Adjusting stepwise p-values in generalized linear models

Chiara Brombin[1] Livio Finos[2] Luigi Salmaso[3]

[1]Department of Statistics
University of Padova

[2]C.I.B.B.
University of Roma "Tor Vergata"

[3]Department of Management and Engineering
University of Padova

Vienna, 8-11 July, 2007

Summary

- Brief introduction (model selection bias and model selection uncertainty)
- A simple way to correct p-values in stepwise procedures
  - Algorithm
  - Theoretical properties
  - Simulation studies (under the null and alternative hypotheses)
- Conclusions

- Several statistical computer packages contain stepwise subroutines for selecting the best subset of regressors and provide a p-value based on the F-to-enter and computed from tables of the F-distribution.
- This distribution is correct only if all previously entered regressors have not been data-steered (see Grechanovsky and Pinsker (1995), Austin and Tu (2004)).
- Resampling methods or bootstrap technique may be a solution (see Harshman and Lundy (2006) and Freedman et al. (1992)).

Type I error is out of control . . .

What's the probability to find a non-real significant model after stepwise regression?

- We generate standard normal distributed independent covariates that are unrelated and independent to the outcome.
- First simulation study: m=10 covariates, n=20 cases, MCM=1000.
- Second simulation study: m=20 covariates, n=30 cases, MCM=1000.

nominal α level	0.01	0.05	0.10	0.20	0.30	0.50
1° simulation	0.187	0.521	0.733	0.892	0.935	0.939
2° simulation	0.530	0.840	0.938	0.983	0.996	0.998

For details see the algorithm described afterwards.

General idea

- The multiplicity arises because of the multitude of models explored by the stepwise method. With m covariates, you can obtain $M = 2^m - 1$ models.
- The p-values tends to be (very) "small" also when $\mathbf{Y}$ is NOT associated with any of the $\mathbf{X}$.
- More covariates and more exaustive research you do, smaller p-values you get ...
- A Bonferroni correction $p \times M$ is valid but the conservativeness of this solution is often unacceptable for both theoretical and practical purposes. In fact
  - the selected p-value is not always the minimum of all possible models, because the research of stepwise methods is not exaustive,
  - M can be very high,
  - the p-values of different models are dependent when part of variables are in common.

Algorithm

This simple algorithm, while controlling the α-level, ensures the unbiasedness and the consistency of the estimated p-values of the selected model (Finos and Salmaso, 2006).

1. Perform a standard stepwise regression (backward or forward) in a lm or glm (e.g. logistic, Poisson, Cox models).
2. Extract the p-value associated to F statistic (test on residual deviance for glm). This p-value is called the observed p-value.
3. Consider a permutation of the response variable y and repeat steps 1) and 2).
4. Carry out B (e.g. 1000 or 5000) independent repetitions of the previous step.
5. The corrected p-value is exactly the fraction of permutation p-values that are less or equal to the observed one.

Test's properties (1)

Define

- $t(\mathbf{Y}, \mathbf{X}) = t(\mathbf{Y})$, the p-value of the model selected by the stepwise procedure for $\mathbf{Y} \sim g(\mathbf{X})$ (t is a test statistic),
- $\mathcal{Y}_{/\mathbf{Y}}$, the permutation sample space or the orbit of $\mathbf{Y}$ that contains $n!$ elements.

The test is invariant with respect to its measure and α-size because (Pesarin F., 2001)

$$f(t(\mathbf{Y}')) = f(t(\mathbf{Y}'')),\ \forall \{\mathbf{Y}', \mathbf{Y}''\} \subset \mathcal{Y}_{/\mathbf{Y}}.$$

i.e. The process has the same distribution for any random permutation of observed $\mathbf{Y}$*.*

Remark: Model selection can just estimate which model is best, based on the single data set; the observed data are conceptualized as random variables and their values would be different if another independent sample were available (see Burnham (2002)).

Test's properties (2)

The test is unbiased because

$$F(t(\mathbf{Y})) > F(t(\mathbf{Y}')),\ \mathbf{Y}' \in \mathcal{Y}_{/\mathbf{Y}}$$

when the vector of observed $\mathbf{Y}$ depends on $\mathbf{X}$.

i.e. Under H_1 the distribution of p-value of selected model is stochastically larger for $\mathbf{Y}$ than for any random permutation.

If the step-wise procedure is consistent (e.g. forward selection considers all the univariate model and backward considers the full model, both are consistent), the test is consistent because

$$t(\mathbf{Y}) \to 0 \quad \text{as} \quad n \to \infty.$$

i.e. The distribution of p-value of selected model tends to be 0 constant, whereas the distribution of p-value of selected model for any pemutation of $\mathbf{Y}$ does not.

Outlines from literature

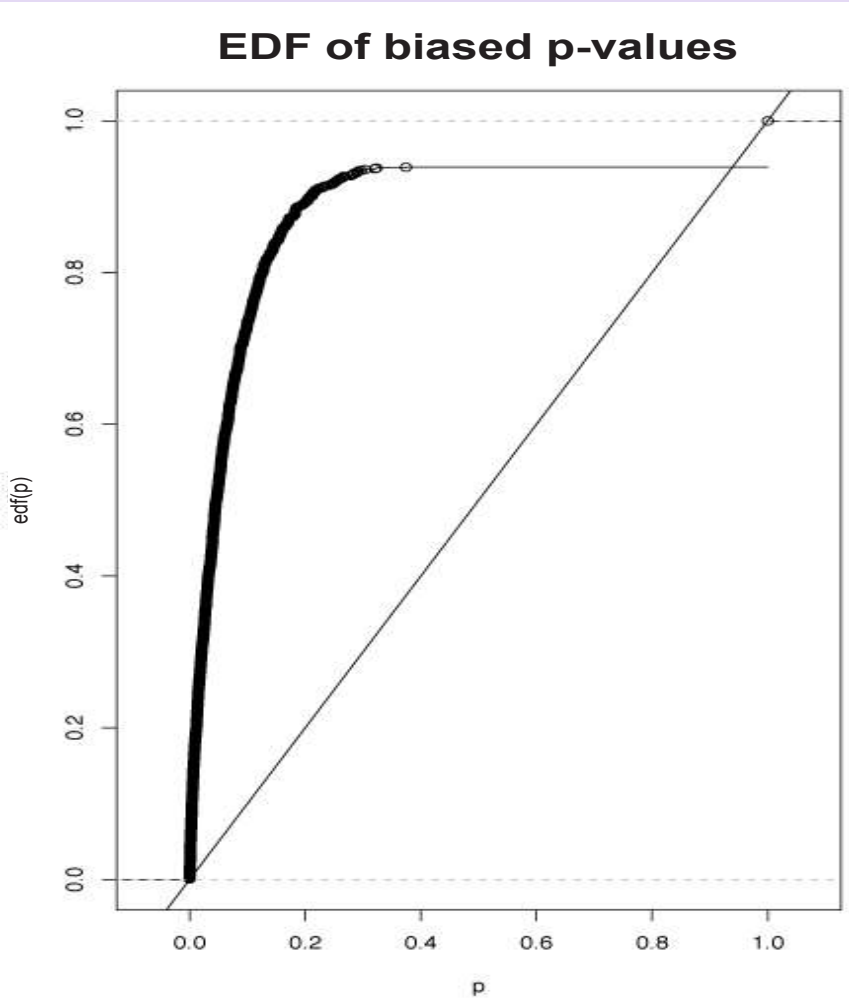
- Copas and Long (1991) propose a correction for forward selection in multiple linear model for orthogonal regressors.
- Grechanovsky and Pinsker (1995) generalize it to general forward selection for linear models.
- Harshman and Lundy (2006) make use of empirical approximate approach for this correction.

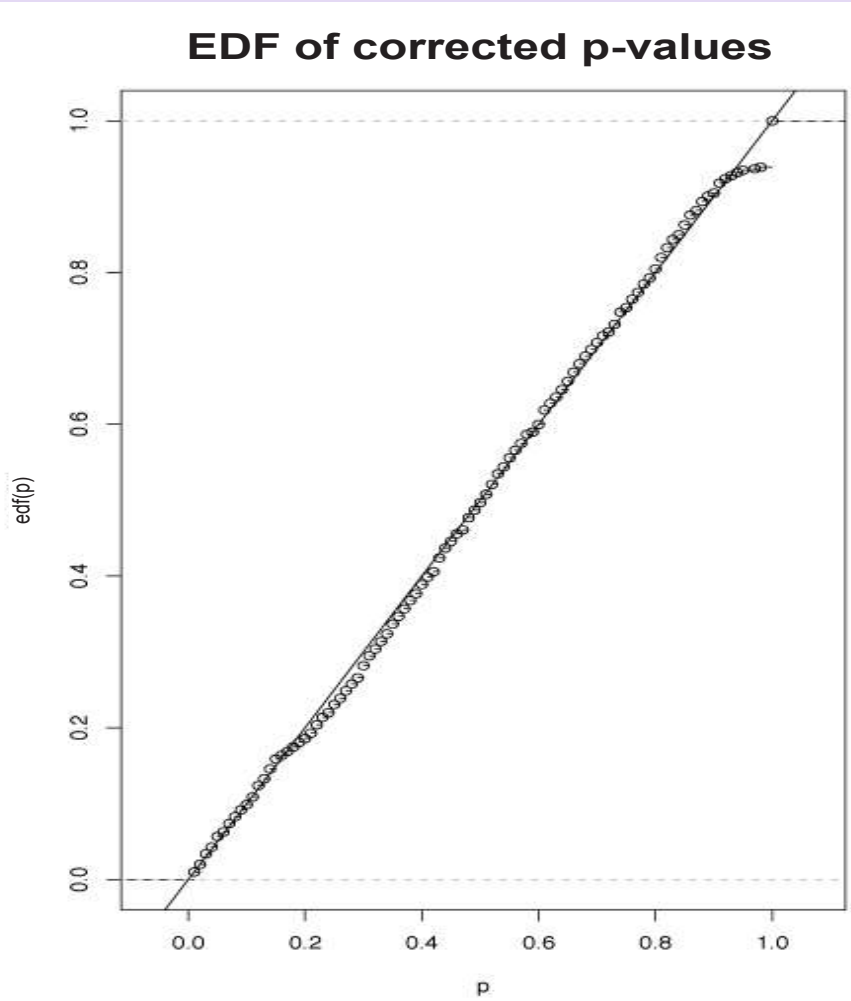
All those works control the Familywise Error Rate (FWE) in a strong sense but are restricted to linear model under forward selection.

The proposed method control the FWE in a weak sense but is valid for any GLM and any stepwise selection method.

Results in the first simulation study

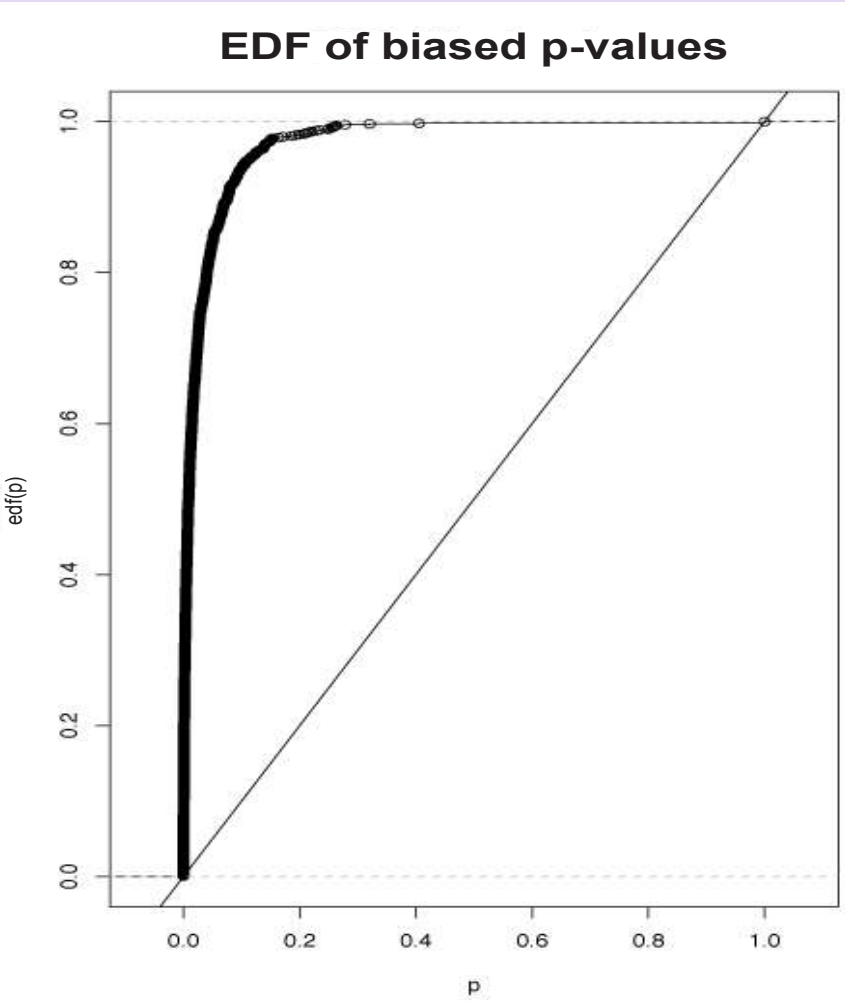
First simulation study: m=10 covariates, n=20 cases, B=100, MCM=1000.

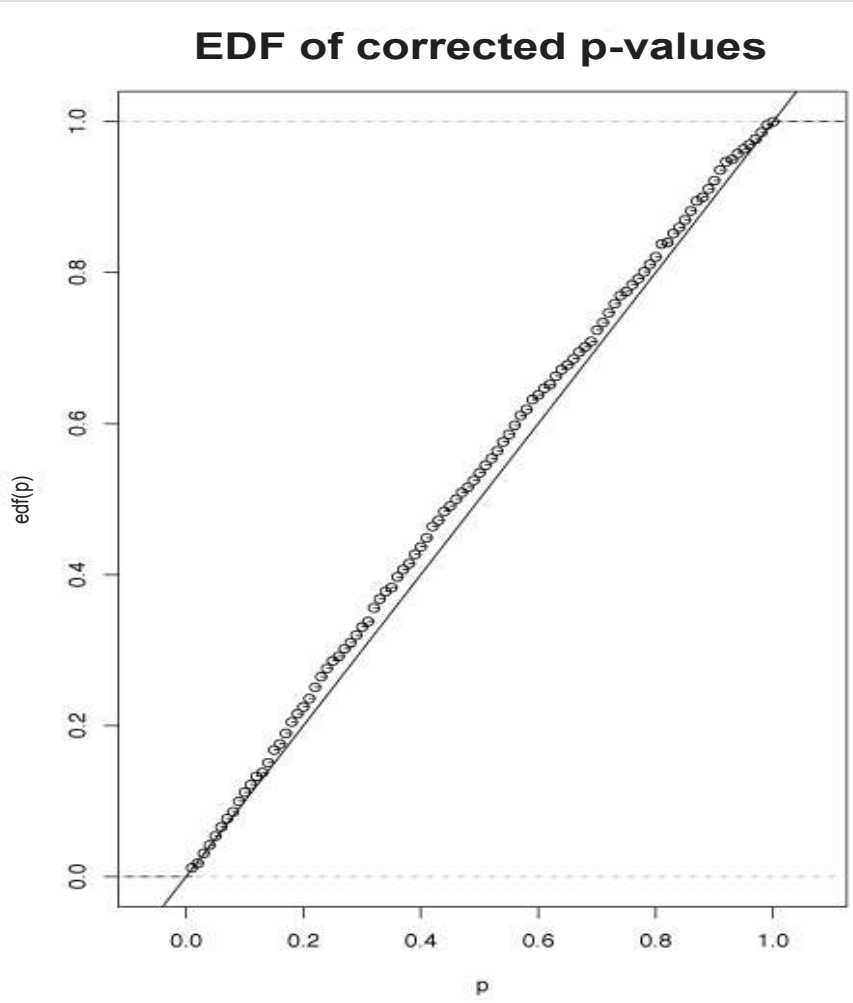




Results in the second simulation study

Second simulation study: m=20 covariates, n=30 cases, B=100, MCM=1000.

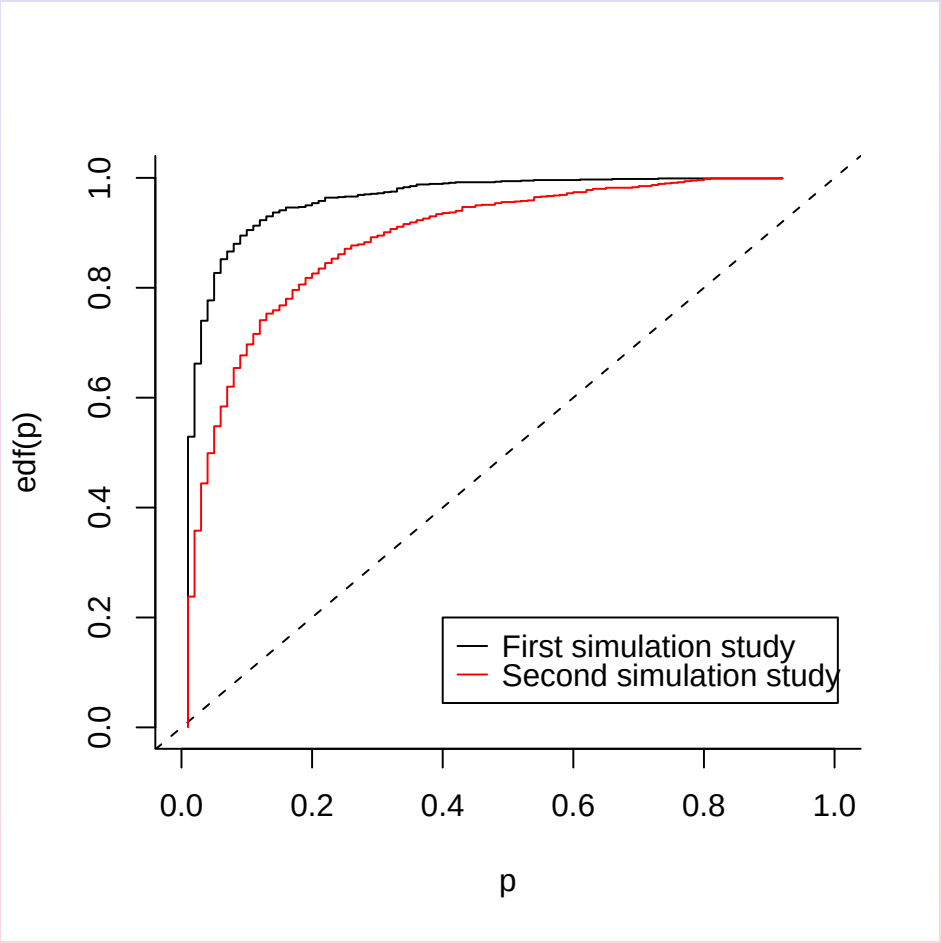




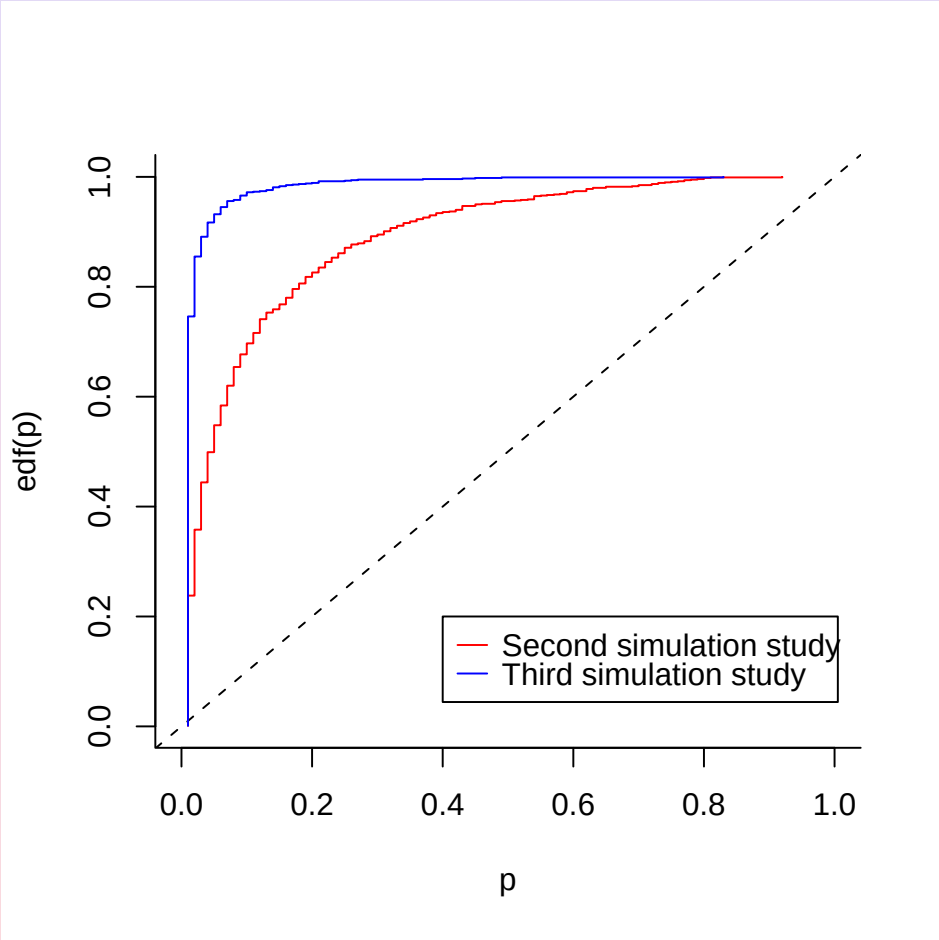
About power ...

- In the first simulation study, we have generated 5 dependent and related to the outcome covariates out of the 10 normally distributed covariates (5 out of the 20 in the second simulation study). The remaining standard normal distributed covariates are generated independent to each other and unrelated and independent to the outcome.
- First simulation study: m=10 covariates, n=30 cases, $\rho = 0.4$, B=100, MCM=1000.
- Second simulation study: m=20 covariates, n=30 cases, $\rho = 0.4$, B=100, MCM=1000.
- Third simulation study: m=20 covariates, n=30 cases, $\rho = 0.6$, B=100, MCM=1000.

Increasing the number of covariates



Increasing the value of the correlation coefficient ρ



The p-values of stepwise regression can be highly biased. In particular

- the evaluation of glm-stepwise must be prudent, mainly when regressors have been data-steered,
- it's possible to correct p-values in a very simple manner,
- our proposal is a nonparametric permutation solution that is exact, flexible and potentially adaptable to most different applications of model selection,
- the correction becomes more severe when many variables are processed by the stepwise machinery.

Open problems

Future developments concern the generalization to

- stepwise canonical correlation / MANOVA,
- discriminant analysis,
- segmentation tree models and general selection and inference methods,
- strong control of FWE *(to be continued . . .)*

R code

```
sim_adjSW<-function(n,m,B,MCMC,seme){
set.seed(seme)
raw_p<-rep(1,MCMC)
adj_p<-rep(1,MCMC)
ps<-as.data.frame(cbind(raw_p,adj_p))
xnam <- paste("X$V", 1:m, sep="")
fmla <- as.formula(paste("y ~ ", paste(xnam, collapse= "+")))
fmlaperm <- as.formula(paste("sample(y) ~ ", paste(xnam, collapse= "+")))
for(j in 1:MCMC){
p<-rep(1,B)
y<-rnorm(n, mean=0, sd=1)
X<-(as.data.frame(matrix(rnorm(n*m, mean=0, sd=1),n,m)))
mod1<-lm(fmla)
result<-summary(step(mod1,trace = 0))
if(is.null(result$fstatistic)==FALSE){if( result$fstatistic[2]>0 & result$fstatistic[3]>0)
p[1]=pf(result$fstatistic[1],result$fstatistic[2],result$fstatistic[3],lower.tail = FALSE)}
for(i in 2:B){
print(cat("."))
y<-sample(y)
mod1<-lm(fmla)
result<-summary(step(mod1,trace = 0))
if(is.null(result$fstatistic)==FALSE){if( result$fstatistic[2]>0 & result$fstatistic[3]>0)
p[i]=pf(result$fstatistic[1],result$fstatistic[2],result$fstatistic[3],lower.tail = FALSE)}
}
ps$adj_p[j]<-(sum(p<=p[1])/B)
ps$raw_p[j]<-p[1]
print(p)
}
return(ps)
}
PS<-sim_adjSW(20,10,100,1000,17)
```

Bibliography

Austin, P. C., Tu, J. V. (2004) Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of Clinical Epidemiology*, **57**, 11381146.

Box, G. E. P., Jenkins, G. M. (1970) *Time series analysis: forecasting and control*. London: Holden-Day.

Burnham, K. P., Anderson, D. R. (2002) *Model selection and multimodelinference: a practical information-theoretic approach*. New York: Springer-Verlag.

Copas, J. B., Long, T. (1991) Estimating the residual variance in orthogonal regression with variable selection. *The statistician*, **40**, 51-59.

Bibliography

Finos, L., Salmaso, L. (2005) A new nonparametric approach for multiplicity control: Optimal Subset procedures. *Computational Statistics*, **20**, 643-654.

Finos, L., Salmaso, L. (2006) Weighted methods controlling the multiplicity when the number of variables is much higher than the number of observations. *Journal of Nonparametric Statistics*, **18**, 245-261.

Freedman, L.S., Pee, D., Midthune, D.N. (1992) The problem of underestimating the residual error variance in forward stepwise regression. *The statistician*, **41**, 405-412.

Grechanovsky, E., Pinsker, I. (1995) Conditional p-values for the F-statistic in a forward selection procedure. *Computational Statistics & Data Analysis*, **20**, 239-263.

Bibliography

Harshman, R. A., Lundy, M. E. (2006) A randomization method of obtaining valid p-values for model changes selected "post hoc". see `http://publish.uwo.ca/ harshman/imps2006.pdf`

Pesarin, F. (2001). *Multivariate Permutation Test With Application To Biostatistics.* Wiley: New York.