



The use of F statistics in regression models from large pools of variables

David Livingstone^{1,2}, Subhash Ajmani² and David W Salt³

¹ChemQuest, Sandown, UK.

²Centre for Molecular Design, University of Portsmouth, UK.

³Department of Mathematics, University of Portsmouth, UK.

Background

- ① Work came about as a result of a review chapter on variable selection
- ① First presented at the EuroQSAR conference in Istanbul – 2004
- ① The simulator has been improved since then (faster) and we have chosen a different approach to modelling the results

Background

- ① Early 1970's – QSAR mostly consisted of multiple linear regression with a few tabulated descriptors
- ① The major exception was the use of molecular connectivity indices (controversial) but still with MLR for modelling

Background

- ① So, at this point QSAR modellers were limited by:
 - ① Modelling tools – MLR has restrictions
 - ① Data – tabulations are always incomplete
- ① Many people began to use other statistical and mathematical methods
- ① Computational chemistry began to provide new descriptors

Today

- ① There are dozens of modelling tools in use – each with their own pros and cons
- ① There is a huge range of descriptors to choose from (Todeschini & Consonni list over 3,100)
- ① But, MLR is attractive:
 - ① Easy to interpret
 - ① Widely available
 - ① Most frequently used

Used and Abused

- ① Some problems in the approach were recognised when John Topliss pointed out the danger of chance correlations
- ① Widely accepted but often misinterpreted
- ① Generates a single model (except GA) when many others may exist
- ① Selection bias (and others) leads to an “incorrect” assignment of significance



“At a rough guess, about 10^5 data sets per day are used as input to multiple regression packages around the world.
.....the results in this paper will be a nasty shock to many users of these packages, though the results have broadly been known for many years.”

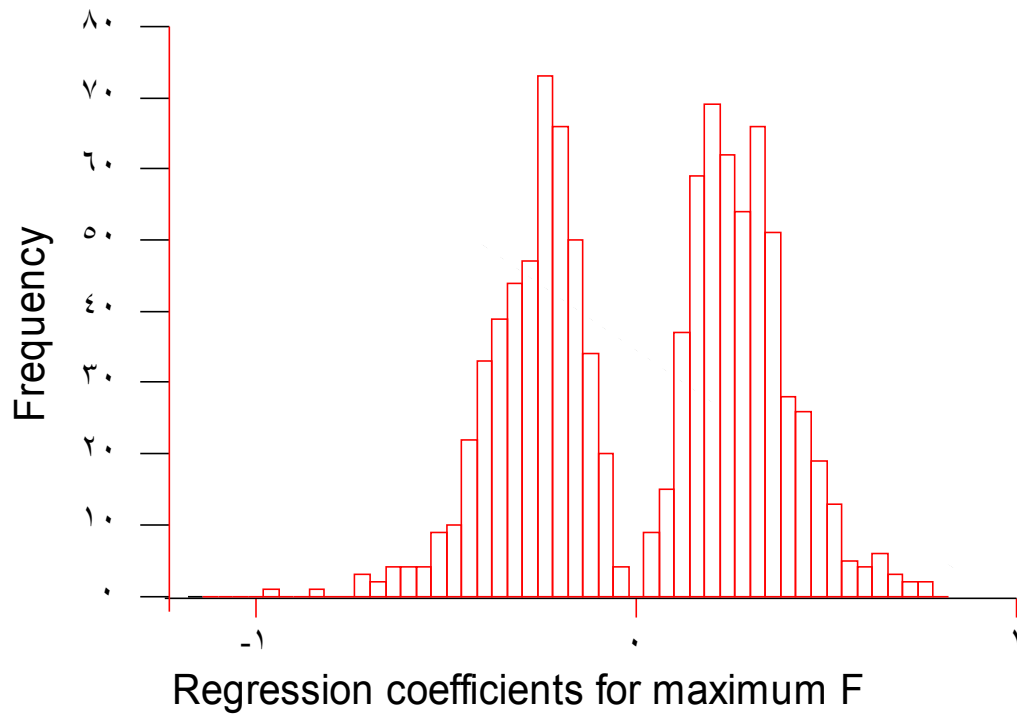
Comments by A.J. Miller on a paper presented by J.B. Copas at the Royal Stats. Soc. In 1983, Regression.Prediction and Shrinkage, J.R.Statist. Soc.B (1983)

Bias in Least Squares

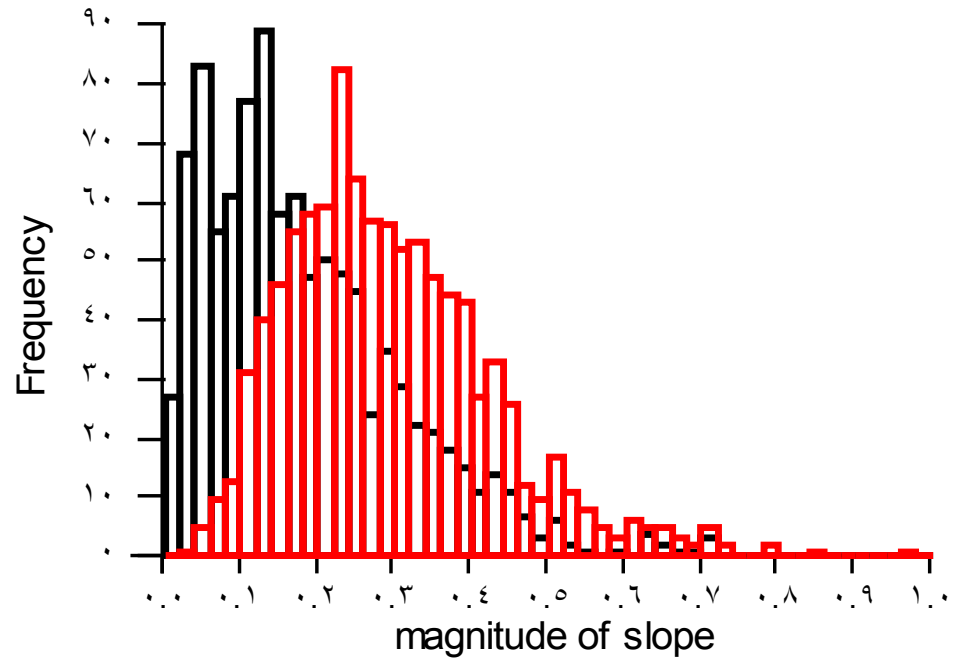
- ① Omission bias – if variables are removed from a “true” model the variance of predictions is reduced – but the remaining regression coefficients are biased
- ① Competition (or selection) bias – if variables are chosen in a supervised fashion then:
 - ① Regression coefficients are inflated
 - ① “significance” is also inflated

Selection bias

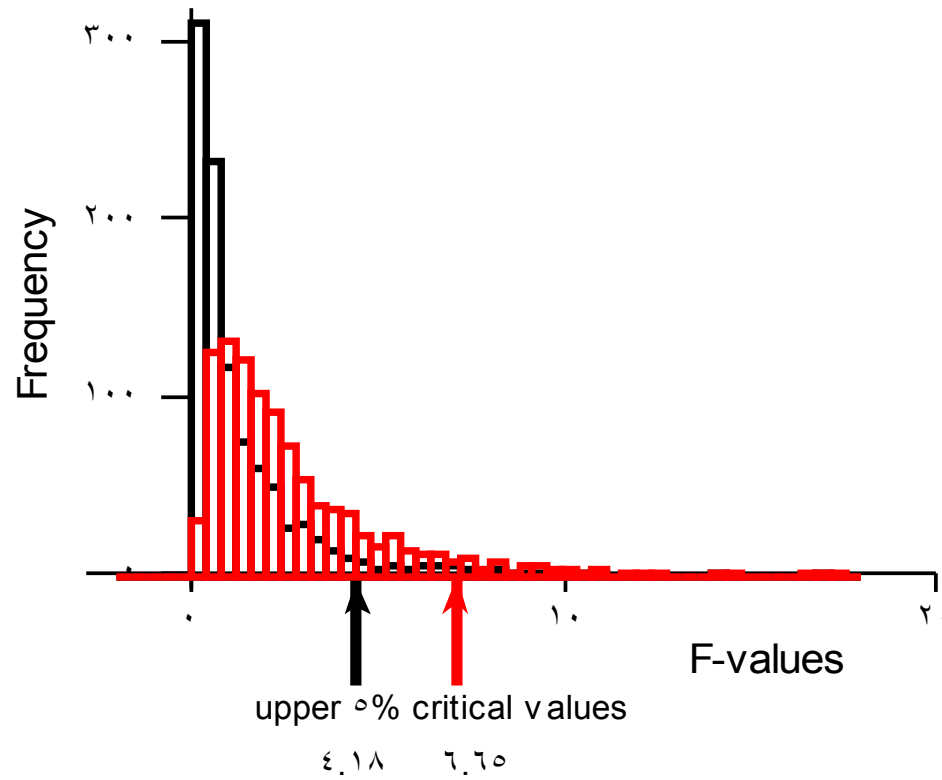
- ① 4 sets of 25 random numbers generated to represent y and 3 x variables
- ① Y regressed against one of the x variables, chosen at random, and the slope and R^2 recorded
- ① Y regressed against all three x variables and the max R^2 recorded
- ① Repeated 1000 times.



Random - Max



Random - Max





So What ?

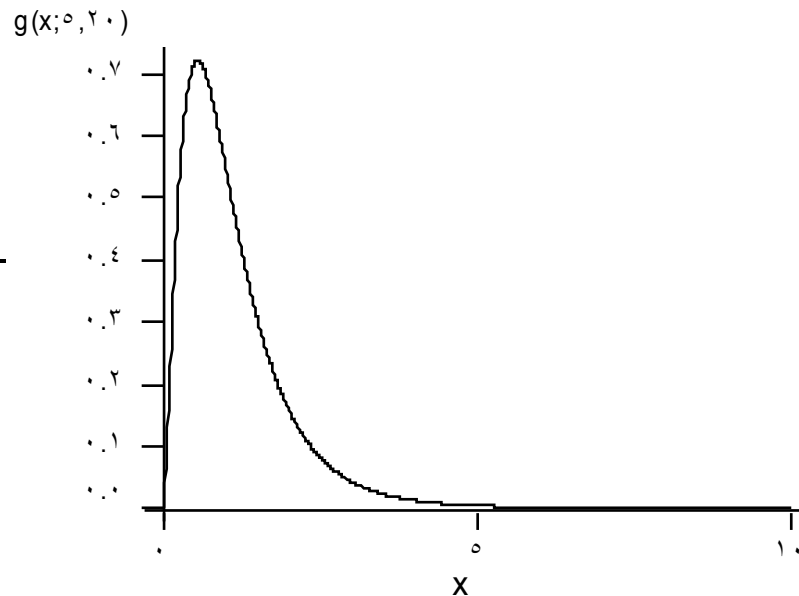
F-distribution

$$g(x; v_1, v_2) = \frac{\Gamma\left(\frac{v_1 + v_2}{2}\right)}{\Gamma\left(\frac{v_1}{2}\right)\Gamma\left(\frac{v_2}{2}\right)} \left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}} x^{\frac{v_1}{2}-1} \left(1 + \frac{v_1}{v_2}x\right)^{-\left(\frac{v_1 + v_2}{2}\right)}$$

where

$$v_1 = p$$

$$v_2 = n - p - 1$$



Distribution of F_{max}

① Theoretical approach

① $p=1$ case.

① Rencher and Pun (1980) Technometrics, 22. 1.

① Simulation approach

① McIntre, Montgomery, Srinivasan and Weitz (1983) J. Marketing Res.

① Both approaches are based on the distribution of R^2

F_{max} Simulations

- ① C++ software which generates random numbers and carries out regressions (PC and SGI versions)
- ① User specification of pool size (k), number of cases (n), size of models (p) and number of simulations
- ① Beware – a combinatorial problem !
 - ① $k=50, p=8$ gives $536 \cdot 10^6$ regressions

F_{max} Simulations

① A range of parameter values were used in the simulations:

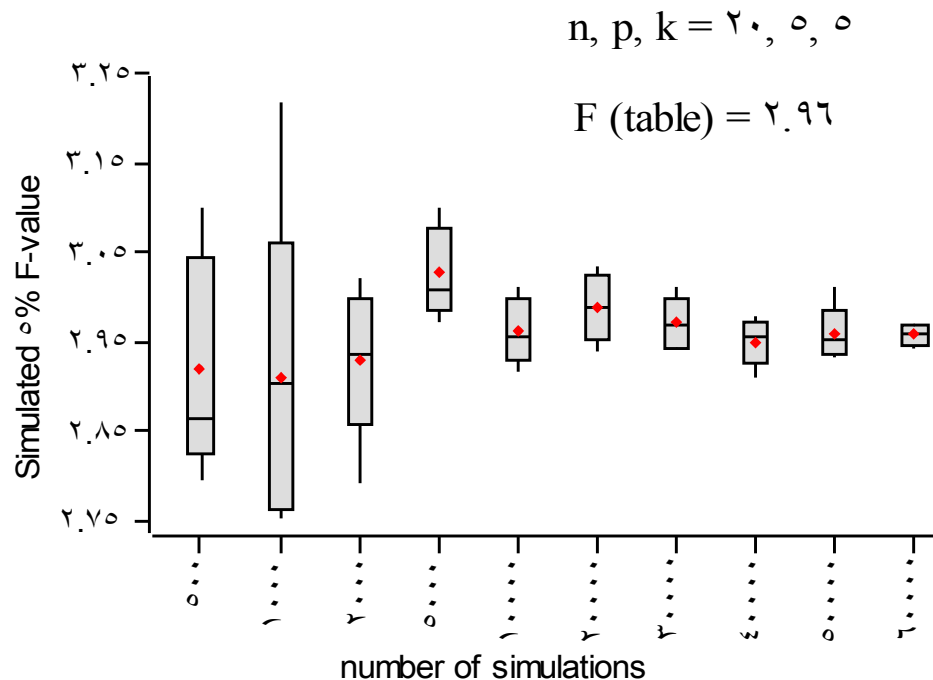
- ① number of observations n - 10,20,30,40,50,75,100
- ① number of variables in model p - 1,2,3,4,5,6,7,8
- ① number of available variables k - 5,10,20,50,100
- ① number of replications - 50000

① Timings

- ① $\text{time(s)} = 0.031(N^{0.951})\exp(0.382p+0.008n)$
- ① $N = k!/(p!(k-p)!)$

No. of Simulations

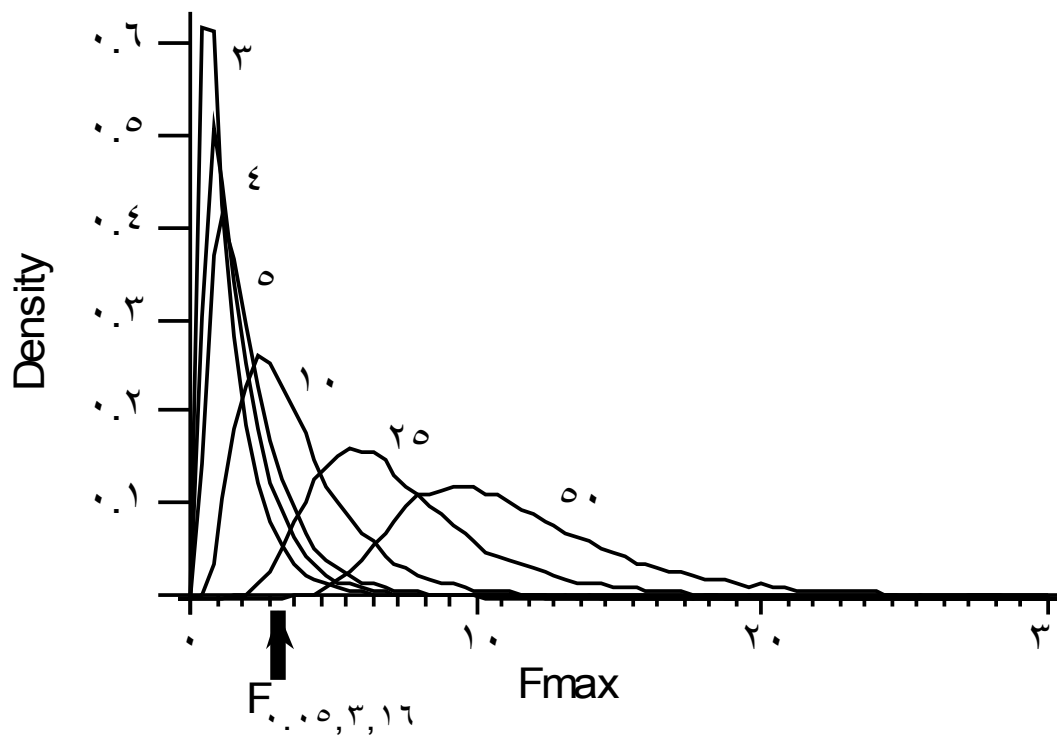
- 1 Early Monte Carlo approaches suffered from small number of replicates
- 1 Trade off between speed and precision



So, is this bias important?

- ① What is the effect of increasing the pool size for “reasonably” small regression models?
- ① How about a regression model of 3 variables for 20 data points?
- ① We ran the simulations for:
 - ① $n = 20$
 - ① $p = 3$
 - ① $k = 3, 4, 5, 10, 25, 50$

F_{max} values from simulations



Antimycin Example

- ① Originally calculated 53 descriptors which reduced to 23 after the removal of correlations (unsupervised)
- ① Supervised selection reduced this to 10 variables
- ① Training set of 16 compounds

Antimycin Example

$$-\log EC_{50} = 0.013mp - 1.97$$

$$R = 0.7 \quad F_{1,12} = 13.0 \quad SE = 0.08$$

$$-\log EC_{50} = 0.016mp + 0.06 \log P - 6.14$$

$$R = 0.86 \quad F_{2,12} = 18.0 \quad SE = 0.40$$

$$-\log EC_{50} = 0.017mp + 0.60 \log P - 0.81ESDL - 7$$

$$R = 0.9 \quad F_{3,12} = 17.0 \quad SE = 0.38$$

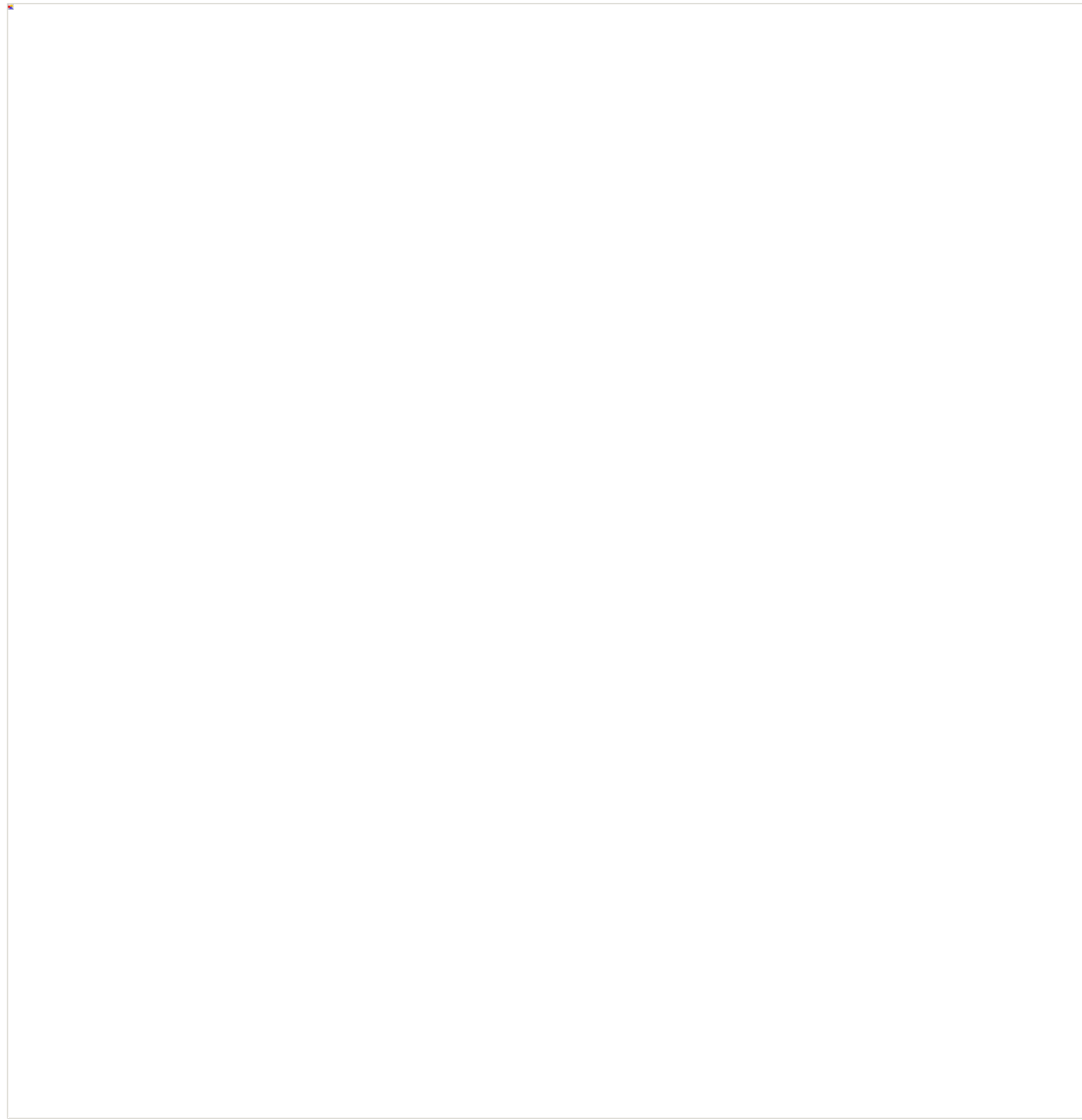
F “Statistics” Comparison

- ① 5 percentage values from tables, 95th percentile from F_{\max} simulations

terms	1	2	3
fit	13.55	18.00	17.50
Table	4.62	3.85	3.54
$F_{\max}(n)$	4.58	3.63	3.48
10 vars	11.06	10.03	9.80
23	13.88	14.85	17.39
53	17.22	21.11	29.73

Modelling the F_{max} Distribution

- ① The problem with simulations is that they probably won't contain the n , p and k combination needed (www.cmd.port.ac.uk)
- ① So, we tried to fit a power function to the simulation results



F_{max} Results

Power function e.g.

$$F_{\max} = \Phi(n, p, k)$$

$$\hat{F}_{\max} = \frac{3.3 n^{3.2} N^{0.2}}{p^{0.8}} e^{\ln(v_r) [\ln(v_r/n) - 0.1]}$$

where $v_r = n - p - 1$

94% of variation in simulated Fmax accounted for.

Variable Selection Inflation Index

① The F_{\max} results are always larger than the tabulated F values

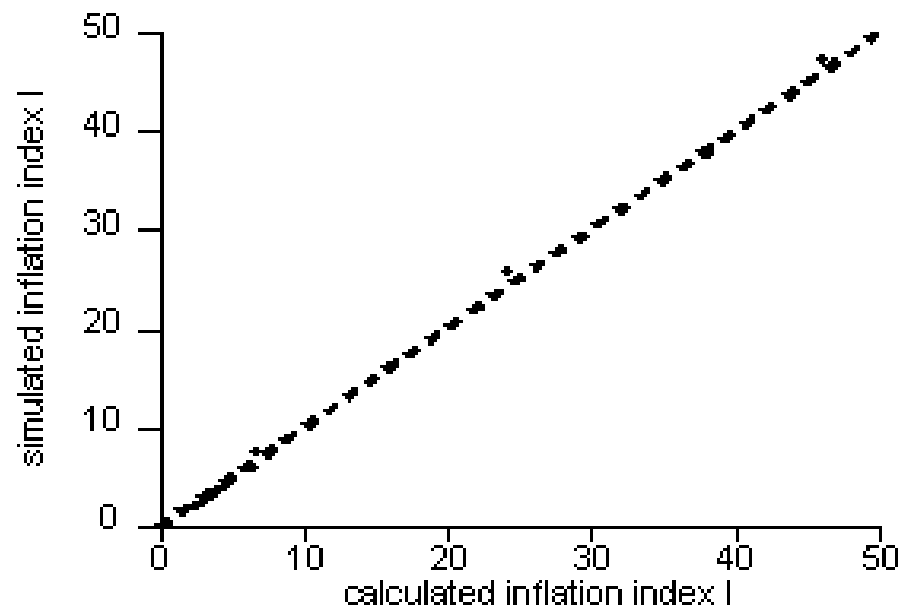
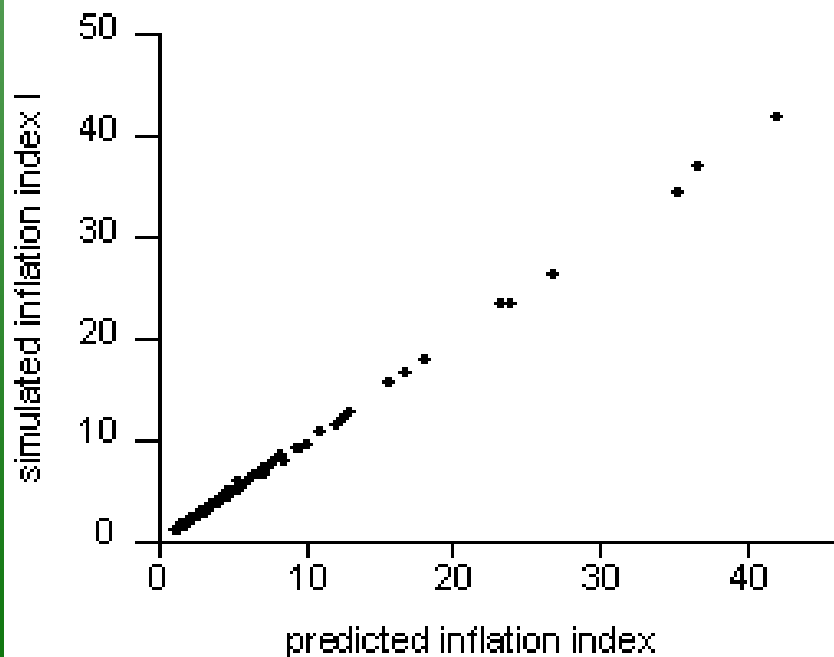
①
$$F_{\max, \nu_1, \nu_2, \dots, \nu} = F_{\nu_1, \nu_2, \dots, \nu} \times I(\nu_1, \nu_2, k)$$

① The inflation index, I , =1 for $p=k$ and is >1 for $p < k$

① Fitted a simple model:

$$I(\nu_1, \nu_2, k) = N^d$$

How well does it work?



Real Data

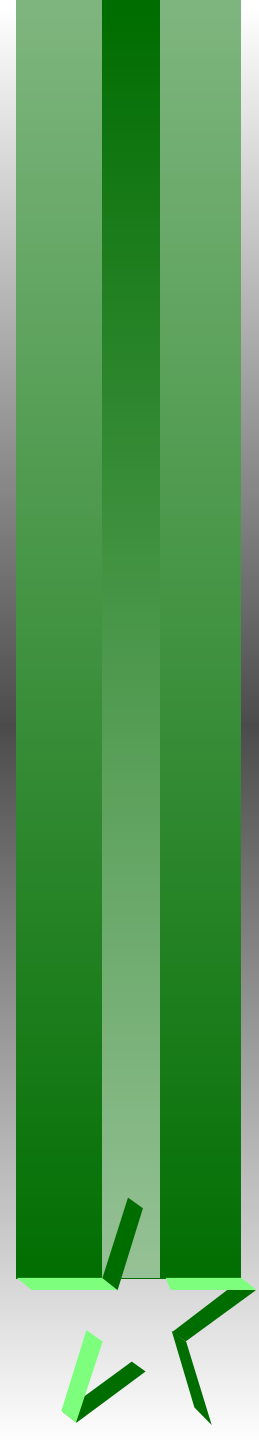
- ① All these results were obtained with random data
- ① It has been shown for stepwise regression that critical values for correlated variables are lower than for random
- ① We took 4 real data sets and scrambled the response, ran best subset and recorded F_{\max} , repeated 2000 times

Real Data

Data Set	n	k	p	Critical F_{\max} random	Critical F_{\max} real	Observed F -ratio	Usual table F value
Selwood	29	53	1	13.77	12.21	16.33	4.21
			2	13.42	10.88	24.64	3.37
			3	13.57	10.70	25.94	2.99
Kappa	35	28	1	11.52	11.03	24.69	4.14
			2	10.22	9.27	166.09	3.29
			3	9.44	8.52	182.22	2.91
Pyrethroid	19	34	1	14.32	12.68	5.09	4.45
			2	15.09	11.92	10.68	3.63
			3	16.75	13.35	9.91	3.18
Damborsky	15	9	1	10.94	11.58	12.40	4.67
			2	9.99	9.82	8.98	3.89
			3	9.61	9.32	8.26	3.59

Conclusions

- ① The results from the simulations suggest that the critical values we should use are considerably higher than the F tables
- ① The simulations are based on random data !
- ① “Complete” tables are impossible to generate by simulation so we have fitted an inflation index
- ① These results are for all subsets – critical F_{\max} will be \leq for the stepwise case



Variable Selection Inflation Index

- 1 Fitted a simple model:

$$I(v_1, v_2, k) = N^d$$

$$d = [a_1 \ln(v_1 + v_2 + 1) + a_2 \ln(v_1) + a_3 [\ln(v_1)]^2 + a_4 \ln(k) + a_5 \ln(v_2) + a_6 \ln(N) + \frac{a_7 \ln(v_2) + a_8 \ln(v_1)}{\ln(v_1 + v_2 + 1)} + a_9 \frac{\ln(v_1)}{\ln(k)} + a_{10}]^2$$

Distribution of F_{max}

① Theoretical approach

Now
$$f = \frac{\text{regression mean sq}}{\text{residual mean sq}} = \frac{R^2/p}{{(1 - R^2)/(n - p - 1)}}$$

$$b(y; a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} y^{a-1} (1 - y)^{b-1} \quad \text{Beta distribution}$$

where

$$y = R^2 \quad a = p/2, \quad b = (n - p - 1)/2 \quad \text{and} \quad 0 < y < 1$$

Distribution of F_{max}

① Theoretical approach

The 95% F_{max} value is the 95th percentile of the distribution of F_{max} .

If for given p and k there are N models

$$f_1, f_2, \dots, f_N \longrightarrow R_1^y, R_2^y, \dots, R_N^y$$

i.e. we can convert the problem to finding the distribution of $R_{max}^2 (= y)$

Distribution of F_{max}

① Theoretical approach: $p=1$

The critical 95% value, $R^2_{0.95}$, can be shown to be the solution of the following equation

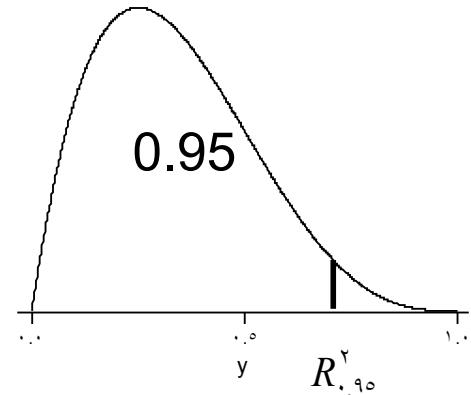
$$\int_0^{R^2_{0.95}} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1} (1-y)^{b-1} dy = (0.95)^{\frac{1}{N}}$$

where

$R^2_{0.95}$

$$y = R^2 \quad a = \frac{p}{r}, \quad b = \frac{(n-p-1)}{r}$$

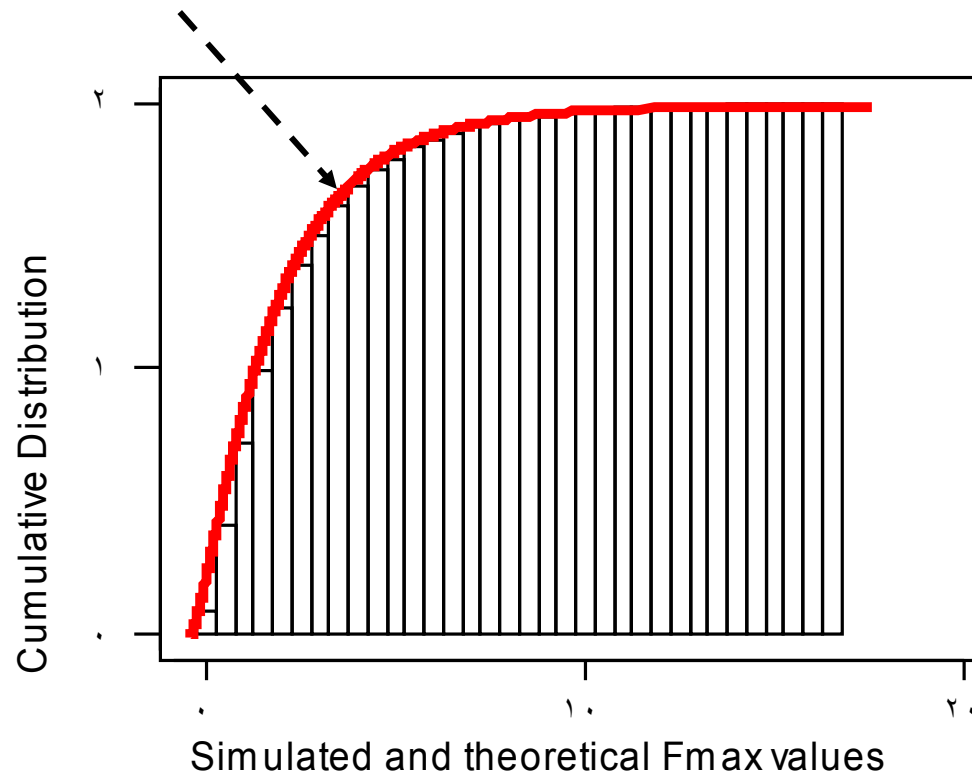
$$a = \frac{1}{r}, \quad b = \frac{(n-r)}{r}$$



and $0 < y < 1$

Distribution of F_{max}

theoretical



Distribution of F_{max}

① Theoretical approach : $p > 1$

The critical 95% value $R^2_{0.95}$ is given by

$$\int_0^{R^2_{0.95}} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1} (1-y)^{b-1} dy = \Phi(a, b, N, 0.95)$$

$$y = R^2 \quad a = \frac{p}{\nu}, \quad b = \frac{(n-p-1)}{\nu} \quad \text{and} \quad 0 < y < 1$$

Distribution of F_{max}

Theoretical approach (continued)

Rencher and Pun, using an argument based on Extreme Value statistics, give the approx. critical value (5%) of R^2 as the solution to

$$\int_0^{R_{.95}^*} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1} (1-y)^{b-1} dy = 1 + \frac{\ln(0.95)}{(\ln N)^{1.4N^{0.4}}}$$

$$N = k!/(p!(k-p)!)$$

where p is the size of the regression model and k is the size of the independent variables “bucket”