ON FISHER'S RAINFALL PROBLEM[1]

"The Influence of Rainfall on the Yield of Wheat at Rothamsted"

By D. S. Robson BU-22-M June, 1951

Problem 1

Given: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + \varepsilon_i \qquad i=1,\ldots,n \qquad \varepsilon_1,\ldots,\varepsilon_n \perp\!\!\!\perp N(0,\sigma)$ [2]

with the least squares solutions

$$Y_i = \bar{y} + \sum_{j=1}^{q} c_j (x_{kj} - \bar{x}_j) \quad , \qquad R^2 = \frac{\sum_1^n (Y_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

Required: Find Pr ($R > R$) under the zero hypothesis

Solution: Let $P = (y_1 - \bar{y}, \ldots, y_n - \bar{y})$, and drop a norm PQ from P to the hyperformed by the q centralized x-vectors $x_1 - \bar{x}_1, \ldots, x_q - \bar{x}_q$. Then the squared length of PQ is

$$\|PQ\|^2 = \sum_1^n (y_i - Y_i)^2$$

and the squared distance from the origin to Q is

$$\|OQ\|^2 = \sum_1^n (Y_i - \bar{y})^2$$

so that

$$\cos^2 QOP = \frac{\|OQ\|^2}{\|OP\|^2} = \frac{\sum_1^n (Y_i - \bar{y})^2}{\sum_1^n (y_i - \bar{y})^2} = R^2$$

and $PQ \perp OQ$.

Hence, [3] $\sum_1^n (Y_i - \bar{y})^2 = \sigma^2 \chi^2_{q-1} \perp\!\!\!\perp \sum_1^n (y_i - Y_i)^2 = \sigma^2 \chi^2_{n-q}$

---

[1] This paper is not self contained; it is meant to be read only in conjunction with Fisher's article.

[2] $\perp\!\!\!\perp$ means "independent."

[3] A proof of this would involve an orthogonal transformation of the x's.

Now, to simplify the notation, take

$$a = \Sigma(Y_i - \bar{y})^2 \quad , \qquad \tau = \Sigma(y_i - \bar{y})^2$$

then the joint frequency function of (a,τ) is

$$g(a,\tau) = \frac{1}{2^{\frac{q}{2}}\sigma^{\frac{q}{2}}\Gamma(\frac{q}{2})} a^{\frac{q}{2}-1} e^{-\frac{a}{2\sigma^2}} \frac{1}{2^{\frac{n-q-1}{2}}\sigma^{n-q-1}\Gamma(\frac{n-q-1}{2})} \tau^{\frac{n-q-1}{2}-1} e^{-\frac{\tau}{2\sigma^2}}$$

$$= \frac{1}{2^{n-1}\sigma^{n-1}\Gamma(\frac{q}{2})\Gamma(\frac{n-q-1}{2})} a^{\frac{q}{2}-1} \tau^{\frac{n-q-1}{2}-1} e^{-\frac{a+\tau}{2\sigma^2}}$$

then take $z = \frac{a}{\tau}$

so that $\Pr(z < t) = \Pr(\frac{a}{\tau} < t) = \Pr(a < \tau t)$

but $\Pr(a < \tau t) = \int_0^{\infty} \int_0^{\tau t} g(a,\tau)\, da\, d\tau$.

Take $a = uv$, $\tau = v$, then $\frac{\delta(a,\tau)}{\delta(u,v)} = v$, and

$\Pr(a < \tau t) = \Pr(u < t)$

$$= \frac{1}{2^{n-1}\sigma^{n-1}\Gamma(\frac{q}{2})\Gamma(\frac{n-q-1}{2})} \int_0^{t} \int_0^{\infty} (uv)^{\frac{q}{2}-1} v^{\frac{n-q-1}{2}-1} e^{\frac{-v(u+1)}{2\sigma^2}} v\, dv\, du$$

$$= \frac{1}{2^{n-1}\sigma^{n-1}\Gamma(\frac{q}{2})\Gamma(\frac{n-q-1}{2})} \int_0^{t} \frac{\Gamma(\frac{n-1}{2})}{(\frac{u+1}{2\sigma^2})^{\frac{n-1}{2}}} u^{\frac{q}{2}-1}\, du$$

$$= \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{q}{2})\Gamma(\frac{n-q-1}{2})} \int_0^t \frac{u^{\frac{q}{2}-1}}{(u+1)^{\frac{n-1}{2}}}\, du \,.$$

Hence,

$$h(z) = \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{q}{2})\Gamma(\frac{n-q-1}{2})} \frac{z^{\frac{q}{2}-1}}{(z+1)^{\frac{n-1}{2}}}$$

but $R^2 = \frac{z}{1+z}$, $1-R^2 = \frac{1}{1+z}$, and $dR^2 = \frac{dz}{(1+z)^2}$

hence,

$$f(R^2)dR^2 = \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{q}{2})\Gamma(\frac{n-q-1}{2})} (R^2)^{\frac{q}{2}-1} (1-R^2)^{\frac{n-q-1}{2}-1} dR^2 \,.$$

Fisher notes that we must have $r_{ij} \neq 1$ for $i \neq j$, $i,j = 1,\ldots,q$, which is another way of saying that we must have linear independence.

$$E(R^2) = \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{q}{2})\Gamma(\frac{n-q-1}{2})} \int_0^1 (R^2)^{\frac{q}{2}} (1-R^2)^{\frac{n-q-1}{2}-1} dR^2$$

$$= \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{q}{2})\Gamma(\frac{n-q-1}{2})} \cdot \frac{\Gamma(\frac{q}{2}+1)\Gamma(\frac{n-q-1}{2})}{\Gamma(\frac{n+1}{2})}$$

$$= \frac{\frac{q}{2}}{\frac{n-1}{2}} = \frac{q}{n-1}$$

Note: $$V(R^2) = \frac{2q(n-q-1)}{(n-1)^2(n+1)}$$

The fr. f. of R is

$$f(R) = \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{q}{2})\Gamma(\frac{n-q-1}{2})} (R^2)^{\frac{q}{2}-1} (1-R^2)^{\frac{n-q-1}{2}-1} 2R$$

$$= \frac{2\Gamma(\frac{n-1}{2})}{\Gamma(\frac{q}{2})\Gamma(\frac{n-q-1}{2})} \; R^{q-1} \; (1-R^2)^{\frac{n-q-1}{2}-1}$$

Fisher, however, chooses to work with my $z^{\frac{1}{2}}$, so take as before

$$R^2 = \frac{z}{1+z} \text{ , or } z = \frac{R^2}{1-R^2}$$

$$w = z^{\frac{1}{2}} , \quad 2wdw = dz$$

then

$$p(u) = \frac{2\Gamma(\frac{n-1}{2})}{\Gamma(\frac{q}{2})\Gamma(\frac{n-q-1}{2})} \quad \frac{w^{q-1}}{(1+w^2)^{\frac{n-1}{2}}}$$

and

$$\Pr(\underline{R} > R) = \Pr\left[w > (\frac{R^2}{1-R^2})^{\frac{1}{2}}\right]$$

$$= \frac{2\Gamma(\frac{n-1}{2})}{\Gamma(\frac{q}{2})\Gamma(\frac{n-q-1}{2})} \int_{(\frac{R^2}{1-R^2})^{\frac{1}{2}}}^{\infty} \frac{w^{q-1}}{(1+w^2)^{(n-1)/2}} \, dw$$

which gives Fisher's expansions for even and odd q.

Q.E.D.

Problem 2

Given: A system of observations,

$$\begin{matrix} y_1 & x_{11} & \cdots & x_{q1} \\ \vdots & \vdots & & \vdots \\ y_m & x_{1m} & \cdots & x_{qm} \end{matrix}$$

where y_j = yield of wheat in bu./acre at Rothamsted in year j

x_{ij} = inches of rainfall at Rothamsted during the i'th period of the j'th year, where the q time periods are of equal length.

Required: To form a prediction equation for the mean yield

$$\bar{w} = c + a_1 r_1 + \cdots + a_q r_q$$

where r_i = inches of rainfall in the i'th interval of time

a_i = increment in yield, bu./acre, per inch of rainfall in the i'th interval of time.

Solution: An interval of time, T, is divided into q equal subintervals. Let $q \to \infty$, then r(t) becomes a continuous function of time t and a(t) becomes a continuous function of time t, $0 \leq t \leq T$, and $\bar{w} = c + \int_0^T ar\, dt$.

Let $T_0, T_1, T_2, \cdots, T_n, \cdots$ be a set of normal orthogonal functions of time t in the region $0 \leq t \leq T$, and compute the Fourier coefficients of the continuous function r(t) associated with this system of $\perp$ functions:

$$r(t) = \sum_0^\infty \rho_i T_i$$

where $\rho_i = \int_0^T r\, T_i\, dt$

since $rT_i = \rho_0 T_0 T_i + \cdots + \rho_i T_i T_i + \cdots$

$$\int_0^T r\, T_i\, dt = \sum_{j=0}^{\infty} \rho_j \int_0^T T_j T_i\, dt$$

and $$\int_0^T T_i T_j = \begin{cases} 0 \text{ for } i \neq j \\ 1 \text{ for } i = j \end{cases} .$$

Similarly, compute the Fourier coefficients

$$\alpha_i = \int_0^T a\, T_i\, dt$$

then take the Cauchy product ar and integrate term by term:

$$ar = \alpha_0\rho_0 T_0^2 + \alpha_0\rho_1 T_0 T_1 + \alpha_1\rho_0 T_1 T_0 + \alpha_0\rho_2 T_0 T_2 + \cdots$$

$$\int_0^T ar\, dt = \alpha_0\rho_0 \int_0^T T_0^2 dt + \alpha_0\rho_1 \int_0^T T_0 T_1 dt + \cdots$$

$$= \alpha_0\rho_0 + \alpha_1\rho_1 + \alpha_2\rho_2 + \cdots$$

so that

$$\bar{w} = c + \alpha_0\rho_0 + \alpha_1\rho_1 + \alpha_2\rho_2 + \cdots$$

The discrete analogue to the preceding argument involves a set of normal orthogonal integral functions $T_0, T_1, T_2, \cdots$ where *

$$\sum_{t=1}^{n} T_i(t) T_j(t) = \begin{cases} 0 \text{ for } i \neq j \\ 1 \text{ for } i = j \end{cases}$$

and $$\rho_i = \sum_{t=1}^{n} r(t)\, T_i(t)$$

* Fisher uses n and q interchangeably.

$$\alpha_i = \sum_{t=1}^{n} a(t)\, T_i(t)$$

Fisher, however, works with a centralized t-variate, thus his

$$t_1 = 1 - \frac{1}{n} \sum_{i=1}^{n} i = 1 - \frac{n+1}{2}$$

$$t_2 = 2 - \frac{1}{n} \sum_{i=1}^{n} i = 2 - \frac{n+1}{2}$$

$$\vdots$$

$$t_n = n - \frac{1}{n} \sum_{i=1}^{n} i = n - \frac{n+1}{2}$$

$$\sum_{i=1}^{n} t_i = \Sigma i - \Sigma i = 0$$

and his $\sum_{1}^{n} t_i^2 = \frac{n(n^2-1)}{12}$ (see BU-6-M "Methods of Curvilinear Regression").

The rationale behind his particular choice of orthogonal functions is adequately demonstrated in the article; he first discusses the simplest case of all, that of an unchanging series of independent values $x_1, \ldots, x_n$ (where x_i = inches of rainfall during the i'th period of the year; this is the unreal case in which rainfall is independent of the time of year). If we suppose that the first term in the expansions, viz., $\rho_0 T_0$ and $\alpha_0 T_0$ are satisfactory measures of the ratio of rainfall and increment of yield then the best estimate of $r(t_i)$ is

(1) $$r(t_i) = \bar{x}, \quad i=1, \ldots, n$$

and we must have

$$\rho_0 T_0(t_i) \equiv \bar{x}$$

but

$$\sum_{1}^{n} T_{0_i}^2 = 1 \quad \text{by orthogonality}$$

and, since T_{0_i} is constant for all i,

$$\sum_{1}^{n} T_{0_i}^2 = n\, T_{0_i}^2 = 1$$

or $$T_{0_i} = \frac{1}{\sqrt{n}}$$

Then in the next simplest case, the case where we presume rainfall rate to change linearly with time, the least squares estimate of r_i is

(2) $$r_i = \bar{x} + bt_i$$

$$= \rho_0 T_{0_i} + \rho_1 T_{1_i}$$

hence, $$\rho_1 = \sum_{1}^{n} r_i T_{1_i} = \bar{x} \Sigma\, T_{1_i} + b\Sigma\, t_i\, T_{1_i}$$

but

(3) $$\Sigma\, T_{1_i} T_{0_i} = \frac{1}{\sqrt{n}} \Sigma\, T_{1_i} = 0$$

so that

$$\rho_1 = b\Sigma\, t_i\, T_{1_i}$$

and from (2),

$$\rho_1 T_{1_i} = b\, t_i$$

so that

$$b\, T_{1_i} \sum_{1}^{n} t_i\, T_{1_i} = bt_i$$

$$T_{1_i} \Sigma t_i\, T_{1_i} = t_i$$

(4) $$T_{1_i} = \frac{t_i}{\Sigma t_i T_{1_i}} \quad \Big| \quad \square, \Sigma$$

$$\Sigma T_{1_i}^2 = \frac{\Sigma t_i^2}{(\Sigma t_i T_{1_i})^2}$$

$$1 = \frac{\Sigma t_i^2}{(\Sigma t_i T_{1_i})^2}$$

$$\Sigma t_i T_{1_i} = \sqrt{\Sigma t_i^2}$$

hence, from (4)

$$T_{1_i} = \frac{t_i}{\sqrt{\Sigma t_i^2}} = \frac{t_i}{\sqrt{\frac{n(n^2-1)}{12}}} = \sqrt{\frac{12}{n(n^2-1)}}\ t_i$$

And, proceeding in this manner, one could construct $T_2, \ldots, T_5$ to conform with the expressions given by Fisher. Note that this choice of T_0 and T_1 makes:

$$\rho_0 = \frac{\sum_{1}^{n} x_i}{\sqrt{n}} \quad (= x_1' \text{ by Fisher's notation})$$

$$\rho_1 = \frac{\Sigma x_i t_i}{\sqrt{\Sigma t_i^2}} \quad (= x_2' \text{ " " " })$$

In terms of matrices, we are constructing an orthogonal matrix T which transforms x to x', thus

$$\begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} T_{0_1} & \cdots & T_{0_n} \\ \vdots & & \vdots \\ T_{n_1} & \cdots & T_{n_n} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

where $$x_1' = \Sigma T_{0_i} x_i = \frac{1}{\sqrt{n}} \Sigma x_i \qquad (= \rho_0)$$

$$x_2' = \Sigma T_{1_i} x_i = \frac{1}{\sqrt{\Sigma t_i^2}} \Sigma t_i x_i \qquad (= \rho_1)$$

and (see BU-6-M "Methods of Curvilinear Regression")

$$x_3' = \Sigma T_{2_i} x_i = \frac{1}{\sqrt{\Sigma(t_i^2 - \frac{1}{n}\Sigma t_i^2)}} (\Sigma x_i t_i^2 - \bar{x}\Sigma t_i^2)$$

$$= \sqrt{\frac{180}{n(n^2-4)(n^2-1)}} (\Sigma x_i t_i^2 - \bar{x}\,\Sigma t_i^2) \qquad (= \rho_2)$$

Fisher remarks that $E(x_1') = \sqrt{n}\, E(x)$, and this is seen from

$$E(x_1') = E\left(\frac{\Sigma x}{\sqrt{n}}\right) = \frac{1}{\sqrt{n}} \sum_1^n E(x) = \sqrt{n}\, E(x) \;;$$

he also notes that

$$V(x_1') = V(x)$$

since

$$V(x_1') = V\left(\frac{\Sigma x}{\sqrt{n}}\right) = V\left(\frac{n\bar{x}}{\sqrt{n}}\right) = nV(\bar{x}) = V(x)$$

and that

$$E(x_i') = 0 , \qquad i \neq 1$$

since

$$E(x_i') = \sum_{j=1}^n T_{ij}\, E(x_j)$$

where

$$\sum_j T_{ij} = 0 , \qquad i \neq 1 \qquad \text{by (3)}$$

and that

$$V(x_i') = V(x) \qquad i=1, \ldots, n$$

since

$$V(x_i') = \sum_{j=1}^n T_{ij}^2\, V(x) = V(x)$$

He writes $(n-p)\mu_1^2 = \sum_{p+1}^{n} x_i'^2$

and this is seen from the fact that

(i) in the case of an unchanging mean, i.e., $r_i = \bar{x}$, $i=1,\ldots,n$

$$\sum_1^n x_i^2 = \frac{1}{n}\left(\sum_1^n x_i\right)^2 + \sum_1^n (x_i-\bar{x})^2$$

but

$$\frac{1}{n}\left(\sum_1^n x_i\right)^2 = x_1'^2$$

so that

$$\sum_1^n (x_i-\bar{x})^2 = \sum_2^n x_i'^2 = (n-1)\, s^2$$

because by $\perp$, $\sum_1^n x_i^2 = \sum_1^n x_i'^2$

(ii) in the case of a linearly changing mean, i.e., $r_i = \bar{x} + bt_i$

$$\sum_1^n x_i^2 = \frac{1}{n}(\Sigma x_i)^2 + \sum_1^n (r_i-\bar{x})^2 + \sum_1^n (x_i-r_i)^2$$

$$= \frac{1}{n}(\Sigma x_i)^2 + \frac{1}{\Sigma t_i^2}(\Sigma x_i t_i)^2 + (n-2)s_{x\cdot t}^2$$

$$= x_1'^2 + x_2'^2 + (n-2)s_{x\cdot t}^2$$

so that

$$(n-2)s_{x\cdot t}^2 = \sum_3^n x_i'^2$$

and so on.

Fisher gives a table of results of an application of his orthogonal polynomials for x = yield, against t = time in years. He took (retaining the previous notation)

$$r_i = \rho_0 T_{0_i} + \rho_1 T_{1_i} + \cdots + \rho_5 T_{5_i}$$

and tested the set of hypotheses : $\rho_1=0, \ldots, \rho_5=0$

He used the average squared residual from the fifth degree polynomial as his "error term", thus:

$$\frac{\rho_1}{s} = \frac{x_1'}{\mu} = \frac{\sqrt{n-6}\,\Sigma x_i t_i}{\sqrt{\Sigma t_i^2 \Sigma (x_i - r_i)^2}} = r_{xt}\sqrt{\frac{n-6}{1-R^2}} = -0.82$$

Note that this is but a slight variation of the method Snedecor uses in testing successively higher degree polynomials; Snedecor would use the error term

$$s^2 = \frac{\Sigma(x_i - \bar{x} - bt_i)^2}{n-2} \qquad (= \mu_1)$$

to get

$$F_{1,(n-2)} = \frac{x_1'^2}{\mu_1^2} = (n-2)\,\frac{r_{xt}^2}{1-r_{xt}^2}$$

or

$$t_{n-2} = \frac{x_1'}{\mu_1} = r_{xt}\sqrt{\frac{n-2}{1-r_{xt}^2}}$$

However, Fisher prefers to use μ_5 as error. Because of the large number of degrees of freedom in this example, Student's t is nearly normal; so Fisher uses the standard normal t as his test distribution. Such a table will give

$$\Pr\left(\left|\frac{x_1'}{\mu}\right| > 0.82\right) = .4122$$

Returning now to the original problem, Fisher computes the numerical values for a', b', ..., f' which are, but for constant multipliers, numerical values for $\rho_0, \rho_1, \ldots, \rho_5$. These coefficients are calculated for each year by the method of summation (see BU-6-M "Methods of Curvilinear

Regression"). There remains the problem of estimating $\alpha_0, \alpha_1, \ldots, \alpha_5$; this is accomplished by computing the least squares regression coefficients for the regression of yield on $a', \ldots, f'$ and then converting the regression coefficients back to $\alpha_0, \ldots, \alpha_5$ (an implicit conversion). The a_i's are computed as

$$a_i = \alpha_0 T_{0_i} + \ldots + \alpha_5 T_{5_i}$$

Fisher does not give the final prediction equation

$$\bar{w} = c + a_1 r_1 + \ldots + a_n r_n \quad .$$