ST7002 Diploma in Statistics, Introduction to Regression

Partial R2 and Sequential/Extra Sum of Squares

A central challenge in multiple linear regression is the isolation and measurement of the (Expected/average) impact on *Y* of a single predictor say *Xa* where this stands for one on the predictor variables *X*1*, ….. Xp.* Indeed it can be said more generally that this is a central challenge in the statistical analysis of all data! The only way to do this unambiguously is to design the data collection such that the variation in *Xa* is ‘perfectly balanced’ with respect to all other *X* variables. A more general challenge is isolating a subset (of size 1, 2..) of *X* variables from the others.

Isolation and measurement can be approached in three ways: (i) via ; or (ii) via (decomposition of ) Sums-of-Squares SS (and hence *R*2.); and (iii) via ‘dropping variables’, but thus generating more coefficients and new SS and new *R*2.. How are all these inter-connected? The key is that each variable must - in general - be interpreted *having regard to the other variables that are in the model*. The Sequential SS shed some light on this, as discussed below.

(There are various *practical* strategies on ‘which to drop’, or the ‘best subset to retain’. Algorithms for ‘Stepwise’ and ‘Best Subsets’ selection are widely available. These ‘automatically’ work on equivalents to Seq SS. So the following discussion is aimed more at the conceptual task of ‘tying things together’ than as a recipe.

Further, although I have not dwelt overly on hypothesis-tests, Seq SS can also be used via *F* tests to ‘test’ hypotheses concerning subsets of *X* variables, for they allow the decomposition of the SS(Reg). Recall that the MINITAB ANOVA provides a test for ‘all coefficients jointly’; rejection of the via *F* test amounts to accepting a hypothesis that ‘at least one’ of the coefficients is non-zero. The t ratios allow tests for coefficients one-by-one.)

It is very easy to “adopt shorthand” and say “the coeff is (eg) 1.36”; so on average if *Xa* increases by one unit, *“Y* will increase by 1.36”. Better is “*Y* will increase by 1.36 if all the others are held constant”. Better again is “*Y* will increase by 1.36 if the following *explicit list* of -variables are held constant”. Ideally “coefficients” should be pronounced “partial coefficients”.

This is tedious, so is ‘partial’ and ‘explicit list’ are often ‘pronounced silently’; this however is the *central source* of confusion in interpreting, and/or communicating, findings. Only if great care has been taken in data collection are these statements *exactly* equivalent. (Variance Inflation factors provide explicit quantification of the extent to which these statements are *approximately* equivalent.) Minitab’s (default) Sums-of-Squares SS (‘Total’ and ‘Regression’ and, hence also, *R*2.) apply *jointly* to all explanatory variables, and have no partial (ie one-by-one) interpretation.

For reasons that are obscure, the word ‘partial’ is used to discuss statements that should only be uttered when accompanied by an *explicit* list of the ***other*** *X* variables; thus coefficients should *always* be interpreted in a ‘partial’ sense. Worse, other jargon terms: ‘Conditional’, ‘Sequential’ SS and ‘Extra’ SSQ are used in the literature. ‘Sequential’ is MINITAB’s preferred word (for such SS).

Note that changing the ‘order’ in which the *list* of *X* variables is presented to MINITAB will have an impact on the *values* of the Seq SS associated with each *X* variable. It will have no impact however on the *values* of the (partial) coefficients, or of *R*2; one set of *X* variables, presented in two *different* orders, will have the *same* set of coefficients, same t-values etc. Naturally, ordering will have no impact on the *sum* of the Seq SS which is SS Reg.

Apart from (partial) *coefficients* the two most useful *additional* ‘quantifiers of a *specific Xa*’ are based on Seq SS for the variable entered *first* and entered *last*. If entered first, the Seq SS is exactly the same as the SSfor a simple linear regression involving *Y* and only *this* *Xa*; ie dropping *all other* variables. Try it yourself. The associated *R*2 (ie 100 × Seq SS(*Xa* entered first) ÷ SS(Total) ) is then the 100 × (simple/regular correlation between *Y* and *Xa* )2.

If entered last, the corresponding Seq SS leads to (ie 100 × Seq SS(*Xa* entered last) ÷ SS(Total) ) is then 100 × (*partial* correlation between *Y* and *Xa* )2 implicitly defining the ‘partial correlation’ and a ‘partial *R*2 ’. If there are already in the regression (ie earlier terms) other *X* variables from which *Xa* ‘could be constructed’ then Seq SS(*Xa* entered last) will be small and ‘partial correlation’ and a ‘partial *R*2 ’ will be small. (Of course VIF will be high if *Xa* ‘could be constructed’ from the other *X* vars, so that’s another signal of the same; the connection to ‘variance inflation’ is technical, but closely related.)

If two variables, *Xa* and *Xb* play very similar roles in a model, it may be desirable to drop one; but which? One criterion is to drop the one that has the lower first-of-two Seq SS when entered both as (*Xa* then *Xb*) and (*Xb* then *Xa*), followed by the others. This coincides of course with preferring the one that has the higher simple correlation with *Y*. If other variables must be in the model *before* ‘one of *Xa* and *Xb*‘ then the choice may be still be made on the basis of ‘the lower first-of-two Seq SS’; but it’s no longer interpretable through ‘simple correlation with *Y’.* An alternative strategy is to form a a single ‘derived’ variable, eg 1/2(*Xa* then *Xb*).

Finally, Seq SS can help to see that adding more explanatory *X* variables must (i) increase the SS(Reg) but (ii) will typically exhibit ‘the law of diminishing returns’. This can be formalised by



Each of these could in principle be computed via Seq SS. All could be described as various versions of ‘partial’ depending on the explicit list of other variables in the model ‘thus far’. But the concept is more important. For example, there is no unique way to order the *X* variables. But it is clear from the above that all lead orderings lead to the same. We touched on this in Week 5.