (t - t_0)^2 \tag{33}$$

gave

$$\begin{aligned} \hat{\alpha}_1 &= -0.363 \pm 0.016 \text{ °C}, \\ \hat{\alpha}_3 &= (3.03 \pm 0.17) \times 10^{-5} \text{ °C / yr}^2, \end{aligned} \tag{34}$$

with $\mathrm{SSR}_2 = 2.224553$, which is almost as small as SSR_3. Figure 3 shows the fit as a dashed curve. I will leave it as an exercise for you to show that the F-test accepts the null hypothesis

$$H_0 : \ \alpha_2^* = 0 \tag{35}$$

at the 95 percent level of significance.

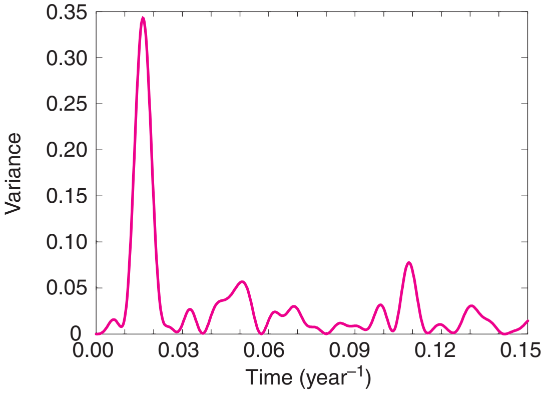

Figure 4. The truncated Fourier power spectrum of the reduced quadratic residuals.

Figure 4 shows the periodogram for the reduced quadratic residuals. The single dominant peak is still at frequency 0.0163 yr^{-1}. Remarkably, the fit gave $R^2 = 0.6947$ with the same number of free parameters as the straight-line fit, which gave $R^2 = 0.6179$. The full quadratic fit, with one additional free parameter, gave $R^2 = 0.7017$, which is not a statistically significant improvement. Significantly, the reduced quadratic mode is monotonically increasing. All this suggests a model of the form

$$\phi(t,\alpha) = \alpha_1 + \alpha_2(t - t_0)^2 + \alpha_3 \sin\left[\frac{2\pi}{\alpha_4}(t + \alpha_5)\right], \qquad (36)$$

with a period of $\alpha_4 \approx 61.4$ years. We will use nonlinear least squares to fit this model in the next installment.

References

1. A.M. Mood and F.A. Graybill, *Introduction to the Theory of Statistics*, McGraw-Hill, New York, 1963.
2. F.A. Graybill, *Theory and Application of the Linear Model*, Duxbury Press, North Scituate, Mass., 1976.
3. P.D. Jones et al., "Global and Hemispheric Temperature Anomalies: Land and Marine Instrumental Records," *Trends: A Compendium of Data on Global Change*, Carbon Dioxide Information Analysis Center, Oak Ridge Nat'l Laboratory, Oak Ridge, Tenn., 2000.

Bert W. Rust is a research mathematician at the National Institute of Standards and Technology. His research interests include ill-posed problems, time-series modeling, nonlinear regression, and observational cosmology. He received his BS in engineering physics and MS in mathematics from the University of Tennessee and his PhD in astronomy from the University of Illinois. Contact him at the Nat'l Inst. of Standards and Technology, 100 Bureau Dr., Stop 8910, Gaithersburg, MD 20899-8910; bwr@cam.nist.gov.