Why do covariate adjustments blow up?

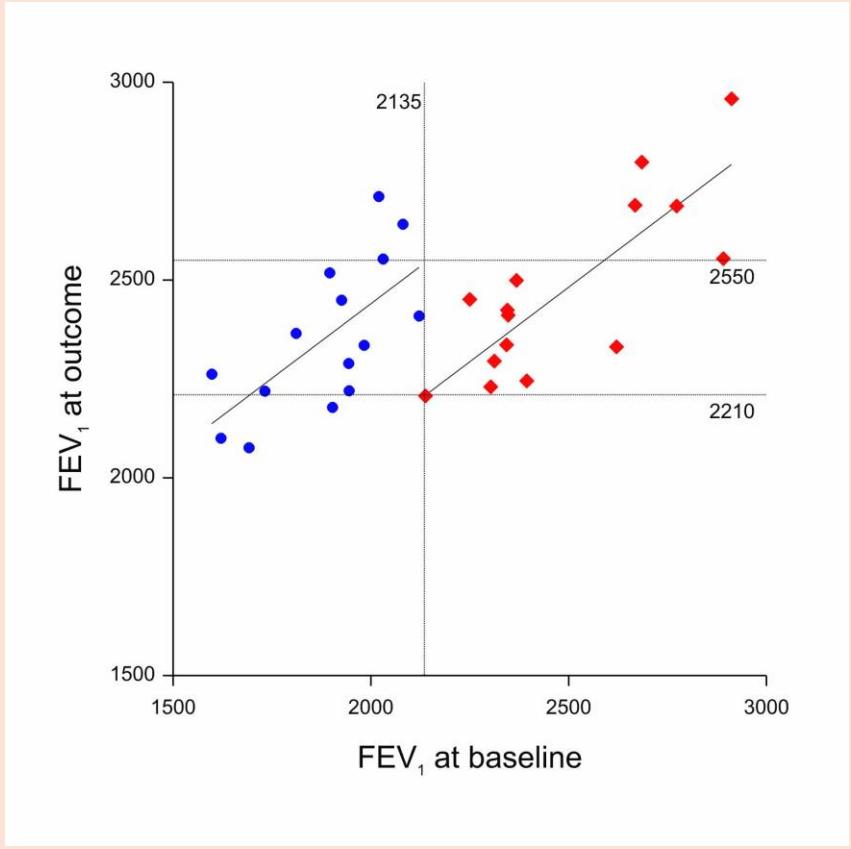
Understanding ANCOVA with variance inflation factors.

Stephen Senn

Consultant Statistician, Edinburgh

Honorary Professor Sheffield & St Andrews Universities

Adjunct Professor, Medical University of Vienna



Acknowledgements

- Thank you for agreeing to my request to invite me!
- This work is joint with Franz König and Martin Posch.
- We thank our colleague Robin Ristl for drawing our attention to the paper by Qu.
- We thank Tim Morris for helpful comments on an early draft of a paper and Alan Agresti for comments on categorical variables.

Outline

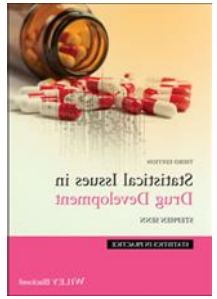
- The three effects of covariate adjustment
  - Mean square error
  - Variance Inflation factors (VIFs)
  - Second order precision.
- A general investigation of VIFs for continuous covariates.
- An example.
- Categorical (including binary) covariates.
- Implications and applications.
- Conclusion.

- NB The context is randomised clinical trials.
- The focus is on the estimate of the treatment effect.
  - Covariate effects are not of interest in themselves.
- I shall only consider adjusting for the main effect of covariates.
- Treatment by covariate interaction is not considered.
- Non-linear cases are not considered.

Three Effects of Adjusting for Covariates

The power of three

The power of three



Chapter 7

- To the extent that the covariate is prognostic of outcome, the expected value of the residual *mean square error* is reduced.
  - A source of 'noise' is identified and removed from the signal.
- To the extent that the covariate is not orthogonal to treatment (that is to say is not balanced), the *variance multiplier* used with the residual mean square error will increase.
  - Ambiguity between what is the effect of treatment and what is the effect of covariate exacts a price in precision.
- Because you have used up a degree of freedom, *second order precision* is worsened.
  - The noise that remains is less but your ability to estimate how much noise remains is reduced.

- For example: confidence limits for the difference between two treatment groups in a parallel group trial.

$$\hat{\delta} \pm SE \times t_{\alpha/2,\nu}$$

$$SE = \sqrt{\lambda\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\hat{\sigma}^2}$$

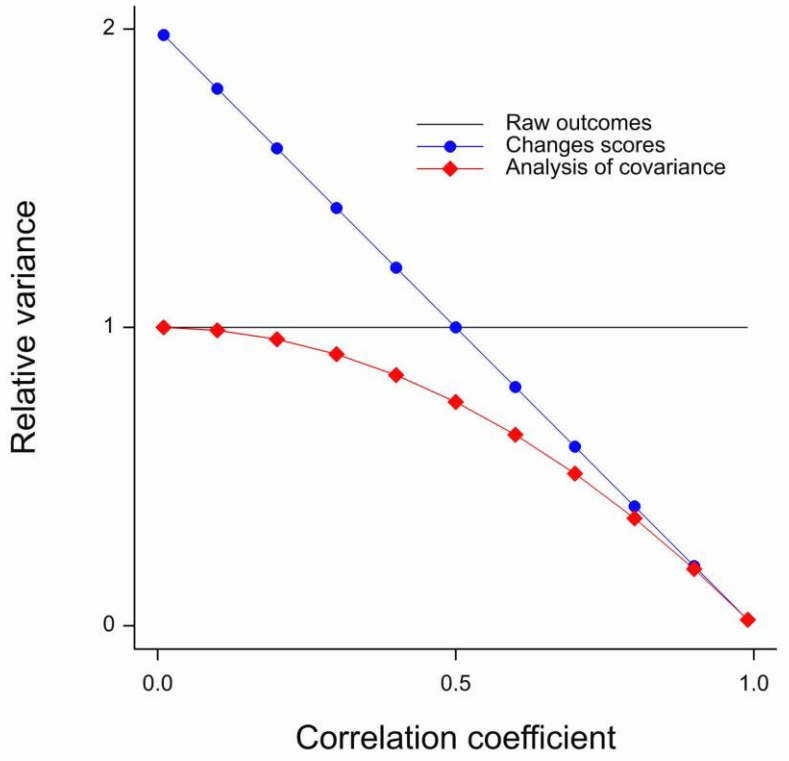
3. Residual degrees of freedom will reduce if you fit covariates and this multiplier will increase.

2. Variance multiplier will increase (λ>1)if you add unbalanced covariates.

1. Mean square error will reduce if you add prognostic covariates.

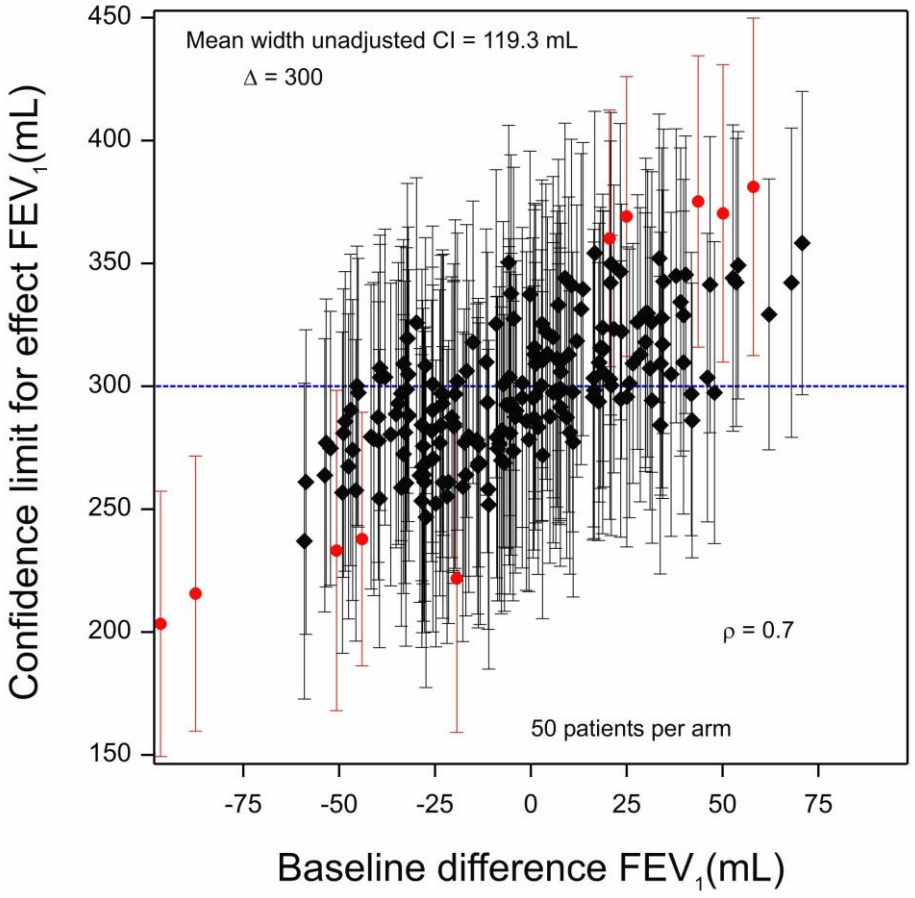
1. Mean square error effect

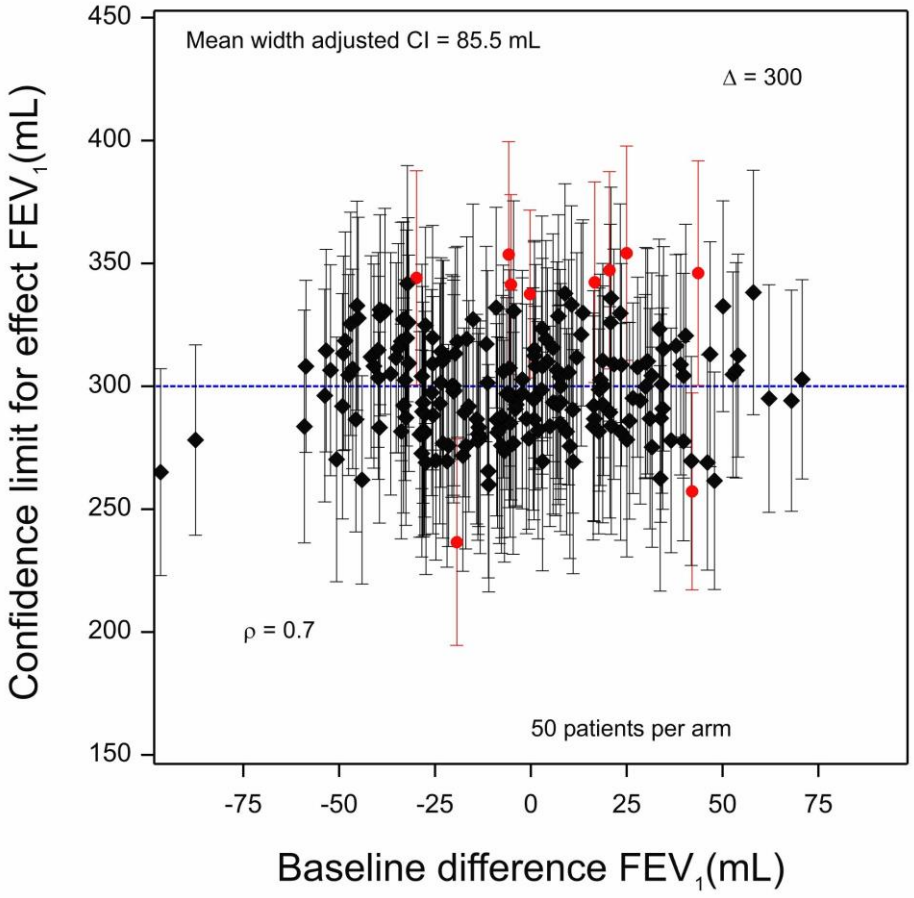
- A simple illustration of this can be given by considering the case where the covariate is a baseline measure corresponding to the outcome measure.
- The diagram shows the effect on the variance of the treatment estimate as a function of the correlation between baseline and outcome.
- Three approaches are used:
  - Ignoring the baseline
  - Using the change score
  - ANCOVA.



1. Mean Square Error Effect (cont)

Illustrated by simulation of 200 trials





2. Variance Inflation Factor

Standard theory of ordinary least squares

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \mathbf{I}_N$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma_\varepsilon^2$$

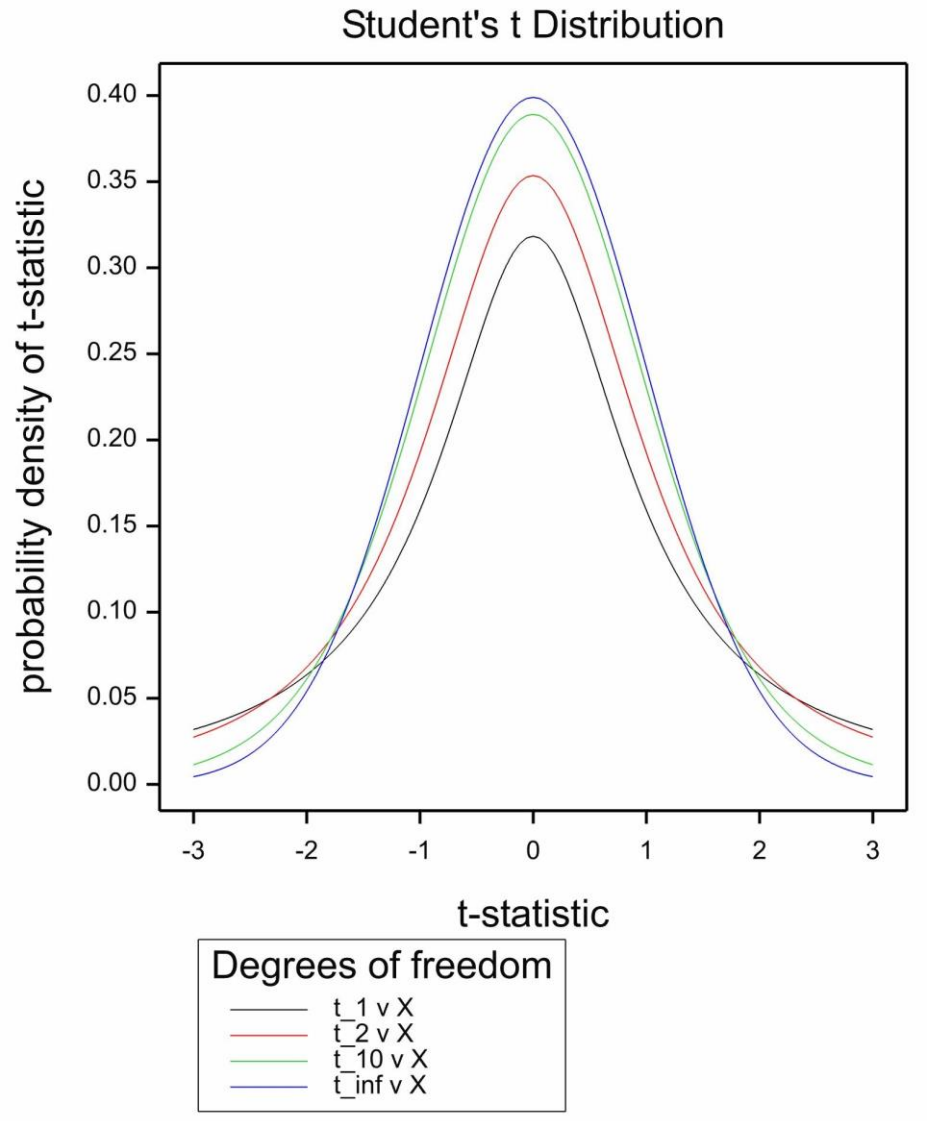
$$= \begin{pmatrix} a_{11} & & & \\ a_{21} & a_{22} & & \\ \vdots & \vdots & \ddots & \\ a_{k1} & a_{k2} & & a_{kk} \end{pmatrix} \sigma_\varepsilon^2$$

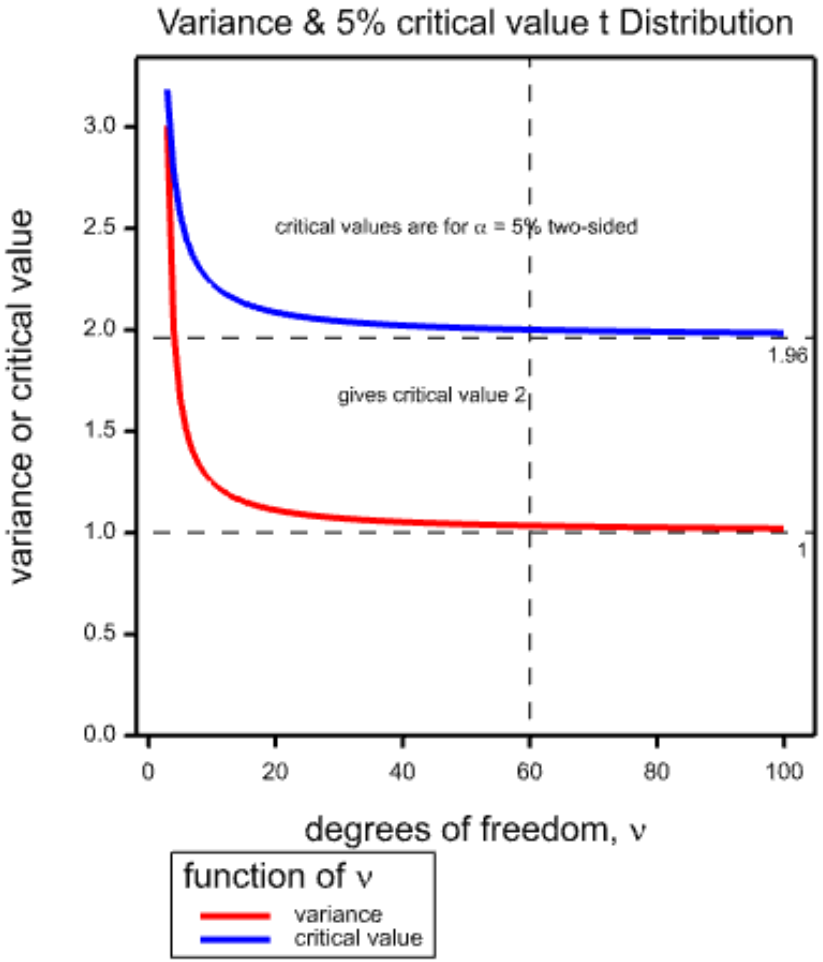
- The variance of the treatment estimate is

$$a_{22}\sigma_\varepsilon^2 = \lambda\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma_\varepsilon^2.$$

- The variance inflation factor is $\lambda \geq 1$ with the lower bound only being obtained if all covariates are orthogonal.
- $\lambda = 1/(1-R_z^2)$ where R_z^2 is the coefficient of determination for the treatment indicator using the covariates to be fitted.
- For a categorical covariate with k+1 categories we have $R_z^2 = Chi_k^2/N$ for the treatment x covariate contingency table.

3. Second order precision





The relationship of the three to design, model and outcome

		Effect		
		Estimated mean square error effect	Imbalance effect or VIF	Second order precision
Influence	Design	No*	Yes	No
	Model	Yes	Yes	Yes
	Outcome	Yes	No	No

*Provided that the the design does not balance for prognostic factors that are not in the model.

The value of looking at it this way

- Although the first and third terms are affected by the choice of model.
  - Expected mean square error is a function of prognostic power of the model.
  - Second order precision is a function of the number of parameters fitted.
- They are ***not*** affected by the design.
- It is thus confusing to include them in the consideration when comparing designs. For example:
  - Completely randomised
  - Median stratified.
- If the model is fixed, the ***design*** then governs the variance inflation factor.

A general investigation of variance inflation factors for continuous covariates

Blowing in the wind

Lines of a derivation

1. Working with a partitioned inverse of the design matrix you can show that the variance inflation factor is given by $\lambda = 1/\left(1-R_Z^2\right)$, where, R_Z^2 is the coefficient of determination of the treatment indicator *Z* regressed on the covariates. Note that the greater the degree of balance, the less "predictive" the covariates are of treatment, the lower the value of the coefficient of determination and the closer λ is to 1.
2. Using the fact that multivariate discriminant analysis is equivalent to linear regression on the treatment indicator, you can show that if the covariates have a multivariate Normal distribution, we have the relationship $\lambda = 1 + \frac{k}{N-k-1} F_{Rao}.$
3. However, given the degrees of freedom of the F distribution, its mean and variance are known.
4. Therefore, since, λ is a linear combination of an F-statistic, we can obtain the mean and variance of λ.
5. The results where there are *k* covariates and *N* subjects are

$$E[\lambda] = 1 + \frac{k}{N-k-3}, \quad k \le N-4.$$

$$Var[\lambda] = \frac{2k(N-3)}{(N-k-3)^2(N-k-5)}, \quad k \le N-6.$$

Notes

- These formulae *do* assume multivariate Normality.
- They do *not* depend on the correlation structure, except that it is assumed that the matrix of *k* predictors is of full rank.
- In fact, even this is not necessary; it is simply required that in the formulae, *k* is used to mean the *rank* of the predictor matrix and not the *number* of covariates.
- The distribution of the outcome variable is completely irrelevant.
- Note that this is a feature of the linear model. Non-linear models *do* engender a dependence on the outcome.

An Example

An incomplete blocks design made even less complete

A highly modified trial in asthma

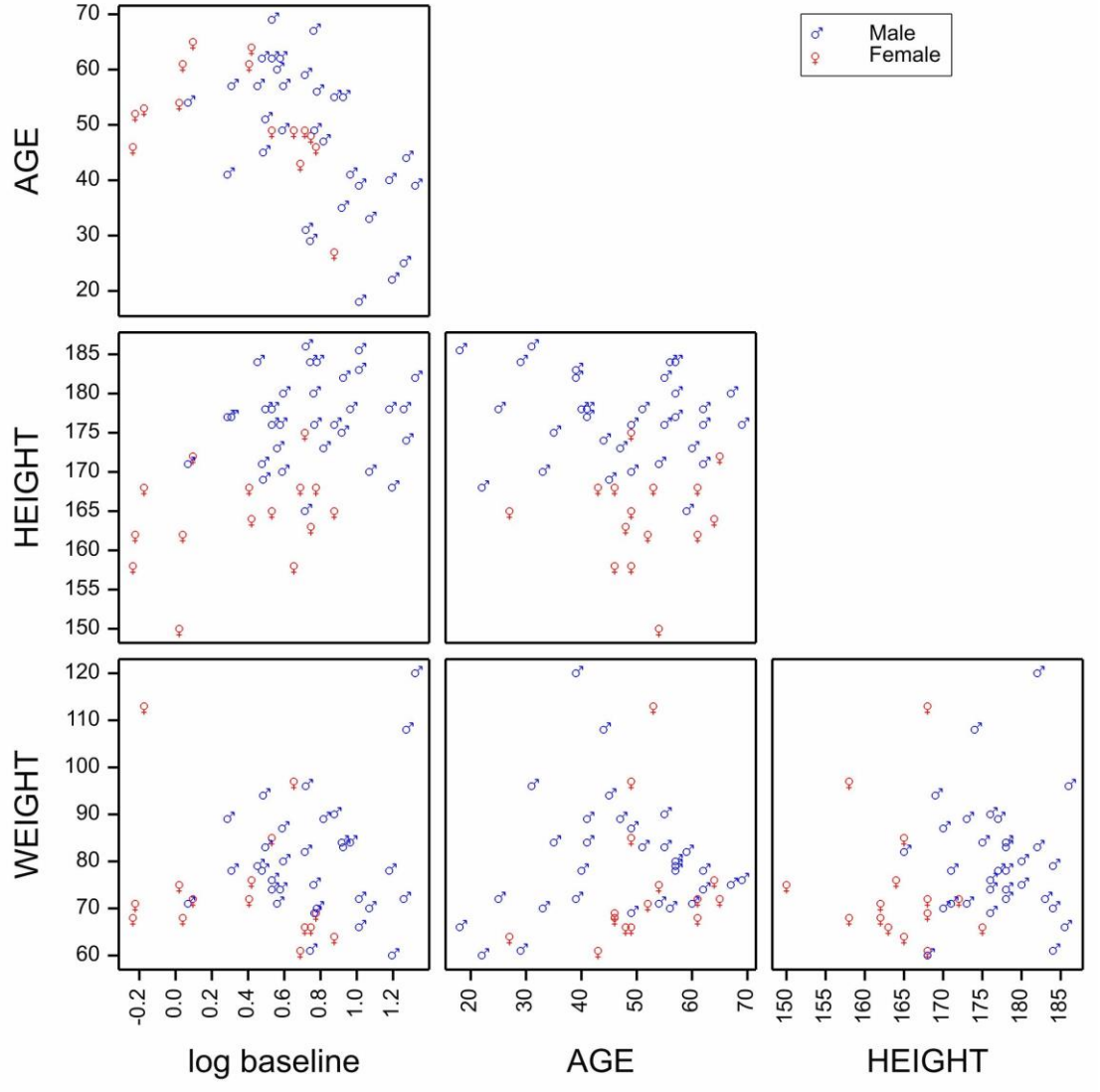
The trial

- This was an incomplete blocks cross-over trial in 7 treatments (3 doses of each of 2 formulations and a placebo), 5 periods and 21 sequences.
- I am only going to use data from period one for two of the treatments: placebo and the highest dose of one of the formulations.
- We thus create a two-arm parallel group design.

The data

- We have data from 46 patients.
- The 5 covariates available are
  - Sex
  - Height
  - Weight
  - Age
  - Log baseline forced expiratory volume in one second, FEV_1.
- Since we are investigating the VIF only, the outcome is not needed.

Baseline Covariate Distribution



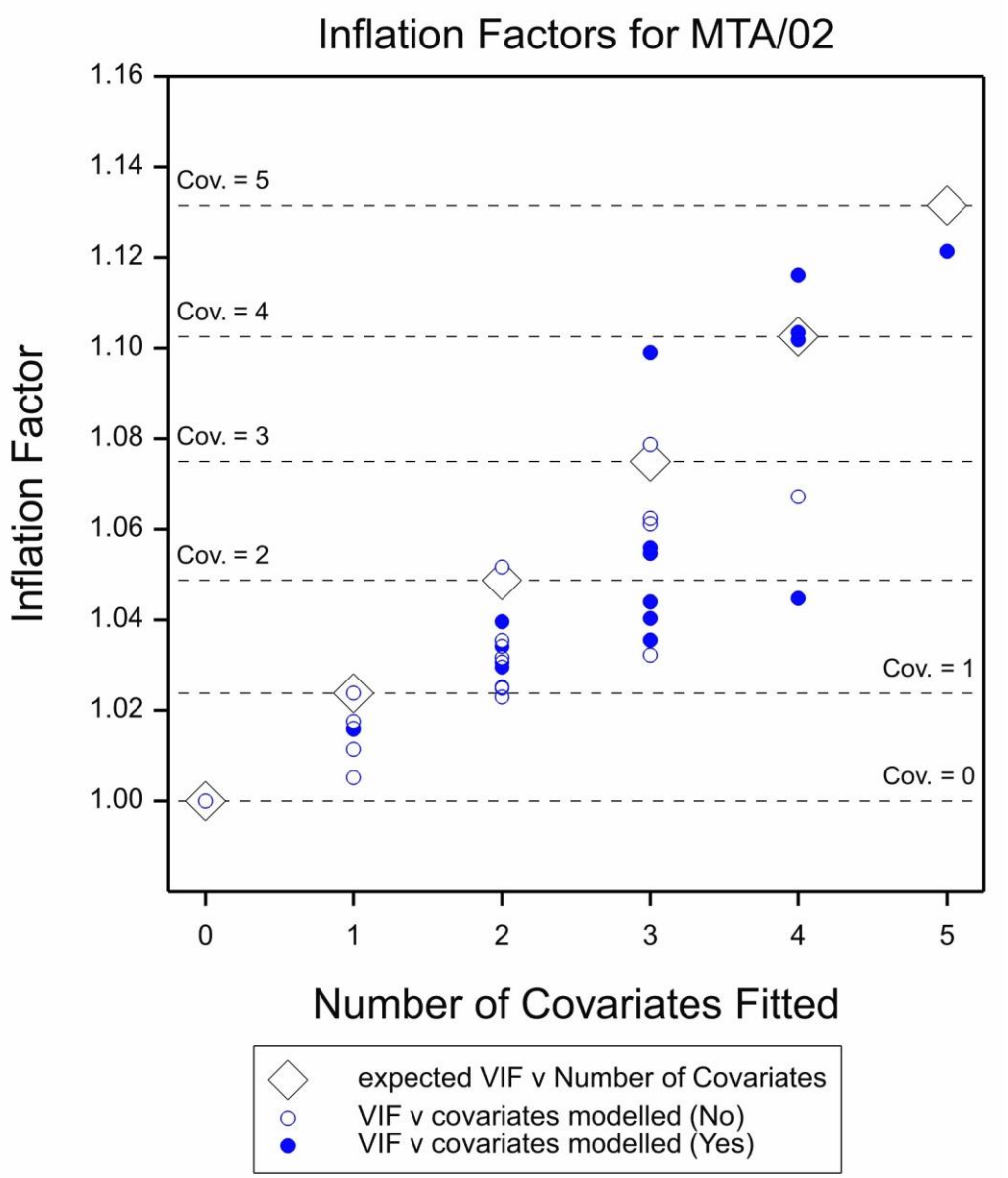
Investigation by simulations

Number of Covariates	Number of models
0	1
1	5
2	10
3	10
4	5
5	1
Any number	$2^5 = 32$

Sampling Plan	Notes
Multivariate Normal	Five dimensions. Observed correlation structure. One variate dichotomised to create 'sex'
Permutation	Covariates fixed. Treatment labels randomly switched.
Bootstrap	Each patient is randomly assigned (with replacement) one of 46 sets of covariates. The treatment is fixed.

Variance Inflation Factors (VIFs) for all 32 models for the actual data

- The VIF is 1 when no covariates are in the model
  - This must be the case
- If all 5 covariates are added, then for this example the VIF is 1.121.
- The expected value for 5 is 1.132.
- On the next few slides various re-sampling approaches will be used to examine the distribution of VIFs and check it against theory.



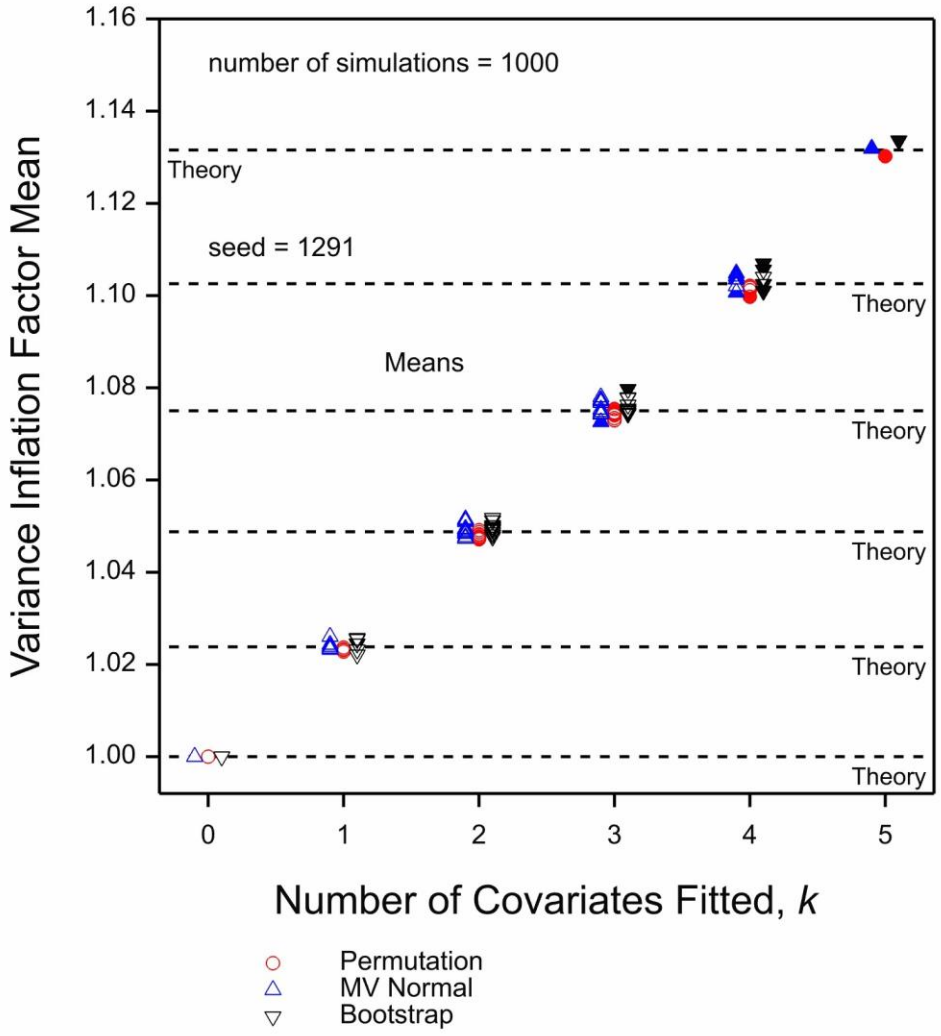
Mean inflation factors for 3 simulation approaches

Theoretical values

$$\frac{N-3}{N-k-3}$$

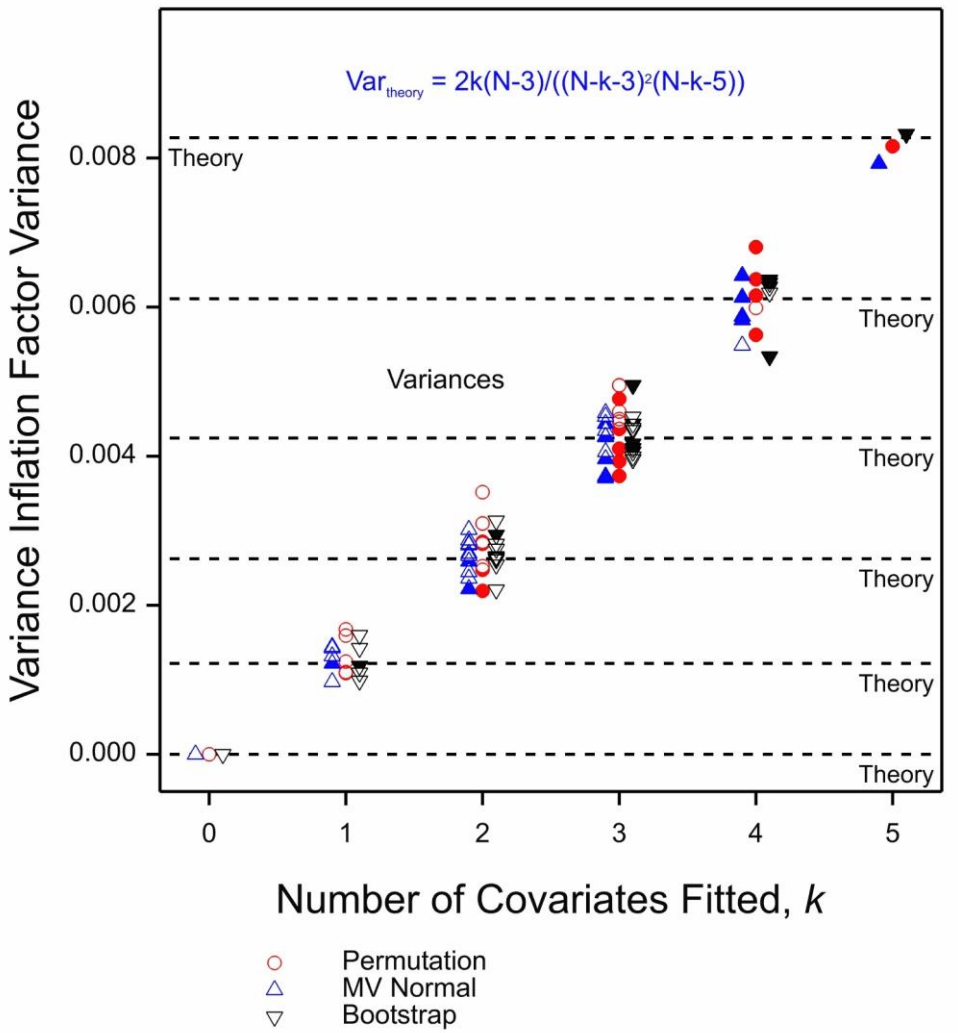
Where *N* is the number of patients and *k* is the number of covariates given by dashed horizontal line.

Symbol open, ***sex*** is not in the model. Symbol closed, ***sex*** is in the model.



Inflation factor variances for 3 simulations approaches

- Again, theoretical values are given by dashed horizontal lines.
- Symbol open, ***sex*** is not in the model. Symbol closed, ***sex*** is in the model.



Conclusions and a Warning

Conclusion

- The formulae work well.
- It does not seem to matter what form of sampling is used.
- It does not seem to matter whether a model includes the binary covariate or not.
  - Caution: for very small samples it will matter a bit.

Warning

- This is only one example!
- However, it gives reasonable hope that the theory will be a good starting point and guide for any simulation.
  - In particular, because for sampling from the multivariate Normal, *any* example will do.

Categorical predictors

This and that

Basic set up

	1	2	3	…	k+1
A	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$		$n_{1,k+1}$
B	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$		$n_{2,k+1}$

Contingency table showing the distribution of the numbers by category and treatment for a categorical predictor.

A binary covariate, such as sex in the example we had is just a special case with k=1.

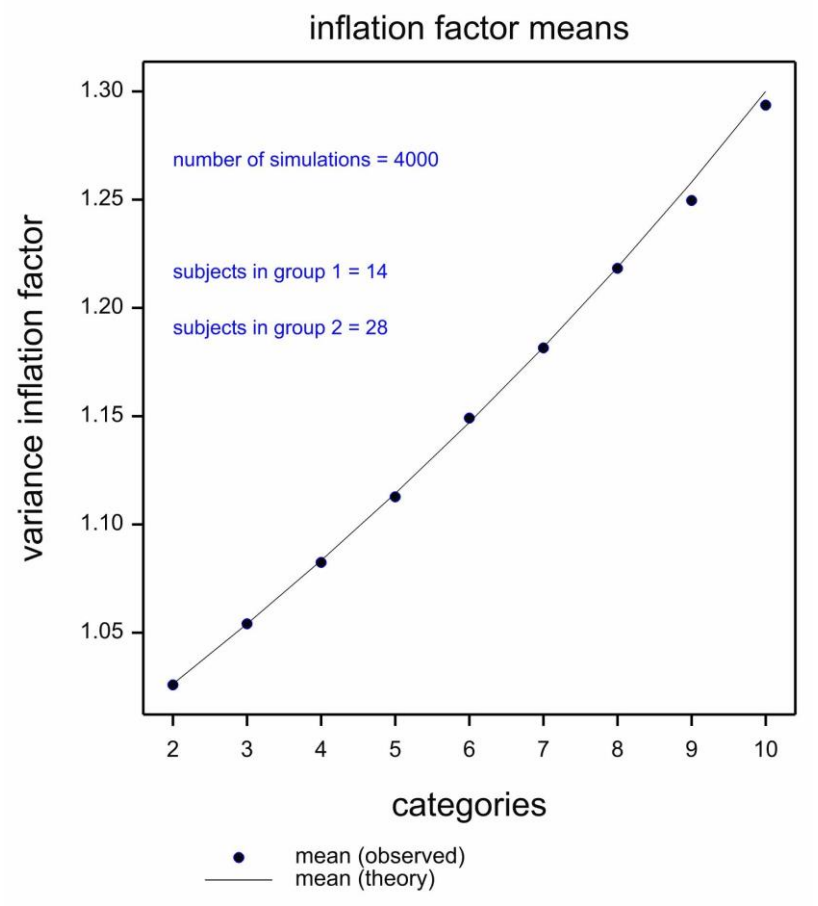
Note that perfect balance implies that the observed numbers are equal to the expected numbers calculated from the margins.

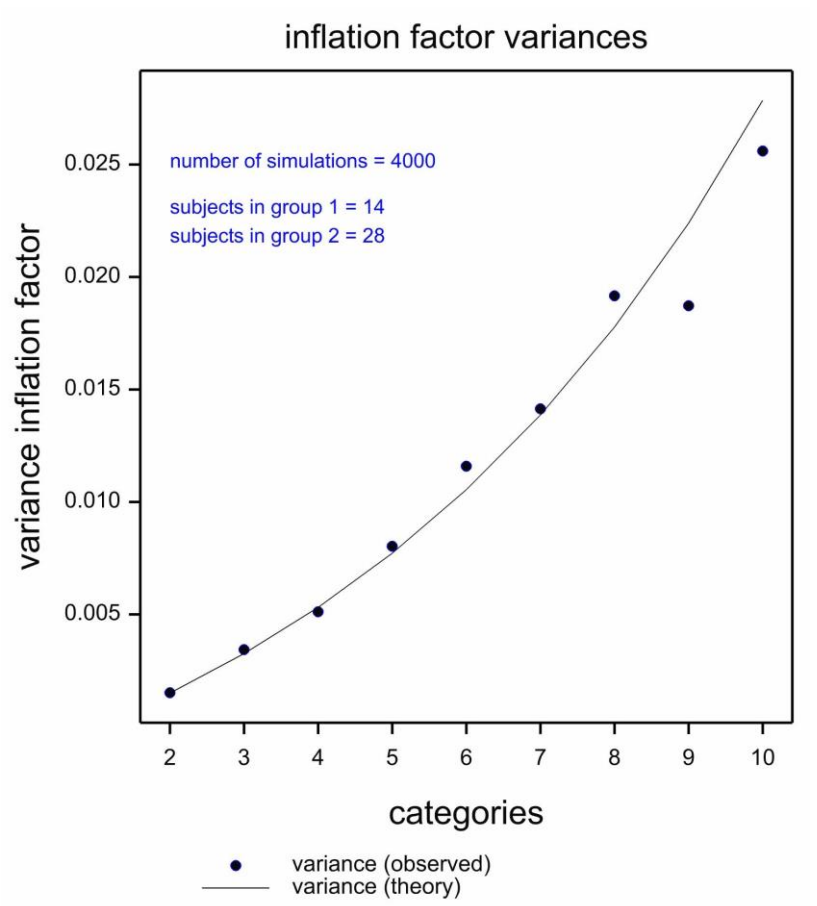
This also implies that the chi-square statistic will be zero.

In fact, it turns out that the variance inflation factor is a function of the chi-square statistic, χ^2.

$$\lambda = \frac{N}{N - \chi^2}$$

Simulation of Variance Inflation for a Categorical Covariate





Implications and Applications

Hopefully, without complications

Some possible questions regarding covariate adjustment will be addressed.

- We assume throughout.
  - There will be no “double-dipping fishing expeditions”.
  - Models used in analysis must be pre-specified.
- We do not claim that we have found perfect answers to these problems.
  - We are just illustrating that the three factor framework can be a useful way of thinking about them.

Example 1. Should you add another covariate?

Assume that you have *k* covariates in the model with ν residual degrees of freedom and wish to add another. As regards the VIFs the relevant consideration is the ratio

$$r_\lambda = \frac{N-k-3}{N-k-4} = \frac{\nu-1}{\nu-2} = 1+\frac{1}{\nu-2}.$$

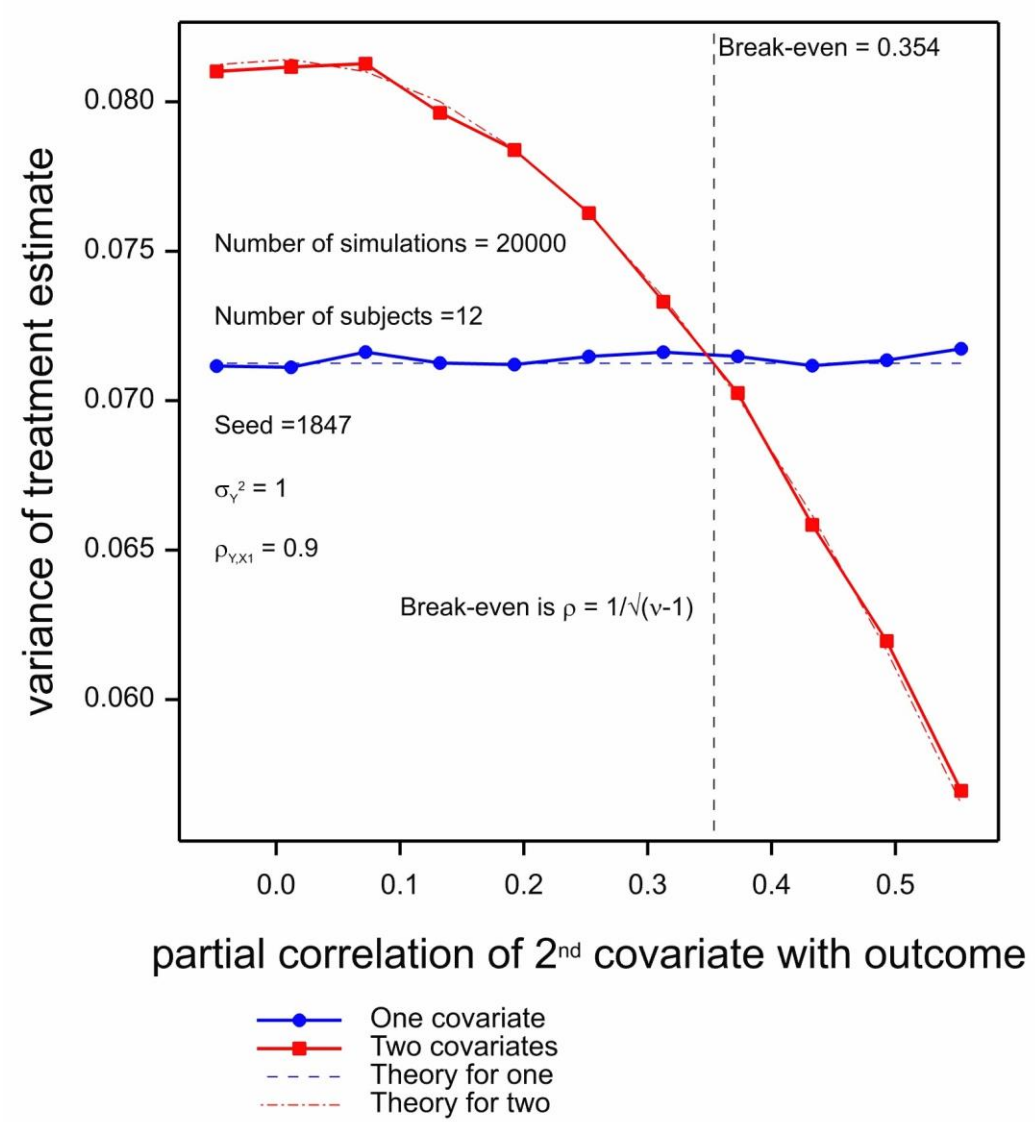
On the other hand, as regards MSE the relevant consideration is their expected ratio,

$$r_{MSE} = 1-\rho^2_{k+1|k},$$

where the partial correlation of the outcome with the new covariate given other is the model is $\rho_{k+1|k}$

The break-even point is

$$\rho_{k+1|k} = \pm\frac{1}{\sqrt{\nu-1}}.$$



Caution

- This ignores the issue of second order precision.
- This is a controversial subject and there are various approaches that we discuss in our paper.
- I don't propose to discuss this further here but suggest that a possible rule of thumb worth investigation is that if

$$\rho_{k+1|k} > \pm\frac{1}{\sqrt{\nu-2}}.$$

it is worth adding a covariate to the model

Example 2: What to do with median stratification?

Basic Set Up

- Suppose that we have one continuous covariate, *X* and a treatment indicator *Z*.
- We try and balance the covariate by using stratification using the predicted median.
- What should we fit?
  - The stratum indicator, *S*?
  - The continuous covariate, *X*?
  - Both?
  - Neither?
- What have we gained by stratification?

Qu's classification

$$Model(A): Y \sim Z$$

$$Model(B): Y \sim Z + X$$

$$Model(C): Y \sim Z + S$$

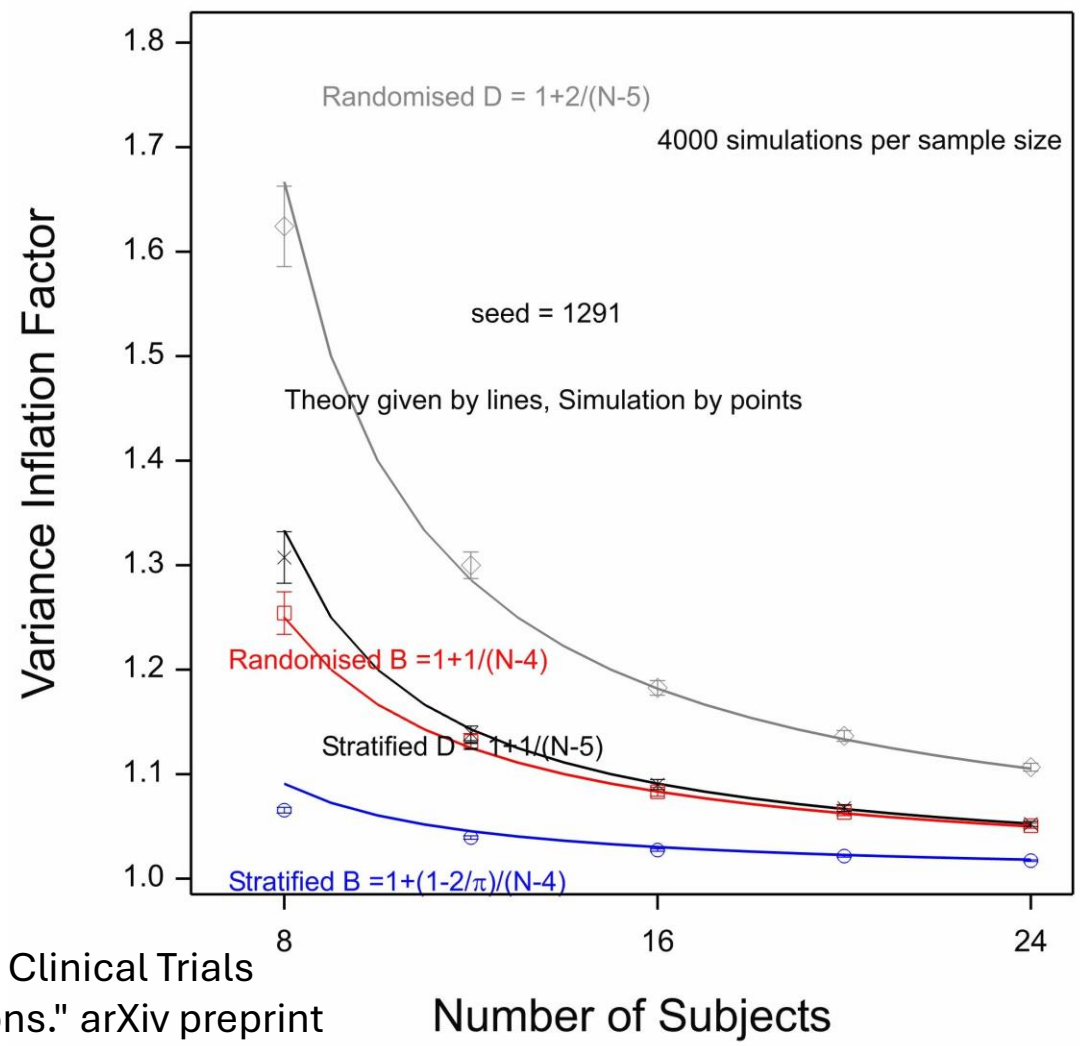
$$Model(D): Y \sim Z + X + S.$$

Qu, Y. (2011). "Issues for stratified randomization based on a factor derived from a continuous baseline variable." Pharm Stat 10(3): 232-235.
See also
Sullivan, T. R., T. P. Morris, B. C. Kahan, A. R. Cuthbert and L. N. Yelland (2024). "Categorisation of continuous covariates for stratified randomisation: How should we adjust?" Stat Med.

Variance inflation factors for stratified and randomised designs

		VIF -1	
Label	**Model**	**Randomised**	**Stratified**
A	Y ~ Z	0	0
B	Y ~ Z + X	$1/(N-4)$	$(1-2/\pi)/(N-4)$
C	Y ~ Z + S	$1/(N-4)$	0
D	Y ~ Z + X + S	$2/(N-5)$	$1/(N-5)$

Uses fact that we are now sampling from a truncated Normal.



Senn, S. J., F. König and M. Posch (2024). "Stratification in Randomised Clinical Trials and Analysis of Covariance: Some Simple Theory and Recommendations." arXiv preprint arXiv:2408.06760.

Example 3. Is the propensity score a good way to adjust for covariates?

- There is a whole heap of simulations investigating the performance of the propensity score compared to analysis of covariance.
- The messages are contradictory and confusing.
- However, the following is the case:
    1. By concentrating on that which is predictive of assignment, PS will maximise the variance inflation factor, λ.
    2. By concentrating on that which is predictive of outcome, ANCOVA will minimise the MSE.
- Obviously, there is a problem if the ANCOVA model is not pre-specified but clearly PS is inadmissible.
- So, the answer is , “no”.

Senn, S. J., E. Graf and A. Caputo (2007). "Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment on exposure." Statistics in Medicine **26(30): 5529-5544.**

Example 4. Should you categorise a covariate, for example age?

- Consider how many categories you might want to create.
- If you have k+1 categories you will have k dummy variables and k is the value that can be used to calculate the expected value of the VIF, λ.
- You can now choose a set of k smooth functions (for example fractional polynomials) to fit a set for which:
  - The expected effect on λ will be the same
  - The effect on second order precision will be the same.
- The only issue that remains is, “can you do better in terms of expected mean square error?”
- Most reasonable modellers will be able to answer ‘yes’.
- So, the answer to the question is ‘no’.

Example 5. Should you replace a set of covariates by a prognostic score based on historical data?

- Using a single prognostic score sets the value of k to 1.
- If there are $k > 1$ covariates then clearly by using them:
  - The expected value of the VIF, λ will be larger
  - Second order precision will be reduced.
- However, the relationship between outcome and the covariates may well vary somewhat from trial to trial.
- Therefore, the prognostic score will be imperfectly constructed.
  - The expected reduction of the mean square error will be greater by using the covariates independently in the model.
- So, the answer is: “it depends on the sample size (& the number of covariates)”. See Example 1.

Siegfried, S., S. Senn and T. Hothorn (2023). "On the relevance of prognostic information for clinical trials: A theoretical quantification." Biom J **65(1).**

Example 6. What is the value of minimisation etc?

- Assuming that the covariates that are used to allocate treatment using minimisation or some superior approach such as Atkinson's algorithm will be modelled whatever the design, there is no advantage as regards:
  - Expected mean square error
  - Second order precision.
- The only benefit is in terms of the variance inflation factor.
- In practice this will be lower than the value for a randomised design but somewhat larger than 1.

Senn, S. J., V. V. Anisimov and V. V. Fedorov (2010). "Comparisons of minimization and Atkinson's algorithm." Statistics in Medicine 29(7-8): 721-730.

Conclusion

The *N* is high and the end is nigh

Conclusions

- We are not saying that there is no room for simulation.
- We are saying that simulation will be more valuable if it is guided by theory.
  - The theory provides a check and *vice versa*.
  - The theory provides a thread that can be used to link different simulations.
  - The theory helps us isolate what particular feature is relevant.
  - The theory frees the simulations to go beyond the standard case the theory covers.
- We think that it useful to look at the three effects on precision of inferences.
- We also suggest that statistics such as power, being derived from variances, are less useful to study than variances themselves.
- To do this, the variance inflation factor is key.

Finally

"The fact is that, useful though simulation is, an analytical solution is always preferable …The relative power of an analytic solution, as compared with a simulation approach to a model, is such that even if a full analytical solution is impossible, such a solution to part of the model, with the remainder investigated by simulation, is preferable by far to a simulation solution to the whole model."

Morgan, B. J. T. (1984). Elements of Simulation, Chapman and Hall.

P7