December 1934

My dear Tedin,

There are a number of interesting points in your letter and I will do my best to sort them out.

Let me take first the discussion on your fourth page of the results of fitting regressions of heading date on

(1) mean temperature for the whole period March to May

and (2) partial regression on the three mean temperatures of these three months.

In this case it is important to note that the second regression includes the first i.e. that any regression formula of the first kind is one of the group of formulae which might have been given by the second method. Consequently the one degree of freedom separated as due to regression by the first method is wholly included in the three degrees of freedom separated by the second method. In fact we may sub-divide the total sum of squares as in the table below.

| | D/f | S.S. | M.S. | ½ log. |
|---|---|---|---|---|
| Single regression | 1 | 506.83 | | |
| Additional regressions | 2 | 13.18 | 6.59 | 2.0941 |
| Remainder | 9 | 5.99 | .6656 | .9477 |
| | 12 | 785.09 | z = | 1.1464 |

Since the residue is 5.99 after fitting the full formula, we may assign this to the remaining nine/ degrees of freedom. The two additional degrees of freedom representing any information which the full formula gives, which is not already included in the simple formula, have diminished the residue sum of squares from 19.17 to 5.99 and consequently ~~this~~ their content ~~to~~ is 13.18. The one degree of freedom of the simple formula of course contains the remainder 506.83.

Now one can settle the question whether treating the mean temperatures of the three months separately makes a significant improvement to the prediction by comparing the two degrees of freedom with the remaining nine, noting that this comparison will be legitimate just because these two groups have nothing in common. Hence, as in the remaining columns of the table, one can obtain z=1.1464 and note that the 1% point for two degrees of freedom against nine = 1.0411. So that the significance of the improvement effected by the ~~failure of~~ further formula just exceeds the 1% level.

In some of the earlier calculations of your letter pp. 3 & 4 you make similar comparisons between degrees of freedom or sets of them which are not mutually exclusive, but on the contrary have something in common. For such tests the z distribution is not valid, because obviously when two quantities have much of their content in common, their ratio ~~can~~ cannot easily be made so great as it will often be when they are wholly independent.

The test which I give on my first page is of course equivdlant to testing whether the ~~free~~ three regression coefficients on the mean temperatures of March, April and May shew or do not shew significant differences interse. But this is not the essential of the test. What is essential is to obtain mutual exclusive groups of degrees of freedom so chosen as to supply the test required.

Now suppose you have taken triple regression on the mean temperatures of March, April and May, and also another triple regression, using three other variates, such as frequencies of Maxima above or minima below chosen temperatures, then the only way it seems to me of getting these two regressions into the same picture is to compare both with the six-fold regression using all six independent variates. For the six degrees of freedom separated by such a six-fold regression must certainly include all that there ic in both the three degrees of freedom picked out by the first method and in the three picked out by the second method, whether these two sets have much in common or not.

Consequently, if the two methods to be tested are called (A) and (B) then a comparison between the results obtained by (A) and (A + B) arranged as on my first page, will enable you to say whether (A+B) supplies any significant information which is not supplied by (A). In other words, given the variates used in (A) do those used in (B) supply any additional information worth having?

The same test may be made comparing (B) with (A+B) and any number of results are possible.

It may be that neither the variates of (A) nor those of (B) give anything worth having in addition to that supplied by the other lot. In such a case all the information comes from what is common to the two lots. It becomes more a matter of convenience than anything else which method should be used in future studies, though one may also be influenced, not with great confidence, but taking a chance, to choose the method which actually leaves the smaller residual error.

In a second class of cases it will turn out that one method, say (B) adds nothing to what is already supplied by (A) but that the reverse is not true, but on the contrary (A+B) is significantly better than (B) alone. In this case (A) contains all that is worth having in (B) and something more in addition, consequently (A) will be chosen in preference to (B) for further work.

Finally, it may appear that both the independent variates used in (A) and those used in (B) supply information beyond that which is common to the two groups. In this case it will be necessary to use the more complex formula (A+B) in preference to either separately. If such a complex formula looks like being too complicated for general use, it may (or may not) be possible to simplify it by throwing out unproductive members of the sets of variates used initially in (A) and (B) (A+B) or to find a different set of fewer than six variates which will do as well as does (A+B).

With respect to the use of accumulated temperature, which I fancy is the official phrase in this country for what you call "sums of warmth~~s~~", would it not be simpler to take the value which gives the mean date correctly, rather than that which minimises the mean square error. If you do this I believe the sum of the squares of the errors divided by eleven should be comparable to the residual sum of squares for a simple regression.

I am very glad to hear how you are getting on, and especially that Bachér is proving useful. The general standard of statistical understanding has always struck me as so high in the Scandinavian countries that I have much looked forward to seeing my methods tried out thoroughly in the various researches in progress in your country.

Yours sincerely,