Department of Economics
University of Southampton
Southampton SO17 1BJ
UK

Discussion Papers in Economics and Econometrics

2000

This paper is available on our website
http://www/soton.ac.uk/~econweb/dp/dp00.html

The Origins of Fixed X Regression

John Aldrich
Department of Economics
University of Southampton
Southampton
SO17 1BJ
UK
Fax 023 80593858
E-mail jca1@soton.ac.uk

Abstract
In 1922 R. A. Fisher introduced the fixed X regression model, synthesising the regression theory of Pearson and Yule with the least squares theory of Gauss. The innovation was based on Fisher's realisation that the distribution associated with the regression coefficient was unaffected by the distribution of X. Subsequently Fisher interpreted the fixed X assumption in terms of his notion of ancillarity. This paper considers these developments against the background of early twentieth century statistical theory.

April 2000

Introduction

In fitting equations to observational data it is routine to use fixed x regression, treating the values of the explanatory variables as non-random, though they vary randomly. The practice was introduced in the 1920s by R. A. Fisher (1890-1962). There were two mathematically distinct theories with different domains of application: an older univariate theory (the "theory of errors") and a newer multivariate theory (the "theory of correlation"). Regression belonged with correlation until Fisher re-located it to the theory of errors.

This paper describes the circumstances in which fixed x regression was established and considers the inferential scheme(s) where it found a justification. Fisher's distribution theory has been celebrated but less has been made of the re-location. The historians Seal (1967) and Hald (1998) see a work of restoration. Hald (p. 616) writes that the statisticians who worked on correlation theory "did not realise that their regression analysis was a version of the linear model and the linear estimation theory" and Seal (p. 16) that "with Fisher (1922) the sampling theory of regression equations returned to the Gaussian model". But the theory could not return to where it had never been–this was a new synthesis. Fisher, however, obscured his own contribution–in the 20s by not explaining it and in the 50s by overlaying it with polemics against Neyman and Karl Pearson.

Section 1 sketches the theory of errors and the theory of correlation as they were understood in Britain at the beginning of the twentieth century; the great days of the first seemed to be over and the future to be with the second. The story proper begins with Slutsky and Karl Pearson and their goodness of fit tests for regression curves (Section 2). Fixed x regression appears when Fisher (1922) corrected their work and at the same time introduced the t-test for regression coefficients (Sections 3 and 4). Sections 5 and 6 survey the new regression presented in *Statistical Methods for Research Workers* (1925) and the continuing work on the

old. Fixed x regression was founded on the claim that the distribution of the test statistics is not affected by the randomness–or otherwise–of the $x's$. M. S. Bartlett (born 1910), the other major contributor to fixed x regression, made good the claim in his "Theory of Statistical Regression" (1933).

In the later 30s a new claim appears grounded in the "theory of estimation": the $x's$ provide "ancillary information" about the regression coefficients. Regression did not fit into Fisher's earliest theory of estimation but an accommodation became possible after 1934, when the theory took a conditional turn with ancillarity. (Sections 7 and 8). Bartlett in 1936 was the first to link regression to conditional inference (Section 9) while Fisher discussed regression in relation to ancillarity in letters of 1939/40. Section 10 considers how Fisher applied the theory of estimation to regression and Section 11 how he made regression serve in the campaign against "repeated sampling from the same population". Finally there are some comments on the obscure history of everyday regression.

1 Theory of errors & theory of correlation

We need some background on the theory of errors and the theory of correlation. For fuller accounts see Seal (1967), Stigler (1986), Hald (1998) and Farebrother (1999). As the sequel is largely a British story, the emphasis is on the British background–Hald (chapter 27) describes a richer continental scene.

The "Gaussian model" of the theory of errors was devised for measurement problems in astronomy. Dynamical theory provides equations $\mu = X\beta$ between the non-stochastic μ and X with β a vector of unknown quantities; the elements of X are observed, but not those of μ; a vector of measurements y deviates from μ by a vector of unobserved errors, ε, distributed

$N(0, \sigma^2 I)$.[1] Estimation is by least squares–the maximum posterior value from a uniform prior; the estimation of the normal mean was a special case. For Pearson and Fisher–our protagonists–this combination of specification and estimation procedure *was* the theory of errors and *was* Gauss; see Aldrich (1997, pp. 162-4). Other contributors did not figure and Gauss's second specification (without normality and treated by the Gauss-Markov theorem) did not impinge on Fisher's thinking, though it was revived in his time by Aitken (1935) and Neyman (1934) and came to be taught with the normal specification as "the linear model".[2]

In the 19th century the theory of errors was also applied to fitting "empirical formulae". Merriman's (1884/1911) textbook (the least squares reference for Pearson and Yule) has many examples. In Merriman the x values are either time trends or quantities selected by the scientist–e.g. recorded depth and depth squared in an equation for water velocity (p. 131).[3] Surprisingly, perhaps, the application of least squares to observational data appears to have begun only in the correlation era. In 1871 Jevons (p. 141, note h) proposed using on price and quantity data the least squares method Bessel had used for trigonometric trends but when economists, such as Moore (1914 & -17), came to fit equations they followed Pearson.

The theory of errors was useful to Galton (1877 & -88) when he formalised his ideas on reversion/regression and correlation but as Karl Pearson (1857-1936) developed these subjects they moved away from the theory. Pearson (1896) presents regression and correlation as aspects of the multinormal distribution. His inference formulae for multinormal regression coefficients were the same–though in different notation–as those developed by Gauss for least squares. For Pearson this was no more than a coincidence as the theories had different subject

[1] The language and notation are modern. Aldrich (1998) discusses varieties of least squares notation.

[2] For Fisher's background in least squares, see Aldrich (1997, pp. 162-3) and for references to the revival, Aldrich (1999). Aitken came from the actuarial branch of trend fitting.

[3] In all Merriman's examples with several right hand side variables, the variables are all functions of a single underlying variable. *Multiple* regression did not come naturally! See Aldrich (1997a).

matter; the x and y of the Gauss model are *not* correlated. In one theory multinormality appears for the (posterior) distribution of β, in the other for the distribution of the observables; see Pearson (1920, pp. 25-7). Pearson's inference technique–see Pearson & Filon (1898, pp. 194-6)–adapted from the theory of errors was a kind of large sample Bayesian theory.[4] Pearson was slow to become involved in the small sample research begun by W. S. Gosset–'Student' (1908a)–and propelled by Fisher (1915). By then Pearson wanted to escape from the restrictions of the normal scheme, not perfect its inference theory.

Yule (1897) extended the linear regression specification to cases of skew correlation and used least squares. Stigler (1986, pp. 348-53) has a throrough treatment of this important development and writes (p. 345) of a *second* "least squares synthesis". This is more plausibly ascribed to Fisher–at least on the theory side. Yule created the modern applications of regression, including causal analysis in the social sciences and these would survive the change in the theory of regression but Yule's conception of regression, as depending on a frequency surface for all the variables, was more Pearsonian than Fisherian.[5]

The limitations of Yule's theory are evident in his study of pauperism. The regressors are not normal and, while Yule uses least squares without hesitation, the normal theory probable errors come with the warning (1899, p. 277)–"so far as they are valid for these cases of non-normal correlation". Yule did not pursue the problems of non-normal correlation; his later work on correlation concentrated on time series where the difficulty is the serial dependence of the observations.

Non-normal correlation was Pearson's project. Applications to the biological, social and physical sciences required a generalisation of his theory of univariate skew curves (1895) but Pearson was dissatisfied with normality and least squares even in their traditional domains:

[4]It was evolving or disintegrating into something else: see Aldrich(1997) for discussion and references.
[5]For Yule's applied regression, see Stigler (1986, pp. 353-8) and Aldrich (1995 and -97a).

he (1902) favoured the method of moments for curve fitting and he (1902a) criticised the use of the normal distribution for observational errors in astronomy.

In early twentieth century Britain the theory of errors was applied and taught but there was no research effort comparable to that in Biometry or Statistics. The students of these more dynamic subjects were not taught the theory of errors. The textbooks mentioned least squares for the sake of their mathematician readers but the books did not teach least squares or insist on it. The Edgeworthian Bowley (1900) mentions least squares twice, noting (p. 284) that his way of rationalising his estimate of the modulus of the normal distribution came from that literature and remarking (p. 177) that least squares might be used for obtaining a relationship between the marriage rate and foreign trade. The Pearsonian Elderton (1908, p. viii) skipped least squares because "the range of its applicability is so limited that there is a growing tendency to put it aside in curve fitting". Yule (1911, p. 230) acknowledged his debt to least squares but did not suggest his readers take up its study. Pearsonian notions, meanwhile, were infiltrating courses for astronomers on the combination of observations; see Brunt (1917, chapters IX and X).

Pearson was crushing the theory of errors so it is remarkable that Seal (p. 16) could write in 1967 that Pearson "must ... be given the credit for extending the Gauss linear model to a much broader class of problems than those of errors of measurement". The story of fixed x regression is of *Fisher* making this improbable perspective possible.[6] Yet even without Fisher the theory was not quite dead: 'Student' (1908) solved the problem of the "probable error of the mean"–writing in Pearson's journal, using Pearson's curves–and the theory was acquiring a role in agricultural experiments. 'Student' was involved and so was F. J. M. Stratton, Fisher's tutor–see E. S. Pearson (1990, p. 47). These were peripheral activities; there was no live

[6] Of course Fisher did not crush the theory of correlation; it lives on as multivariate analysis where both his and Pearson's names are hallowed.

centre.

2 Regression curves & goodness of fit

Testing equations was a Pearsonian project–his goodness of fit test (1900) applied to his non-linear regression specification (1905). However the initiator was E. E. Slutsky (1880-1948), an economist correlator based in Kiev.[7]

A "general theory of skew correlation and non-linear regression" needs a surface relating to univariate skew curves as the multinormal surface relates to the normal curve. Pearson (1905, §4) reported that the full theory of skew correlation surfaces "has not yet been worked out owing to difficulties of analysis, and their complete discussion must be postponed" so the regression curve had to serve.[8] Pearson does not write down a specification but something commodious is implied, perhaps:

$$Y = \mu(X) + \varepsilon(X)$$

where the regression curve, $\mu(X) = E(Y|X)$, may be linear, quadratic, ..., or quartic; my symbol $\varepsilon(X)$ marks the possibility that the distribution of the deviation can vary with X; in particular skewness and scedasticity can vary.[9]

The data on X is grouped and consists of repeated values of x–an "array"–with associated values of y. If Y_p is associated with x_p and μ_p is its expected value, then the regression curve of y on x expresses the relationship between μ_p and x_p.[10] There are n_p replicates of x_p where the

[7] Seneta (1988) has biographical information.

[8] Pearson (1923) reviewed his and his assistants' years of struggle with this problem. The paper turned out to be an obituary for the skew-correlation project.

[9] Blyth (1994) gives an interesting account of Pearson's project and his data analysis from the perspective of Bjerve & Doksum's (1993) theory of correlation curves.

[10] I need a magnifying glass to read Pearson's–or Slutsky's–notation so I have simplified it.

numbers n_p are random variables. Pearson does not restrict the distribution of Y_p around μ_p; the normal homoscedastic case was *not* to be assumed. He (§4) gives a string of propositions leading up to the probable error of the "correlation ratio"; this is a measure of correlation which in the case of linear regression equals the correlation coefficient. Pearson proposed testing for linearity by comparing the two quantities and Blakemore (1905) developed the suggestion. No probable errors are given for the regression coefficients which are estimated by the methods of moments.[11]

Slutsky (1913) proposed a χ^2 test for the skew correlation set-up. Denote by $\overline{y}_p$ the mean of the ys associated with μ_p and by e_p the deviation of $\overline{y}_p$ from its expected value μ_p . Appealing to one of Pearson's propositions, Slutsky states that the standard deviation of e_p is given by $\sigma_p/\surd n_p$ where σ_p is the standard deviation of y in the p-th array. Appealing to another, he states "Now it is known that there is no correlation between the deviations in the mean of an x-array and in the mean of a second x-array". Slutsky changes Pearson's specification: he retains heteroscedasticity but assumes that each Y_p is normally distributed around the appropriate μ_p. Slutsky concludes that the quantity

$$\chi^2 = \sum \frac{n_p(\overline{y}_p - \mu_p)^2}{\sigma_p^2}$$

is distributed as chi-square with the number of degrees of freedom equal to the number of arrays. In the test statistic estimates replace unknowns.

The first of Slutsky's examples is the cubic Pearson (1905, §9) fitted to the height and age of 2272 girls classified into 20 age groups. The other is a linear relationship fitted to 124 observations classified into 11 groups on the price of rye in pairs of adjacent months (the

[11] Moore, who attended Pearson's lectures, compares his polynomial regression curves by the estimate of standard error (1914, pp. 78-80). He does not give probable errors for the estimated parameters.

first fitting of an autoregressive model to time series data?). As the numbers in the groups are small, heteroscedasticity cannot be established, so for this case Slutsky reworked the test assuming homoscedasticity.

Pearson (1914, p. xxxii) gave a "word of caution" about Slutsky's procedure then developed his own ideas in a 1916 paper. Pearson (1916, p. 256) warns that "very fallacious results" can be reached by Slutsky's test. Pearson objected to the presumption of normality but adopted the assumption nevertheless. He was content with replacing population quantities with sample quantities but considered Slutsky's replacements unsatisfactory. In the case of a random array size the $\sigma_p/\surd n_p$ quantity could be improved on but Pearson (p. 248) had a more general objection: Slutsky's practice of estimating σ_p and n_p from data on the p-th array but estimating μ_p from all the data was arbitrary–all the quantities should be estimated from all the data. The realised value n_p is replaced by an estimate obtained from fitting a distribution to the x values and σ_p is estimated from the entire sample by using the heteroscedasticity relationship between σ_p and x. Compromises are necessary when working with Slutsky's price data (pp. 250-3) but the full scheme is demonstrated on the abundant height/age data (pp. 253-6).

Slutsky had outlined his test in a letter to Pearson in 1912.[12]. He told Pearson that "quite analogous" would be a criterion to be applied to the physical sciences for testing whether a given system of measurements can reasonably be supposed to correspond to a certain functional relationship. Slutsky's published paper considers only observational (statistical) data but Pearson's "On the Application of 'Goodness of Fit' Tables to Test Regression Curves and Theoretical Curves used to Describe Observational or Experimental Data" considered both and rejected the "analogy".

[12]It is in the Pearson archive at UCL, though without Pearson's reply.

Pearson (p. 256) writes of the physicist making a few measurements of a variate A for each of a series of a variate B: "there is no question in the ordinary sense of a frequency surface". For the category of "physical, technical and astronomical measurements", Pearson's (pp. 256-8) procedure is the same as Slutsky's except that σ_p is estimated by a different method. The important difference between the physical and statistical cases is that in the former, the numbers in each array are non-stochastic. Pearson (p. 247) remarks "It is singular that the goodness of fit theory can actually be applied with greater accuracy to test physical laws than to test regression lines." It would be interesting to know the reasoning behind Slutsky's "analogy"–to know whether *he* invented fixed x regression.

3 Fisher on the fit of regression formulae

In 1922 Fisher was working on several projects–χ^2 theory, agricultural meteorology, genetics, the theory of estimation and the analysis of variance. The projects were more interrelated than they sound. Regression goodness of fit was a minor division of χ^2 theory. Fisher's reconstruction of Pearson's contingency table theory brought him recognition from the statisticians, for Bowley and Yule had an interest in the outcome, but his regression work made no immediate mark.[13]

Goodness of fit is the main business of the 1922 regression paper; see Hald (§27.6) for further discussion. Fisher moves freely between specifications and his arguments are not fully spelt out but his core model is

$$y \sim N(X\beta,\, \sigma^2 I)$$

[13] Box (1978, chapters 3-5) describes Fisher's activities in his first years at Rothamsted. There are treatments of individual topics in Fienberg & Hinkley (1980).

where the N rows of X (the x's may be powers of some underlying variable) comprise a distinct vectors, the p-th of which is replicated n_p times. Fisher focusses on the homoscedastic version of Slutsky's statistic, viz

$$\chi^2 = \sum \frac{n_p(\overline{y}_p - \mu_p)^2}{\sigma^2}$$

Fisher treats explicitly only the statistical case of random n's, though it is obvious that the results also hold for the physical case. The key step is in establishing the distributions of the components of the χ^2 statistic. Fisher (1922, p. 598) writes:

> For such samples of n_p , therefore, the mean, $\overline{y}_p$, will vary about the same mean m_p [my μ_p], and since this mean of $\overline{y}_p$ is independent of the number in the array, m_p [my μ_p] will be the mean of all values of $\overline{y}_p$ from random samples, however the number n_p may vary.

Fisher takes $\surd n_p(\overline{y}_p - \mu_p)$ to be normal with mean zero and standard deviation σ, so the numerator of Slutsky's statistic is σ^2 times a χ^2 . When μ_p has to be estimated, a degrees of freedom adjustment is necessary; in this regression set-up there is the further distributional complication associated with s^2 replacing σ^2 in the test statistic:

$$\chi^2 = \sum \frac{n_p(\overline{y}_p - \widehat{\mu}_p)^2}{s^2}$$

The estimate s^2 is obtained by combining the within-array estimates of σ^2 ; the combining rule is based on a marginal maximum likelihood argument; s^2 has a χ^2 distribution with $N-k$ degrees of freedom. Fisher derives the exact distribution of the test statistic and identifies it as a Pearson Type VI curve–as distinct from the Type III appropriate when σ^2 is known. When

Fisher (1924/8, p. 812 & 1925, pp. 214-8) presented the test in the format of the analysis of variance he introduced the numbers of degrees of freedom associated with the deviation of the array mean from the formula and rescaled the statistic to become $F(a-k,\ N-k)$. The *Statistical Methods (*1925) has tables for $z = \frac{1}{2}lnF$.

Fisher compared his statistic with Slutsky's and with Pearson's statistic for the experimental case. His main point was that neither Slutsky nor Pearson adjusted the degrees of freedom for the estimated parameters; the need for such adjustment was the theme of Fisher's χ^2 work–see e.g. his (1922b).[14] It is clear throughout that for Fisher the observational and experimental cases should *not* be treated differently: e.g. he (p. 607) comments on the limitation of the analysis to the case of groups of y-values corresponding to identical values of x, "little statistical or physical data is strictly of this kind although the former may in favourable cases be confidently grouped, so as to simulate [this] kind of data." When he illustrated the method in *Statistical Methods* (1925, Ex. 42 of §44 in all editions) it was for a (non-randomised) experiment on the influence of temperature on the number of eye facets in drosophila.

4 The distribution of regression coefficients

In his note to the 1950 reprint of "The Goodness of Fit of Regression Formulae, and the Distribution of Regression Coefficients" Fisher remarks that the second topic is linked to the first "only by arising also in regression data". The technical analysis is unrelated–there are no replications and a different distribution is involved–and the work came about in a different way. Box (1978, p. 115) and E. S. Pearson (1990, p. 48). describes how Gosset wrote Fisher in April 1922:

[14] See Lancaster (1969, chapter 1), Fienberg (1980) and Hald (§27.4).

I want to know what is the frequency distribution of $r\sigma_x/\sigma_y$ for small samples, in my work I want that more than the r distribution now happily solved ..."

Fisher was the obvious person to ask about the regression coefficient in the bivariate normal; in 1915 he had "solved" the r distribution–indeed in 1924 and -28 he would solve the partial and multiple problems.

Fisher (1922, p. 598) speaks of "an exact solution of the distribution of the regression coefficients" but the distribution he gives is of the t-ratio for a Gaussian unknown quantity. Yet everything in the paper–except the argument–indicates that the discussion was not restricted to the traditional applications of the theory of errors: the language of "regression coefficients", the bundling with the regression goodness of fit test, the reference (p. 612) to "agricultural meteorology" where x's are weather variables–as in Hooker (1907)–and finally the emphasis in the statement (p. 611), "the accuracy of the regression coefficients is only affected by the correlations which appear *in the sample*", which is pointless unless there is a population of $x's$.

The argument follows what have become familiar lines. Fisher deals with the simple and then the multiple regression model

$$y \sim N(X\beta,\ \sigma^2 I)$$

assuming X *given*. The assumption slips out when Fisher (p. 608) writes of simple regression–the regressor is expressed in deviations from the mean–that the least squares estimates "are orthogonal functions, in that given the values of x observed, their sampling variation is independent."

Gosset's letters show he was disappointed. Through 1922 he kept asking for the marginal distribution for b; in November he told Fisher that the proof of the distribution of b is limited to "given $\overline{x}$ and σ_x". Fisher's letters have not survived, however there is an answer in Fisher (1925b).[15] Fisher (p. 96) begins his derivation of the distribution of the t-ratio, emphasising that he is "confining attention to samples having the same value of x". The work done, he (p. 99) reflects:

> The quantity t involves no hypothetical quantities, being calculable wholly from the observations. It is the point of the method, as of 'Student's' original treatment of the probable error of the mean, to obtain a quantity of known distribution expressible in terms of the observations only. If we had found the distribution of b for samples varying in the values of x observed, we should have been obliged to express the distribution in terms of the unknown standard deviation σ_x in the population sampled; moreover since σ_x is unknown, we should have been obliged to substitute for it an estimate based on $S(x-\overline{x})^2$; the inexactitude of the estimate would have vitiated our solution, and required us to make allowance for the sampling variation of $S(x-\overline{x})^2$; finally this process, when allowance had been accurately made would lead us back to the 'Student's' distribution found above. The proof given above has, however the advantage that it is valid whatever may be the distribution of x, provided that y is normally and equally variable in each array, and the regression of y on x is linear in the population sampled.

While it is not clear that the proof is "valid", the answer to Gosset is clear: the marginal distribution of b is no use on its own and the usable form–the t-ratio–is available whether x is

[15] The distribution theory of this paper and that of the "epoch-making" (1924/8) on the z distribution are reviewed by Hald (pp. 669-74).

normally distributed or not.

Perhaps this was Fisher's case already in 1922. In a letter of 1954–see Bennett (1990, p. 214) he refers to the "application" of the principle invoked in the discussion of the goodness of fit test to the distribution of regression coefficients.

5 The idea of regression

Kendall (1951, p. 12) noticed a "subtle" change in outlook on regression analysis in the early 20s which is "difficult to trace in the literature". An important part of the change was a new idea of regression. In *Statistical Methods for Research Workers* Fisher (1925, p. 114) writes:

> The idea of regression is usually introduced in connection with the theory of correlation, but it is in reality a more general, and, in some respects a simpler idea, and the regression coefficients are of interest and scientific importance in many classes of data where the correlation coefficient, if used at all, is an artificial concept of no real utility.

Excluding Fisher's work, "usually" can be read as "invariably". Regression and correlation were aspects of a joint distribution.

Fisher took the situations for which Pearson and Yule had used correlation/regression and Merriman least squares and treated them together. Yule (1909, p. 722) had described correlation as "an application to the purposes of statistical investigation" of least squares. Least squares was outside Statistics and Yule did not conflate the situations where least squares was traditionally used with those to which correlation/regression was appropriate, nor conflate the sampling theories. Fisher did both. As "classes of data" he made no distinction between observational and experimental material: his examples (pp. 114-36) of x and y include age

and height of children, height of fathers and sons, fertilisers and yield, time and yield, position and rainfall. One sampling theory does for all. Curiously Fisher did not invent a new term for this extended conception or use a term from the theory of errors; he used "regression". Yule (1897, p. 814) had once preferred the colourless "characteristic line" to "regression" with its unwanted associations of biological "stepping back".[16]

Fisher's (pp. 114-5) only restrictions on the use of the model arise from the "very different relations" the independent and dependent variables bear to the regression line. If errors occur in the former, the regression line will be altered; if they occur in the latter, the regression line will not be altered, provided the errors "balance in the averages"; so the errors in variables case was *not* covered. Secondly "the regression function does not depend on the frequency distribution of the independent variable, so that a true regression line may be obtained even when the age groups are arbitrarily selected ..." On the other hand a selection of the dependent variate will "change the regression line altogether".

The book has a chapter on correlation as an aspect of the bivariate normal; evidently correlation coefficients may be of "of interest and scientific importance". The results of his 1915, -21 and -24 papers are presented and illustrated. The examples illustrating the significance of a correlation and a partial correlation (pp. 158-161) are from agricultural meteorology and Yule's work on pauperism. Fisher mentions that the partial correlation depends on the "assumption that the variates correlated (but not necessarily those eliminated) are normally distributed." Fisher was attached to the idea of investigating the existence of dependence between variables by testing hypotheses about the correlation coefficient rather than the regression coefficient. The t-test that modern packages always offer–of $\beta = 0$–appears in the *Methods* as a test on the correlation coefficient, a coefficient only meaningful in the bivariate

[16] Of course interest in "stepping back" continued but not as part of regression analysis. See Stigler (1997).

normal setting.

The *Methods* was a very busy book and the reviewers, including 'Student' and E. S. Pearson, had plenty to discuss without mentioning regression. The book went through 14 editions and it came to be recognised as epoch-making– its silver jubilee marked by an issue of *JASA*. Yet it was a poor platform for the new regression. Fisher (1925, p. 16) had discovered that the same few distributions turn up "again and again" and his book consists of a few tables each prefaced by a chapter surveying its many uses. The idea of regression appears in the chapter on the t-distribution and the regression goodness of fit test in the chapter on the z-distribution. The methods are not documented; "references" are listed but–apart from the data sources–not referred to. The crucial (1924/8) and (1925b) do not appear as they were not published in time for the first edition. Fisher's early readers had to discover for themselves that the regression a's and b's are least squares/maximum likelihood values; after 1934 there is a historical note (§5) mentioning Gauss, least squares and maximum likelihood.[17] Fisher never presented the techniques and the underlying theory together; Hotelling proposed a collaboration on such a presentation but nothing was produced.

The new regression was crowded out of Fisher's empirical work. In 1922 he had mentioned "agricultural meteorology" as an application of regression and his first job at Rothamsted was analysing historical data on yields and weather. The task may have inspired the new regression but the main product–the orthogonal polynomials of (1921a) and (1925)–belonged to the old least squares, to fitting empirical formulae.[18] In the most ambitious study "The Influence of Rainfall on the Yield of Wheat at Rothamsted" Fisher (1924a, p. 96) did not regress yield on weekly rainfall directly but made an ingenious use of orthogonal polynomials

[17] Schultz (1929, p. 86) complained "it is to be regretted that Dr Fisher did not see fit clearly to separate the propositions which are due to him from the general body of statistical theory."

[18] Hald (§25.7) places Fisher's work in a literature which goes back to Chebyshev.

in a discrete approximation to a continuous time formulation in which yield depends on the entire past rainfall record. The regressors are time trends!

"Studies in Crop Variation I" (1921a) analysed historical data but Study II–Fisher & Mackenzie (1923)–analysed Fisher's own experiments and observational studies were soon eclipsed by experiments in the work of Fisher and other statisticians. The new experimentation was not Merriman's or Pearson's for now randomisation was involved. Fisher's work on experiments did not affect his regression theory–it was already done and the later conditional inference theory owed nothing to experiments–but there was probably influence the other way. The randomised experiment set-up resembles Pearsonian regression, with the statistician randomising not nature. The analysis of variance that follows is fixed x analysis.

The new regression went forward without further contributions from Fisher. In Econometrics, the field where regression was most used, a practical synthesis of regression and least squares had been proceeding independently with Ezekiel & Tolley (1923) and others applying least squares algorithms to Pearson/Yule regression.[19] Ezekiel's *Methods of Correlation Analysis* (1930) was based on Fisher's methods; Moore's work of fifteen years before had been based on Pearson's.[20] The first adequate account of Fisher's regression theory appeared in Koopmans's *Linear Regression Analysis of Economic Time Series* (1937).

6 Regression old & new

Pearson did not visibly react to the new regression. He (1923, -23a) continued to publish on frequency surfaces though the grand theory projected twenty years would never arrive. More surprisingly he began working in the vein of 'Student' (1908a) and Fisher (1915). His

[19] For more information on the Gauss, Pearson, Yule and Fisher formalisms see Aldrich (1998).

[20] Ezekiel's book had an influence back on Fisher for it seemed to induce the publication of the fiducial argument–see Aldrich (2000).

first contributions, Pearson (1925, -26), gave Gosset what he had wanted from Fisher–the marginal distribution of b for the bivariate normal. By the end of the decade there was a complete account of the statistics for the multinormal distribution that came into use in the 1890s. Pearson, Fisher (1924 & -28) and Wishart (1928) and Wishart's student Bartlett all contributed.

The results were consolidated in Bartlett's "On the Theory of Statistical Regression" (1933). Part I surveyed the statistics associated with the multivariate normal distribution. Though Bartlett was born into the new regression, he was not satisfied with the treatment in Fisher (1925b): "[Fisher] seems to suggest that ... his test holds under somewhat wider conditions than he assumed." Part II considered which of the results survive if all that is normal is the conditional distribution of one of the variables. Crucial to the analysis were factorisations of the joint distribution. Amongst numerous results Bartlett (p. 278) showed that the t-test of significance of b is "valid, with no restrictions on x ". Bartlett had shown how Fisher's regression theory could be integrated with Yule's, with all the t's crossed.[21]

Pearson (1931, cxxxii-cxl) gave a very negative evaluation of Student's t-work and did not mention the extension to regression. However he conceded the goodness of fit point, writing (1934, p. li) that Fisher's test applies whether the array totals "are kept the same or vary in a random manner". Pearson finally engaged Fisherian regression–without mentioning Fisher–in a very long comment on Welch (1935) and Kolodziejczyk (1935). These papers applied the test theory of Neyman and (E. S.) Pearson (1928 and -33) and used Fisher's regression results but Welch was explicitly concerned with fixed x regression–his y and x have a joint distribution– while Kolodziejczyk's "linear hypothesis" belongs with Neyman (1934) and descends from Markov's statement of the theory of errors–with normality restored. Pearson (1935) argued

[21] Sampson (1974) presents Barlett's results for the multinormal distribution in modern notation. Curiously his tale of two regressions does not mention Bartlett's interest in integrating the two regressions.

that the generality of the "Welch-Kolodziejczyk frequency surface"–the frequency surface for which the normal regression model is the conditional distribution–is illusory for the only important case is the bivariate normal which is best treated by *not* using the Fisher apparatus.

7 Estimation: population, information & sufficiency

Fisher's first justification for fixed x regression was the distribution theory he developed for the tests proposed or inspired by Pearson and 'Student'. Later justifications derived from his own "theory of estimation". Fisher worked on this theory while he worked on fixed x regression but the two would not fit. Developments in the theory eventually solved the problem but it seems that the problem never set the pace. The solution rationalised a practice Fisher knew was right; it is possible that the rightness of the practice guided his thinking on conditional inference but the influence cannot be seen in what he wrote.

"The theory of estimation" is more a theory of the information that estimation–in its popular sense–exploits. The "Mathematical Foundations of Theoretical Statistics" (1922a, p. 311) describes the statistician's task as the "reduction of data"–ideally without loss of information. The statistician specifies a "hypothetical infinite population" to which the observed sample is referred and calculates a statistic which "should summarise the whole of the relevant information supplied by the sample". This is the supreme "criterion of sufficiency" (p. 316); when such a statistic is found "the problem of estimation is completely solved" (p. 315).[22]

The application of sufficiency to regression was problematic. In the regression paper Fisher (1922, p. 598) reflected on his handling of the randomness of n (see Section 3 above)

> we have not attempted to eliminate known quantities, given by the sample, from

[22] See Aldrich (1997) for an account of the paper and the three estimation criteria.

> the distribution formulae of the statistics studied, but only the unknown quantities–parameters of the population from which the sample is drawn–which have to be estimated somewhat inexactly from the given sample.

A footnote ties the point to the "the problem of estimation":

> Statistics whose sampling distribution depends upon other statistics given by the sample cannot, in the strict sense, fulfil the Criterion of Sufficiency. In certain cases evidently no statistic exists, which strictly fulfils this criterion. In these cases statistics obtained by the Method of Maximum Likelihood appear to fulfil the Criterion of Efficiency; the extension of this criterion to finite samples thus takes a new importance.

Fisher's "Theory of Statistical Estimation" (1925a) extended "efficiency" to finite samples–measured by the information in a statistic's sampling distribution–but it did nothing about the inelligibility due to the use of a conditional distribution. In his note to the 1950 reprint, Fisher described the "Theory" as "more compact and businesslike" than the "Foundations"; it was because it shelved many of the problems. The notion Bartlett crystallised as "quasi-sufficiency"–see Section 9 below–better addresses the regression difficulty.

Fisher applied "efficiency" and "consistency"–the two lesser criteria–to regression in an unpublished critique (1924/5) of N. R. Campbell's (1924) alternative to least squares.[23] Fisher states that both methods are consistent and asymptotically normal under general conditions but that least squares is more efficient. If the errors in y are normally distributed, "it may be shown" that the estimate b has 100% efficiency. In the 1922 theory "efficiency" is a large sample property delivered by maximum likelihood.and "showing" presumably used the

[23] Campbell's method was a variant of the method of averages–see Farebrother (1999, pp. 236-7).

fact that least squares is maximum likelihood–in fixed or random x situations. Fisher gives two examples quantifying the inefficiency of Campbell's method. In the first, illustrative of experimental work, the x is a (non-stochastic) arithmetic progression and the other, illustrative of "observational studies", x is a normally distributed. The second analysis is curious because Fisher does *not* condition on x when he calculates the variance of the estimator.

The *infinite* in "hypothetical infinite population" drew immediate fire and Fisher (1925a, p. 700) replied with a clarification. In the 50s–see Section 11–he pressed himself to clarify the *hypothetical.* He (1922a, p. 313) had originally written: "any such set of numbers [observations] are a random sample from the totality of numbers produced by the same matrix of causal conditions". Of course any hypothesised population had to face a "rigorous and objective test of the adequacy with which the proposed population represents the whole of the available facts" but that was the end of it.

Some students paused over the hypotheticalness of the regression population. Working and Hotelling (1929, p. 82), who made the first extension to Fisher's regression t-results, were fitting time trends by least squares:

> The fiction is conventionally adopted that the sampling might be repeated indefinitely with new and independent values of the random part of y, but with the same fundamental trend.

Koopmans (1937, pp. 1-8) spent more time on the interpretation of the fixed x population. He sent his book to Fisher but they seem not to have discussed the regression population. Koopmans later had a strong influence on econometric thinking on the subject but that is another story–see Aldrich (1993).

8 Ancillary information

Ancillarity reconciled regression with the theory of estimation. The notion had been trailed in Fisher (1925a, p. 724) but only became prominent in "Two New Properties of Mathematical Likelihood" (1934) and "The Logic of Inductive Inference" (1935).[24] An ancillary provides help in reducing the "loss of accuracy" associated with the use of a single estimate; the loss is the difference between the information in the entire sample and in the estimate. The ancillaries that materialised in 1934 were for the location and location/scale families. For the location case he showed how conditioning on the "configuration" leads to the full recovery of the information lost; the configuration is the set of $n-1$ differences between the median and the other observations.

Fisher's practice was to work through "trivial but representative" (1956, p. 158) problems, without proving or even stating precisely any theorem of which they are representative instances. To help fix these notions of loss and recovery I present an argument which underlies much that he wrote but which he seems never to have written down. The formulation is from Kalbfleisch (1982, p. 78).

The information in the sample X is

$$I_X(\theta) = -E\,\frac{\partial^2 \ln f_X(x;\theta)}{\partial\theta^2}$$

In the case of interest there is no single sufficient statistic. If T is the maximum likelihood estimator of θ, then $I_T(\theta)$ the information in T, calculated from the sampling distribution of T, will be less than that in the sample, $I_X(\theta)$. Fisher calls the difference the information "lost" in using T rather than X.

[24] Hinkley (1980 and -80a) and Hald (pp. 729-33) discuss relevant aspects of Fisher's theory of estimation.

Suppose there is a statistic A (for ancillary) such that (T, A) is jointly sufficient for θ and the distribution of A is free from θ. Consider now the information in the conditional distribution of T given the realised value of A:

$$\begin{aligned} I_{T|A=a}(\theta) &= -E\left[\frac{\partial^2 \ln f_{T|A}(t;x;\theta)}{\partial\theta^2}|A=a\right] \\ &= -E\left[\frac{\partial^2 \ln f_{T,A}(t;x;\theta)}{\partial\theta^2}|A=a\right] \end{aligned}$$

since $f_A(a)$, the density of A is free from θ.

Average these conditional informations across A and use the joint sufficiency of (T, A) to obtain the information measure for the sample.

$$EI_{T|A}(\theta) = I_{T,A}(\theta) = I_X(\theta)$$

Fisher's (1934, p. 303) gloss is that "The process of taking account of the distribution of our estimates in samples of the particular configuration [A for the location problem] observed has therefore recovered the whole of the information available."

Fisher (1935, p. 48) emphasises two further points: ancillary statistics tell us nothing about the value of the parameter; their function is to tell us what "reliance" to place on the estimate. Regression is not mentioned but the ideas seem obviously applicable. Fisher also initiated a second life for ancillarity with an example showing that ancillarity is "useful not only in questions of estimation proper" (p. 78). This is the test for independence in the $2X2$ table, obtained by conditioning on the margins (p. 48):

> If it be admitted that these marginal frequencies by themselves supply no information ... as to the proportionality of the frequencies in the body of the table

> we may recognize the information they supply as wholly ancillary; and therefore recognize that we are concerned only with the relative probabilities of occurrence of the different ways in which the table may be filled in, subject to these marginal frequencies.

Information loss and recovery do not figure. Fisher eventually applied to regression both this estimation-free notion of ancillarity (and information) and the original notion–see Sections 10 and 11 below–but in the 30s only Bartlett published on ancillarity in relation to regression.

9 Quasi-sufficiency & statistical regression

For the next few years Bartlett wrote more about conditional inference than Fisher; Fraser (1992) has a brief review. The first paper, "Statistical Information and Properties of Sufficiency" (1936), is the most relevant to regression. Bartlett took Fisher's idea away from "the theory of estimation"–he did not share Fisher's "reduction of data" viewpoint nor his enthusiasm for information calculations in small sample work.[25]

Bartlett (p. 131) re-states Fisher's analysis of the location problem; the distribution of each item in the sample S is of the form $f(x-m)$, the chance of a configuration C is independent of the parameter m and there is a T such that

$$\begin{aligned} p(S \,|\, m) &= p(S \,|\, C, m)\, p(C) \\ &= p(T \,|\, C, m)\, p(C) \end{aligned}$$

"Hence all the information on m is given by $T \,|\, C$". Such estimates as T are called *quasi-*

[25] Fisher and Bartlett had a protacted argument over fiducial inference: see Bartlett (1965) and by Zabell (1992, pp. 377-8). Bartlett reflects on his general relations with Fisher in Bartlett (1982) and Olkin (1989).

sufficient statistics by analogy with the factorisation condition for sufficiency.

Bartlett thus made explicit the scheme implicit in Fisher (1934)–or rather half-explicit for the reader has to define quasi-sufficiency in general. That done, Bartlett (p 135) points out:

> The important practical illustration of the use of *quasi*-sufficient statistics occurs in the theory of statistical regression. In the simplest case [σ^2 known] our estimate b_{yx} of the coefficient β_{yx} is accompanied by a specification of the value of $\sum(x-\overline{x})^2$ obtained, the distribution of $b_{yx}|\sum(x-\overline{x})^2$ being normal (for normal y), whatever the distribution of x;

Behind the correspondence between $(b_{yx}, \sum(x-\overline{x})^2)$ and (T, C) is a certain amount of calculation which is not given; it draws on the factorisation analysis of the 1933 paper on statistical regression–see Section 6 above. I do not think the sufficiency of b in the fixed x regression model had been noted before.

Bartlett's later papers show the influence of Neyman & Pearson (1933) as well as of Fisher: as Fraser (p. 110) notes, the 1937 paper "uses concepts and theory from both the Fisher and Neyman-Pearson schools in a manner that might now be called unified".[26] Bartlett returned to quasi-sufficiency and the regression example after Welch (1939) identified a conflict between conditioning and power. Welch (p. 66) had concluded "that certain methods, for which properties analogous to those of sufficiency have been claimed, do not satisfy conditions which I think they should, if these claims are to be upheld". The "claims" were Fisher's, the "conditions" related to Neyman and Pearson's power but Bartlett felt the criticism.

Bartlett (1939, p. 392) did not directly defend conditioning but turned the objection by showing how, if the size of the test is varied with the value of the ancillary, conditional tests can achieve maximum power for a given unconditional (long run) size. He illustrates as follows:

[26] Earlier Bartlett (1933a) had tried to explain Fisher and Jeffreys to one another.

> The orthodox theory is to consider the conditional statistic $b|\sum(x-\overline{x})^2$ ... Suppose for the sake of argument that the true variance of ...y_x was known to be unity, and the x's are such that $\sum(x-\overline{x})^2 = 1$ on Mondays and 1.44 on Tuesdays. Then for an 0.025 significance level (one tail), the usual practice would be to take 1.96 as the significance level for b (from $b_0 = 0$) on Mondays, and $1.96/1.2 = 1.633$ on Tuesdays. The power of the test in relation to the alternative that $b_1 = 3.92$ is 0.9860. But if we were satisfied with adjusting the significance level to be 0.025 merely in the long run for Mondays and Tuesdays together, we may raise the power of the test to its maximum value of 0.9878 by taking the Monday significance level at $b = 1.87$ ($\alpha = 0.0307$) and the Tuesday level at $b = 1.723$ ($\alpha = 0.0194$).

Bartlett returned to regression and conditioning in an obituary of E. S. Pearson. He (1981, p. 3) mentioned a "formidable logical criticism" of the concept of power: in regression the conditional power for a test about β_{yx} depends on $\sum(x-\overline{x})^2$ and "we usually consider it irrelevant to ask whether we can obtain a better procedure based on "absolute power" by considering the sampling variation of $\sum(x-\overline{x})^2$."

In the 30s neither Fisher nor Bartlett articulated what Birnbaum (1962) called the "conditionality principle": see Cox & Hinkley (pp. 38-9). However Bartlett wrote as though he accepted it while Fisher was more equivocal–there was a world beyond "questions of estimation proper" but information extraction came first.

10 Regression & ancillary information

Fisher probably regarded the application of ancillarity to regression as obvious but his earliest mention of the application that I have seen is in a 1939 letter to Jeffreys (see Bennett, 1990,

p. 173):

> I regard regression work ... as a good example of ancillary information, in that the precision of the regression does not really depend on the number in the sample, but only on the sum of squares of the independent variate, or, in general, on the dispersion ... In fact the whole work is completely independent of how they may be distributed in the population sampled ...

Fisher made the same point to Darmois in August 1940. In an earlier letter Fisher (see Bennett, 1990, p. 70) criticised Bartlett's use of the "phrase 'conditional sufficiency"'.[27] Fisher never referred to Bartlett's treatment of regression but he was surely aware of it.

The Fisher-Darmois correspondence (see Bennett, pp. 65-79) is particularly rich and many of the ideas sketched there went into "Conclusions Fiduciaires" (1948) and then into *Statistical Methods & Scientific Inference*. One of the innovations of "Conclusions Fiduciaires" was a definition of ancillarity had been implicitly defined by its role in recovering lost information but o was the definition (p. 193):

> Tout ensemble de statistiques dont la distribution simultanée est indépendante des paramètres, est appelé un ensemble "ancillaire" des statistiques.

When Fisher (1935) wrote that marginal frequencies "supply no information" he presumably envisaged a notion like that embodied in this definition.

In "Conclusions Fiduciaires" the technique of the theory of estimation was applied to regression. The paper's second example (p. 197) considers the bivariate normal regression model with known variance σ^2 In 1922 Fisher had written that the least squares estimator b is normally distributed with variance σ^2 $/A$, where A is the sum of squared deviations of x.

[27] Bartlett did not use the term though years later Cox and Hinkley (1974, p. 32) did.

He now supposed that x is normally distributed with known variance α, so that A is α times a χ^2 with $N-1$ degrees of freedom. With this specification the marginal distribution of b is a non-central t with $N-1$ degrees of freedom with parameters which are functions of the known quantities α and σ^2 and the unknown β. Fisher (p 21) states that there is less information in this unconditional distribution than in the normals of which it is a mixture. By ignoring the value of A and using the value of α and the sample size N, the information has been reduced in the ratio $N/N+2$.

Fisher obtains this value by applying results from his first example which was itself based on a weighting argument that went back to 1925. However a free-standing argument can be based on the 'implicit theorem' of Section 8 above, identifying β with θ and (b, A) with (T, A). The information in a sample of size N conditional on the value of A is A/σ^2. Taking the expectation of these conditional informations yields $\alpha(N-1)/\sigma^2$ as the information in the entire sample. If now we compute the information in b from its marginal distribution we obtain a smaller value: the information in the sample is reduced in the ratio $N/N+2$.

These information calculations–unlike most of the paper–did not find their way into *Statistical Methods & Scientific Inference*. Indeed regression does not appear in the estimation chapter but in the chapter on "misapprehensions about tests of significance".

11 A multiplicity of populations

Fisher's campaign against the Neyman-Pearson theory of testing and the notion of repeated sampling from a fixed population was a reply to criticisms of his tests for the $2X2$ table and the Behrens-Fisher problem. Thus on the latter he (1946, p. 713) wrote–misguidedly–against Bartlett:

> I am quite aware that Bartlett, following Neyman, feels bound to identify the population of samples envisaged in tests of significance with those generated by repeated sampling of a fixed hypothetical population ...

In the polemics of the 1950s Fisher argued that the criticised work must be correct because it follows the regression pattern which everyone knows is correct. The business is testing and the estimation theory aspect disappears.

The article, "Statistical Methods and Scientific Induction", and the book, *Statistical Methods & Scientific Inference*, stress the *hypotheticalness* of the statistician's population. The root difficulty with the formula, "repeated sampling from the same population", is that there is "a multiplicity of populations to each of which we can regard our sample as belonging" (1955, p. 71). In an acceptance samping (quality control) situation the population has an "objective reality" but in the natural sciences the population is a "product of the statistician's imagination" and "the first to come to mind may be quite misleading" (1956, pp. 77 & 78). Fisher was criticising Neyman but the formulation in the "Foundations"–see Section 7 above–was just as vulnerable.

Setting up the regression model, Fisher (1955, p 71) states "The qualitative data may also tell us how x is distributed with or without specific parameters; this information is irrelevant." He (p. 72) continues

> The normal distribution of b about β with variance σ^2/A does not correspond with any realistic process of samping for acceptance but to a population of samples in all relevant respects like that observed, neither more precise nor less precise, and which therefore we think it appropriate to select in specifying the precision of the estimate b. In relation to the value of β the value A is known as an *ancillary statistic*.

However there is no appeal to information calculations.

In the book Fisher does not use the word "ancillary" with regresssion; perhaps to make the attack on repeated sampling less dependent on the theory of estimation which was not widely accepted. He (1956, p. 82) presented regression adding the fiducial distribution of β and this pointed introduction:

> A case which illustrates well how misleading the advice is to base the calculations on repeated sampling from the same population, if such advice were taken literally, is that of data suitable for the estimation of a coefficient of linear regression.

The regression material appears in the book's chapter on "misapprehensions" about significance tests and the material is organised around the t-distribution, i.e. Fisher's first regression theory where it is shown that the distribution is not affected by the distribution of x. The "advice" for testing and fiducial inference resulting from failure to condition is not going to be "misleading"–it is going to be the same.

For Bartlett (1939) and for Fisher (1948) conditioning had to be related to power or information. In 1956 Fisher (p. 84) makes a more direct appeal:

> To judge of the precision of a given value of b, by reference to a mixture of samples having different values of A, and therefore different precisions for the values of b they supply, is erroneous because these other samples throw no light on the precision of that value which we have observed.

This is an eloquent amplification of the 1922 proposition: "the accuracy of the regression coefficients is only affected by the correlations which appear *in the sample*".

Fisher wrote about fixed x regression over a period of more than thirty years. He produced three justifications: from t distribution theory, from the theory of estimation and from the

application of the conditionality principle to the choice from the multiplicity of populations. He saw these justifications not as alternatives but as mutually reinforcing.

12 Retrospects

In 1956 Fisher gave a very lean history of fixed x. He wipes his own contribution and the Pearsonian past of regression and bivariate distributions. Like Seal (see Introduction), Fisher (p. 84) describes a return–from Pearson (1925) to Gauss (1809):

> [I]n repeated sampling from the bivariate distribution of x and y, the value of A would vary from sample to sample. The distribution of $(b - \beta)$ would no longer be normal, and before we knew what is was, the distribution of A, which in turn depends on that of x would have to be investigated. Indeed, at any early stage Karl Pearson did attempt the problem of the precision of a regression coefficient in this way, assuming x to be normally distributed. The right way had, however been demonstrated many years before by Gauss, and his method only lacked for completeness the use of 'Student"s distribution, appropriate for samples of rather small numbers of observations.

In his Foreword Fisher (p. 3) had remarked that Pearson cared little for the past, instancing the "Gaussian tradition of least square techniques". In the 20s, when Pearson's regression was being claimed for the Gaussian tradition, Gauss was not to be seen. Only Fisher's first paper (1912) has working references to the literature of that tradition.

In 1956 Fisher thought fixed x regression beyond dispute. It was clearly a presence but on what terms? A full answer would be a project in itself. Section 6 gave some views from the Rothamsted/University College 'inside' and I will add some examples from outside to

illustrate further possibilities. Hotelling was a born again Fisherian of the 1920s, an early contributor to regression t-theory and should have been an insider.[28] Yet when he (1940, pp. 276-7) weighed the merits of the fixed x and joint multinormality assumptions in the regressor/predictor selection problem he did not consider Fisher's distribution agument:

> The advantages of exactness and of freedom from the somewhat special trivariate normal assumption are obtained at the expense of sacrificing the precise applicability of the results to other sets of values of the predictors.

This recalls Yule on normal theory probable errors (Section 1 above): they are not perfect but they work–after a fashion.

Cramér (1946) noticed the distribution theory argument or at least part of it. His chapter 29 considers regression inference for the multinormal distribution case and chapter 37 regression with non-random x's. Cramér (p. 550) records the "formal identity" of the t-results in the two cases; he noticed Part I of Bartlett (1933) but did not mention the results in Part II.

The conditioning arguments were less visible and less noticed. The underlying theory of estimation was not accepted, understood or even widely known. Thus Hotelling (1948, p. 867) complained after reading Kendall's *Advanced Theory:* "it is still not clear what the statistician is supposed to do with ancillary statistics".[29] Before 1955/6 the regression applications were footnotes not headlines.[30] Yet modern views on conditioning in regression–see Barndorff-Nielsen & Cox (1994, p. 39) and Gelman et al (1995, p. 235)–are linked to Fisher and Bartlett through Cox (1958) and Savage (1962). Cox re-stated Fisher's point about the multiplicity of populations, the weighing machine example (p. 360) is a parable for regression and the

[28] See Aldrich (2000) for an account of Hotelling's relations with Fisher.

[29] Kendall discusses ancillarity with the theory of estimation (p. 32); it is not linked to conditional testing which is applied to regression (pp. 127 & 156).

[30] The earliest reference in Barndorff-Nielsen's (1978, p. 36) historical note on regression in relation to ancillarity is Fisher (1956).

references include Bartlett (1939).[31] Savage (p. 19) gives a Bayesian view of ancillarity and regression, referring to Cox.

The fixed x assumption was not a central issue in Anglo-American Statistics but from the 30s into the 70s econometricians made a profession out of *not* fixing x–with errors in variables and simultaneity.[32] They had to discuss when the fixed x practice would pass–see Aldrich (1993)–but their discussions did not influence the literature treated here. A matter the econometricians discussed is the possible *causal* nature of the relation between y and x. From Yule onwards regression was used for investigating causal relations but in the statistical tradition the causal interest was not intrinsic to the statistical analysis but a thing apart.

References

Aitken, A. C. (1935) On Least Squares and Linear Combinations of Observations, *Proceedings of the Royal Society of Edinburgh*, **55**, 42-48.

Aldrich, J. (1993) Cowles Exogeneity and CORE Exogeneity, Southampton University, Department of Economics, Discussion Paper 9308.

_____ (1995) Correlations Genuine and Spurious in Pearson and Yule, *Statistical Science*, **10**, 364-376.

_____ (1997) R. A. Fisher and the Making of Maximum Likelihood 1912-22, *Statistical Science*, **12**, 162-176.

_____ (1997a) Multiple Regression Grumbles, in D. Conniffe (ed) *Roy Geary 1896-1983 Irish Statistician*, Dublin, Oaktree Press.

_____ (1998) Doing Least Squares: Perspectives from Gauss and Yule, *International Statistical Review*, **66**, 61-81.

[31] See Reid (1994) for Cox's own account of the background to his paper.
[32] Morgan (1990) describes their work

_____ (1999) Determinacy in the Linear Model: Gauss to Bose and Koopmans, *International Statistical Review*, **67**, 211-219.

_____ (2000) R. A. Fisher's 'Inverse Probability' of 1930, to appear in *International Statistical Review*.

Barndorff-Nielsen, O. (1978) *Information and Exponential Families*, Chichester: Wiley.

Barndorff-Nielsen, O. E. & D. R. Cox (1994) *Inference and Asymptotics*, London: Chapman & Hall.

Bartlett, M. S. (1933) On the Theory of Statistical Regression, *Proceedings of the Royal Society of Edinburgh*, **53**, 260-283.

_____ (1933a) Probability and Chance in the Theory of Statistics, *Proceedings of the Royal Society A*, **141**, 518-534.

_____ (1936) Statistical Information and Properties of Sufficiency, *Proceedings of the Royal Society A*, **154**, 124-137.

_____ (1937) Properties of Sufficiency and Statistical Tests, *Proceedings of the Royal Society A*, **160**, 268-282.

_____ (1939) A Note on the Interpretation of Quasi-sufficiency, *Biometrika*, **31**, 391-2.

_____ (1965) R. A. Fisher and the First Fifty Years of Statistical Methodology, *Journal of the American Statistical Association*, **60,** 395-409.

_____ (1981) Egon Sharpe Pearson, 1895-1980, *Biometrika,* **68**, 1-7.

_____ (1982) Chance and Change, in J. Gani (ed) *The Making of Statisticians*, New York: Springer-Verlag.

Bennett, J. H. (1971) (ed) *Collected Papers of R. A. Fisher*, Adelaide: Adelaide University Press.

_____ (1990) (ed) *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*, Oxford, University Press.

Birnbaum, A. (1962) On the Foundations of Statistical Inference, *Journal of the American Statistical Association*, **57**, 269-306.

Bjerve, S. & K. A. Doksum (1993) Correlation Curves: Measures of Association as Functions of Covariates, *Annals of Statistics*, **21**, 890-902.

Blakeman, J. (1905) On Tests for Linearity of Regression in Frequency Distributions, *Biometrika*, **4**, 332-350.

Blyth, S. (1994) Karl Pearson and the Correlation Curve, *International Statistical Review,* **62**, 393-403.

Bowley, A. L. (1900) *Elements of Statistics*, London: King.

Box, J. F. (1978) *R. A. Fisher: The Life of a Scientis*t, New York: Wiley.

Brunt, D. (1917) *The Combination of Observations*, Cambridge: Cambridge University Press.

Campbell, N. (1924) The Adjustment of Observations, *Philosophical Magazine*, **47**, 816-826.

Cox, D. R. (1958) Some Problems Connected with Statistical Inference, *Annals of Mathematical Statistics*, **29**, 357-372.

Cox, D. R. & D. V. Hinkley (1974) *Theoretical Statistics*, London: Chapman & Hall.

Cramér, H. (1946) *Mathematical Methods of Statistics*, Princeton: Princeton University Press.

Ezekiel, M. (1930) *Methods of Correlation Analysis*, New York: Wiley.

Farebrother, R. W. (1999) *Fitting Linear Relationships: A History of the Calculus of Observations,* New York: Springer.

Fienberg, S. E. (1980) Fisher's Contribution to the Analysis of Categorical Data, pp. 75-84 of Fienberg & Hinkley (1980).

Fienberg, S. E. & D. V. Hinkley (1980) (eds) *R. A. Fisher: An Appreciation*, New York: Springer.

Fisher, R. A. (1912) On an Absolute Criterion for Fitting Frequency Curves, *Messenger of Mathematics*, **41**, 155-160. CP1.

_____ (1915) Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population, *Biometrika*, **10**, 507-521. CP4.

_____ (1921) On the 'Probable Error' of a Coefficient of Correlation Deduced from a Small Sample, *Metron*, **1**, 3-32. CP14.

_____ (1921a) Studies in Crop Variation. I. An Examination of the Yield of Dressed Grain from Broadbalk, *Journal of Agricultural Science*, **11**, 107-135. CP15.

_____ (1922) The Goodness of Fit of Regression Formulae, and the Distribution of Regression Coefficients, *Journal of the Royal Statistical Society*, **85**, 597-612. CP20.

_____ (1922a) On the Mathematical Foundations of Theoretical Statistics *Philosophical Transactions of the Royal Society, A*, **222**, 309-368. CP18.

_____ (1922b) On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P, *Journal of the Royal Statistical Society*, **85**, 87-94. CP19.

_____ (1924) The Distribution of the Partial Correlation Coefficient, *Metron*, **3**, 329-332. CP35.

_____ (1924a) The Influence of Rainfall on the Yield of Wheat at Rothamsted, *Philosophical Transactions of the Royal Society, B*, **213**, 89-142. CP37.

_____ (1924/5) Note on Dr Campbell's Alternative to the Method of Least Squares, unpublished manuscript. In Barr-Smith Library of the University of Adelaide.

_____ (1924/8) On a Distribution Yielding the Error Functions of Several Well Known Statistics, *Proceedings of the International Congress of Mathematics, Toronto,* **2**, 805-813.

CP36.

_____ (1925) *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd.

_____ (1925a) Theory of Statistical Estimation, *Proceedings of the Cambridge Philosophical Society*, **22**, 700-725. CP42.

_____ (1925b) Applications of 'Student's' Distribution, *Metron*, **5**, 90-104. CP43.

_____ (1928) The General Sampling Distribution of the Multiple Correlation Coefficient, *Proceedings of the Royal Society, A*, **121**, 654-673. CP61.

_____ (1934) Two New Properties of Mathematical Likelihood, *Proceedings of the Royal Society, A*, **144**, 285-307. CP108.

_____ (1935) The Logic of Inductive Inference, (with discussion) *Journal of the Royal Statistical Society,* **98**, 39-82. CP124.

_____ (1946) Testing the Difference between Two Means of Observations of Unequal Precision, *Nature*, **15**, 713. CP207.

_____ (1948) Conclusions Fiduciaires, *Annales de l'Institute Henri Poincaré*, **10**, 191-213. CP222.

_____ (1955) Statistical Methods and Scientific Induction, *Journal of the Royal Statistical Society, B*, **17**, 69-78. CP261.

_____ (1956) *Statistical Methods and Scientific Inference*, Edinburgh: Oliver & Boyd.

Fisher, R. A. & W. A. Mackenzie (1923) Studies in Crop Variation. II. The Manurial Response of Different Potato Varieties, , *Journal of Agricultural Science*, **13**, 311-320. CP32.

Fraser, D. A. S. (1992) Introduction to reprint of Bartlett (1937), pp. 109-112 of S. Kotz & N. L. Johnson (eds) *Breakthroughs in Statistics, volume 1*, New York: Springer.

Galton, F. (1877) Typical Laws of Heredity, *Nature*, **15**, 492-495, 512-514, 532-533.

_____ (1886) Family Likeness in Stature, *Proceedings of the Royal Society*, **40**, 42-73.

Gauss, K. F. (1809) *Theoria Motus Corporum Coelestium*, English translation by C. H. Davis, reprinted 1963, Dover, New York

Gelman, A & J. B. Carlin, H. S. Stern & D. B. Rubin (1995) *Bayesian Data Analysis*, London: Chapman & Hall.

Hald, A. (1998) *A History of Mathematical Statistics from 1750 to 1930*, New York: Wiley.

Hinkley, D. V. (1980) Theory of Statistical Estimation: the 1925 Paper, pp. 85-94 in Fienberg & Hinkley (1980).

_____ (1980a) Fisher's Development of Conditional Inference, pp. 101-108 in Fienberg & Hinkley (1980).

Hooker, R. H. (1907) Correlation of the Weather and Crops, *Journal of the Royal Statistical Society*, **70**, 1-51.

Hotelling, H. (1940) The Selection of Variates for Use in Prediction with Some Comments on the Problem of Nuisance Parameters, *Annals of Mathematical Statistics*, **11**, 271-283.

_____ (1948) Review of *The Advanced Theory of Statistics vol. 2*, by M. G. Kendall, *Bulletin of the American Mathematical Society*, **54**, 863-868.

Jevons, W. S. (1871) *The Theory of Political Economy*, London: Macmillan.

Kalbfleisch, J. (1982) Ancillary Statistics, pp. 77-81 of S. Kotz & N. L. Johnson (eds) *Encyclopedia of Statistical Science*, volume 1, New York: Wiley.

Kendall,.M. G. (1946) *The Advanced Theory of Statistics vol. 2*, London: Griffin.

_____ (1951) Regression, Structure and Functional Relationship. Part I, *Biometrika*, **38**, 11-25.

Kolodziejczyk, S. (1935) On an Important Class of Statistical Hypotheses, *Biometrika*, **27**, 161-190.

Koopmans, T. C. (1937) *Linear Regression Analysis of Economic Time Series*, Haarlem: Bohn.

Lancaster, H. O. (1969) *The Chi-squared Distribution*, New York: Wiley.

Merriman, M. (1884/1911) *A Textbook on the Method of Least Squares*, New York, Wiley. References to the eighth edition 1911.

Moore, H. L. (1914) *Economic Cycles–their Law and Cause*, New York: Macmillan.

_____ (1917) *Forecasting the Yield and the Price of Cotton*, New York: Macmillan.

Morgan, M. S. (1990) *A History of Econometric Ideas*, New York, Cambridge University Press.

Neyman, J. (1934) On the Two Different Aspects of the Representative Method, (with discussion), *Journal of the Royal Statistical Society*, **97**, 558-625.

Neyman, J. & E. S. Pearson (1928) On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference, Biometrika, 20A, 175-240 & 264-294.

_____ (1933) On the Problem of the Most Efficient Tests of Statistical Hypotheses, *Philosophical Transactions of the Royal Society*, **231**, 298-337.

Olkin, I (1989) A Conversation with Maurice Bartlett, *Statistical Science*, **4**, 151-163.

Pearson, E. S. (1926) Review of *Statistical Methods for Research Workers* by R. A. Fisher, *Science Progress*, **20**, 733-734.

_____ (1990) *'Student', A Statistical Biography of William Sealy Gosset*, Edited and Augmented by R. L. Plackett with the Assistance of G. A. Barnard, Oxford: Oxford University Press.

Pearson, K. (1895) Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material, *Philosophical Transactions of the Royal Society A*, **186**, 343–414.

_____ (1896) Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia, *Philosophical Transactions of the Royal Society A*, **187**, 253-318.

_____ (1900) On the Criterion that a Given System of Deviations from the Probable in the Case of Correlated System of Variables is such that it can be reasonably Supposed to have

Arisen from Random Sampling, *Philosophical Magazine*, **50**, 157-175.

_____ (1902) On the Systematic Fitting of Curves to Observations and Measurements, Parts I & II, *Biometrika*, **1**, 265-303 and **2**, 1-23.

_____ (1902a) On the Mathematical Theory of Errors of Judgement, with Special Reference to the Personal Equation, *Philosophical Transactions of the Royal Society A*, **198**, 235-299.

_____ (1905) On the General Theory of Skew Correlation and Non-Linear Regression, *Drapers' Company Research Memoirs, Biometric Series II,* Cambridge: Cambridge University Press.

_____ (1914) *Tables for Statisticians and Biometricians*, Cambridge: Cambridge University Press.

_____ (1916) On the Application of 'Goodness of Fit' Tables to Test Regression Curves and Theoretical Curves used to Describe Observational or Experimental Data, *Biometrika*, **11**, 239-61.

_____ (1920) Notes on the History of Correlation, *Biometrika*, **13**, 25-45.

_____ (1923) Notes on Skew Frequency Surfaces, *Biometrika*, **15**, 222-230.

_____ (1923a) On Non-Skew Frequency Surfaces, *Biometrika*, **15**, 231-244.

_____ (1925) Further Contributions to the Theory of Small Samples, *Biometrika*, **17**, 176-199.

_____ (1926) Researches on the Mode of Distribution of the Constants of Samples taken at Random from a Bivariate Normal Population, *Proceedings of the Royal Society A*, **112**, 1-14.

_____ (1931) *Tables for Statisticians and Biometricians, Part II*, Cambridge: Cambridge University Press.

_____ (1934) *Tables of the Incomplete Beta-Function*, Cambridge: Cambridge University Press.

_____ (1935) Thoughts Suggested by the Papers of Messrs Welch and Kolodziejczyk, *Biometrika*, **27**, 227-259.

_____ & Filon L. N. G. (1898) Mathematical Contributions to the Theory of Evolution IV. On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation, *Philosophical Transactions of the Royal Society A,* **191**, 229-311.

Reid, N. (1994) A Conversation with Sir David Cox, *Statistical Science*, **9**, 439-455.

Sampson, A. R. (1974) A Tale of Two Regressions, *Journal of the American Statistical Association*, **69**, 682-689.

Savage, L. J. (1962) Subjective Probability and Statistical Practice, pp. 9-35 of L. J. Savage et al *The Foundations of Statistical Inference: A Discussion*, London: Methuen.

Schultz, H. (1929) Applications of the Theory of Error to the Interpretation of Trends: Discussion, *Journal of the American Statistical Association*, *Supplement: Proceedings of the American Statistical Association*, **24**, 86-89.

Seal, H. (1967) The Historical Development of the Gauss Linear Model, *Biometrika*, **54**, 1-24.

Seneta, E. (1988) Slutsky (Slutskii), Evgenii Evgenievich, pp. 512-515 of S. Kotz & N. L. Johnson (eds) *Encyclopedia of Statistical Science*, volume 8, New York: Wiley.

Slutsky, E. (1913) On the Criterion of Goodness of Fit of the Regression Lines and on the Best Method of Fitting Them to the Data, J*ournal of the Royal Statistical Society*, **77,** 78-84.

Stigler S. M. (1986) *A History of Statistics.* Cambridge, Mass.: Belknap Press.

_____ (1997) Regression Towards the Mean, Historically Considered, *Statistical Methods in Medical Research*, **6**, 103-114.

'Student' (1908) The Probable Error of a Mean, *Biometrika*, **6**, 1-25.

_____ (1908a) Probable Error of a Correlation Coefficient, *Biometrika*, **6**, 302-310.

_____ (1926) Review of *Statistical Methods for Research Workers* by R. A. Fisher, *Eugenics Review*, **18**, 148-150.

Tolley, H. R. & M. J. B. Ezekiel (1923) A Method of Handling Multiple Correlation Problems, *Journal of the American Statistical Association*, **18**, 993-1003.

Welch, B. L. (1935) Some Problems in the Analysis of Regression Among k Samples of Two Variables, *Biometrika*, **27**, 145-160.

_____ (1939) On Confidence Limits and Sufficiency, *Annals of Mathematical Statistics*, **10**, 58-69.

Wishart, J. (1928) The Generalised Product Moment Distribution in Samples from a Normal Multivariate Population, *Biometrika*, **20, A**, 32-52.

Working, H. & H. Hotelling (1929) Applications of the Theory of Error to the Interpretation of Trends, *Journal of the American Statistical Association*, **24**, 73-85.

Yule, G. U. (1897) On the Significance of Bravais' Formulae for Regression, &c., in the Case of Skew Correlation, *Proceedings of the Royal Society*, **60**, 477-489.

_____ (1899) An Investigation into the Causes of Changes in Pauperism in England, etc., *Journal of the Royal Statistical Society*, **62**, 249-296.

_____ (1909) The Applications of the Method of Correlation to Social and Economic Statistics, *Journal of the Royal Statistical Society*, **72**, 721-730.

_____ (1911) *An Introduction to the Theory of Statistics,* London: Griffin.