November 6, 2015

Multimodel Inference:
Understanding AIC relative variable importance values

Kenneth P. Burnham
Colorado State University
Fort Collins, Colorado 80523

Abstract

The goal of this material is to present extended theory and interpretation for the variable importance weights in multimodel information theoretic (IT) inference. We show that these statistics are a valid measure of relative predictor variable importance under all subsets model selection. This became of concern to the author upon realizing that the validity and value of variable importance weights, given in Burnham and Anderson (2002), has been called into question recently. An aspect of the "validity" issue that seems to have been raised is that these importance weights did not relate to regression parameter estimates, which historically are the basis for inference about predictor importance given a single model. The material here answers these issues with theory that clarifies the relative variable importance values in question and shows they are valid.

Background

Assessment of the relative importance of variables has often been based only on the best model (e.g., often selected using a stepwise testing procedure). Variables in that best model are considered "important," while excluded variables are considered not important. More refined importance measures can be based on actual parameter estimates from the fitted model. However, if model selection is part of inference we should also want a measure of variable "importance" that is computed based on all the models in the candidate set. The solution in the literature involves Akaike weights (or model probabilites, if Bayesian) summed for all models containing predictor variable x_j, $j = 1, \ldots, R$; denote these sums as $w_+(j)$. The predictor variable with the largest predictor weight, $w_+(j)$, is thus estimated to be the most important of the predictors. The variable with the smallest sum is estimated to be the least important predictor, and possibly one of no importance at all.

These variable importance values are best considered as only relative importances, because the tie to actual predictive performace is indirect (but real, as shown below). The "relative" aspect of this variable importance inference is also totally tied to the set of models used, including the form of the models and the form in which variables are used and the sample size, n. Inferences could change if interaction terms are included, or the model set changes in other ways, or there were more data.

A context where these variable importance weights is sensible is all-subsets selection for generalized linear models (usually simple linear and for standard assumptions including residual variance normality and homogeneity). Then for any predictor variable x_j there are 2^{R-1} models that do not contain the variable and the same basic 2^{R-1} models agumented by adding predictor variable x_j. These two subsets are the same size and their union is the full set of all 2^R models being considered. [Note: to fully understand what is below the reader should know material in Burnham and Anderson (2002), especially Sections 4.2.2 and 6.9.8; it might suffice to have studied Burnham and Anderson (2004), or Anderson (2008)].

Some background notation and formulae:

$$\text{AIC} = -2\log(\mathcal{L}(\hat{\theta} \mid data)) + 2K$$

$$\Delta_i = \text{AIC}_i - \text{AIC}_{min}$$

$$\mathcal{L}(g_i \mid data) = \exp(-\Delta_i/2) = \text{ the likelihood of model } g_i \text{ .}$$

Note that only relative likelihoods have inferential meaning, hence Δ_i can be replaced by just AIC_i for theory, but often not when some actual calculations are to be made. Akaike weights:

$$w_i = \frac{\exp(-\Delta_i/2)}{\sum_{r=1}^{R} \exp(-\Delta_r/2)}. \qquad 1$$

New Theory

This w_i can also be viewed as the proportion of total evidence in support of model g_i as being the Kullback-Leibler (KL) best model. There is a sense in which these can be considered the probability that model g_i is the KL best model. It simplifies what is below to use this probability terminology; alternatively it can be viewed as "proportion of evidence." Lastly,

$$w_+(j) = \sum_{i \text{ for } x_j \in g_i} w_i$$

which is the sum of model weights for the subset of all models that contain predictor variable x_j. The sum of model weights for the subset of all models that do not contain predictor variable x_j is

$$w_-(j) = \sum_{i \text{ for } x_j \notin g_i} w_i \quad .$$

In the underlying model-likelihoods, β_j (plus the other parameters) has been estimated by maximum likelihood (ML) for each model g_i involved in $w_+(j)$. The estimate $\hat{\beta}_j$ varies by model; call it $\hat{\beta}_j(i)$ for parameter $\beta_j(i)$ in model g_i. The $\beta_j(i)$ do not appear in $w_-(j)$. So, to assess the "importance" of predictor x_j some form of contrast of $w_+(j)$ to $w_-(j)$ is needed. Importance here really means the probability, or weight of evidence, that predictor x_j is included in the KL best model. Because $w_+(j) + w_-(j) = 1$ it suffices to simply note the value of $w_+(j)$; the larger it is the more important is predictor x_j. Thus, relative importance among the predictors is quantified directly by $w_+(j)$ (provided there is a *balanced model set*) because this is the probability that x_j is "important," as regards prediction based on the set of fitted models. Thus, "important" is to be understood here in this sense.

In general, for $\beta_j z_j$, appearing in some or all of the models in the model set, $w_+(j)$ is the probability that the term $\beta_j z_j$ is in the KL best model. Here, z_j can be a predictor, such as x_j or $(x_j)^2$ or $x_j z_j$, for z_j a second recorded covariate. Considered fitting a polynomial in x; terms like $\beta_1 x$, $\beta_2 x^2$, and $\beta_3 x^3$ are in the models. Then, for exsample, $w_+(2)$ relates to the importance of term $\beta_2 x^2$ in the set of models, not the importance of predictor x, as x is clearly "important" given that it is the only covariate in the models. So, using the $w_+(j)$ as relative variable importance measures applies only as a special case and depends on the model set.

The KL best model (the target for AIC selection) includes just those predictors x_j where the associated β_j can be estimated with enough precision that including x_j improves fitted model predictive performance. A predictor is excluded if the associated partial regression parameter is poorly estimated, such that including the predictor does not improve fitted model predictive performance. This does not require $\beta_j = 0$; in fact, in IT model selection inference we assume $\beta_j \neq 0$. This same sum, but of model posterior probabilites (Bayesian fitted model weights analogous to $w_+(j)$), is used in Bayesian model selection, and has been used and considered in published papers (see, e.g., Hoeting et al. 1999, Viallevfont et al. 2001, Barbieri and Berger 2004, and Hooten and Hobbs 2015). However, we maintain that the sometimes used Bayesian interpretation that their $w_+(j) = \Pr\{\beta_j = 0\}$ should really be $w_+(j) = \Pr\{x_j \text{ is in the true model}\}$.

While $w_+(j)$ depends on $\hat{\beta}_j(i)$ for the subset of all models that include x_j, and is defacto compared to $w_-(j)$, because they sum to 1, an explicit representation of this validity will be interesting and helpful. The math is easier to do by working with the odds ratio (OR), which is the logical basis for evidentual inference in the likelihood paradigm; see e.g., Burnham and Anderson (2014):

$$\frac{w_+(j)}{w_-(j)} \equiv \frac{w_+(j)}{1-w_+(j)} = \text{OR} . \qquad 2$$

A simplified notation is used for the summation over subsets (wrt predictor x_j):

$$\text{OR} = \frac{\sum\limits_{i \text{ for } x_j \in g_i} w_i}{\sum\limits_{i \text{ for } x_j \notin g_i} w_i} \equiv \frac{\sum\limits_{in} w_i}{\sum\limits_{out} w_i} \ . \qquad 3$$

Let K_i be the parameter count for model i in the subset of the simpler models. Then for the same model with x_j added its parameter count is $K_i + 1$. Various things that are constants over the model set will drop out when we express OR in terms of likelihoods (note, only AIC is considered here; results for AIC_c are given in **Supplement A**):

$$\text{OR} = \frac{\sum\limits_{in} \mathcal{L}(\cdot \mid g_i)(\text{e}^{-K_i})(\text{e}^{-1})}{\sum\limits_{out} \mathcal{L}(\cdot \mid g_i)(\text{e}^{-K_i})} \ .$$

Simplfied notation is used for likelihoods: $\mathcal{L}(\cdot \mid g_i)$ denotes the usual parametric likelihood conditional on model g_i.

The numerator and denominator sums are over different models. Likelihoods have been maximized over all relevant parameters; $\hat{\beta}_j(i)$ only occur in numerator models. So further denote this as (again, with simplfied notation for the likelihood)

$$\text{OR} = \frac{\sum\limits_{in} \mathcal{L}(\hat{\beta}_j(i))(\text{e}^{-K_i})(\text{e}^{-1})}{\sum\limits_{out} \mathcal{L}(\beta_j(i)=0)(\text{e}^{-K_i})} \ . \qquad 4$$

Further Theory

A little algebraic re-expression gives

$$\text{OR} = \frac{\sum\limits_{in} \left(\frac{\mathcal{L}(\hat{\beta}_j(i))}{\mathcal{L}(\beta_j(i)=0)} \right) \left(\mathcal{L}(\beta_j(i)=0)(\text{e}^{-K_i}) \right) (\text{e}^{-1})}{\sum\limits_{out} \mathcal{L}(\beta_j(i)=0)(\text{e}^{-K_i})} \ .$$

The quantity

$$\frac{\mathcal{L}(\hat{\beta}_j(i))}{\mathcal{L}(\beta_j(i)=0)} = 1/\text{LR}$$

is the inverse of the standard likelihood ratio (LR) used to test $\beta_j(i) = 0$ as a null hypothesis model. This is another odds, i.e., evidence ratio. LR reflects only the information about the "importance" of $\beta_j(i)$, hence importance of x_j as a predictor in the given alternative model (i.e., given the set of other predictors in these two models). This ratio is ≥ 1; we can define $2\log(1/\mathrm{LR}) = \chi_i^2$. So defined, and considered as a random variable, this statistic is approximately a chi-square random variable. This fact is not needed to get a re-expression for OR (it is useful later):

$$\mathrm{OR} = (\mathrm{e}^{-1}) \frac{\sum_{in} \left(\mathrm{e}^{\frac{1}{2}\chi_i^2} \right) \left(\mathcal{L}(\beta_j(i)=0)(\mathrm{e}^{-K_i}) \right)}{\sum_{out} \mathcal{L}(\beta_j(i)=0)(\mathrm{e}^{-K_i})} .$$

Now $\mathrm{e}^{\frac{1}{2}\chi_i^2}$ is from a pair of models, so keeping it unchanged we can just as well do the numerator sum over the subset of "out:"

$$\mathrm{OR} = (\mathrm{e}^{-1}) \frac{\sum_{out} \left(\mathrm{e}^{\frac{1}{2}\chi_i^2} \right) \left(\mathcal{L}(\beta_j(i)=0)(\mathrm{e}^{-K_i}) \right)}{\sum_{out} \mathcal{L}(\beta_j(i)=0)(\mathrm{e}^{-K_i})} .$$

Define

$$w_{i,out} = \frac{\mathcal{L}(\beta_j(i)=0)(\mathrm{e}^{-K_i})}{\sum_{out} \mathcal{L}(\beta_j(i)=0)(\mathrm{e}^{-K_i})} .$$

For the subset of models that exclude predictor x_j these are model weights; they sum to 1. Hence, write OR as

$$\mathrm{OR} = (\mathrm{e}^{-1}) \left(\sum_{out} \left(\mathrm{e}^{\frac{1}{2}\chi_i^2} \right) w_{i,out} \right). \qquad 5$$

.

All the information about the importance of x_j is in the set of χ_i^2 as each one is a function of $\hat{\beta}_j(i)$. Thus, $w_+(j)$ functionally is only dependent on the importance of the $\hat{\beta}_j(i)$ (which are from the full subset of models that include x_j). Using the notation χ^2 does not make that statistic a chi-square random variable. So far, it is just a transformation on the likelihood ratios. However, it's probability distribution is closely approximated as a 1 df chi-square. Moreover, it is numerically well approximated by the Wald statistic, which is

$$[\hat{\beta}_j(i) / \hat{\mathrm{se}}(\hat{\beta}_j(i))]^2.$$

Hence, we could use the (very good at large n) approximation

$$e^{\frac{1}{2}\chi_i^2} \doteq \exp\left(\frac{1}{2}\left\{\frac{\hat{\beta}_j(i)}{\hat{\text{se}}(\hat{\beta}_j(i))}\right\}^2\right)$$

to emphasize that the odds ratio (OR), and $w_+(j)$, depend on estimated partial regression parameters and their estimated precision.

The back transform from OR to $w_+(j)$, from equation 2 is

$$w_+(j) = \frac{\text{OR}}{1+\text{OR}} \equiv \left(1 + \frac{1}{\text{OR}}\right)^{-1}; \qquad 6$$

it is monotonic.

Discussion

We note that these formulae for OR (under AIC) do not depend on sample size. Thus bounds and other special values given below for $w_+(j)$ apply at any sample size (even if AIC is used when AIC_c should be used). The minimum possible value for OR occurs when all $\chi_i^2 = 0$ (all original LR = 1); the result is $OR=e^{-1}$ and $w_+(j) = 1/(1+e) = 0.2689$.

Using the chi-square assumption a more useful bound may be when we set each χ_i^2 to 1. This corresponds to the expected chi-square under the case when all the $\beta_j(i)$=0. The result:

$$\text{OR} = (e^{-1})(e^{\frac{1}{2}}) = e^{-\frac{1}{2}}, \quad \text{and} \quad w_+(j) = \frac{1}{1+\sqrt{e}} = 0.3775\,.$$

Rounding a bit, this suggests on average we may need $w_+(j) \geq 0.4$ to begin to consider x_j as possibly important. Even if, for example, $w_+(j) = 0.4$ were the case, AIC still might favor having x_j in a particular model if for that model the specific evidence ratio, 1/LR, is large.

The evidence ratio OR equals 1 if all $\chi_i^2 = 2$, which provides the point at which AIC gives a tie between two models different by 1 parameter when the simpler model is nested in the more general model. In this case $w_+(j) = 0.5$. Barbieri and Berger (2004) give special note of this "median" model which is the model based on just the predictors that have $w_+(j) \geq 0.5$.

It is important to realize that these results about $w_+(j)$ reflect their being a weighted average over the full set of all possible models; hence, they do not apply to any specific model. Rather, variable importance values represent an average property of predictors for the model selection context wherein they are computed. That context is the specific set of models including their form and the covariates used. Because the set of $w_+(j)$, $j = 1, \ldots, R$, do not depend on any specific model of the model set they are suitable as a basis to measure the relative predictive importance of these R predictors in the specific context of application with a balanced model

set. This interpretation should also be evident because $w_+(j)$ is the probability that predictor x_j is in the KL best model (or for a Bayesian, it is the probability that x_j is in the true model).

There is a lack of sensitivity to rank the predictors when any two or more $w_+(j)$ values equal 1, or are very near to 1. Moreover, the better way to assess estimated predictor importance is to use the set of statistics $t_j = |\hat{\beta}_j / \hat{\text{se}}(\hat{\beta}_j)|$, $j = 1, \ldots, R$, either directly or transformed to evidence ratios (Burnham and Anderson 2014). However, this requires selecting a particular model, or getting the full model averaged results. Either of those approaches require first having the model weights, hence in effect, first having the $w_+(j)$ values. Supplements B and C give numerical exmples of these ideas.

The theory presented here about variable importance weights, in multimodel information theoretic inference, provides the mathematical proof that these statistics are a valid measure of predictor variable importance under all subsets model selection.

Acknowledgements

I thank David R. Anderson for helpful inputs and reviews of this work, and Gary C. White for his comments on some specific ideas. Marc J. Mazerolle kindly ran his R program *AICcmodavg* on the fat data to compute for me the unconditional standard errors of the model averaged $\hat{\beta}_{ma}$ (AICcmodavg: model selection and multimodel inference based on (Q) AIC(c), MJ Mazerolle - R package version, 2011). This work was in response to comments on web sites stating that these variable importance values were not valid and could not possibly be correct or useful. I thank Joseph P. Ceradini for being instrumental in bringing such web-site comments to my attention, and providing proof reading comments.

References

Anderson, D. R. 2008. Model based inference in the life sciences: A primer on evidence. Springer, New York, NY.

Barbieri, M. M., and J. O. Berger. 2004. Optimal prediction model selection. The Annals of Statististics 32:870-897.

Burnham, Kenneth P., and David R. Anderson. 2002. *Model Selection and Multimodel Inference*: *A Practical Information-Theoretical Approach* (2nd ed.). Springer, New York, NY.

Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference: understanding AIC and BIC in model selection. Sociological Methods and Research 33:261-305.

Burnham, K. P., and D. R. Anderson. 2014. P-values are only an index to evidence: 20th- vs. 21st-century statistical science. Ecology 95:627-630.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Bayesian model averaging: A tutorial. Statistical Science 14:382-417.

Hooten, M. B., and N. T. Hobbs. 2015. A guide to Bayesian model selection for ecologists. Ecological Monographs 85:3-28

Murray, K, and M. M. Conner. 2009, Methods to quantify variable importance: implications for analysis of noisy ecological data. Ecology 90:348-355.

Viallevfont, V, A. E. Raftery, and S. Richardson. 2001. Variable selection and Bayesian model averaging in case-control studies. Statistics in Medicine 20:3215-3230.

Supplement A

Results above do not apply to AIC_c wherein the formula is less simple and the weights are $w_{ci,out}$ and are based on AIC_c, but otherwise computed in the same way as the $w_{i,out}$ are computed. The generalized formula is

$$OR_c = (e^{-1}) \left(\sum_{out} \left(e^{\frac{1}{2}\chi_i^2} \right) (w_{ci,out}) Z_i \right),$$

where

$$Z_i = \exp\left(K_i \left(\frac{n}{n-K_i-1} - \frac{n}{n-K_i-2} \right) \right) < 1.$$

In the limit as n gets large $Z_i = 1$, $w_{ci,out} = w_{i,out}$, and hence $OR_c = OR$. However, any sort of exact results are unclear given the complexity of the terms in OR_c. Bounds are possible because

$$\sum_{out} w_{ci,out} = 1 \quad \text{and} \quad Z_i < 1.$$

So if all the $\chi_i^2 =$ a constant q (≥ 0), then

$$OR_c(q) < (e^{-1})\, e^{\frac{1}{2}q} = OR(q).$$

Let $w_{c+}(j)$ be the relative variable importance when AIC_c is used, rather than AIC.
For $q = 0$ (LR = 1), we have minimum of $w_{c+}(j) < 0.268 = 1/(1+e) =$ minimum of $w_+(j)$.

The same methods apply easily to BIC via BIC model weights in the same sort of all-subsets model selection. The analogous OR and variable importance weights are structurally the same except for the one factor e^{-1}. It becomes $\exp(-\frac{1}{2}\ln(n)) = 1/\sqrt{n}$. Therefore the minimum possible variable importance value under BIC, *with uniform priors on the models*, is

$$\frac{1}{1+\sqrt{n}} \quad .$$

The issue of the prior over the model set is critical here; it is usually uniform by default.

As a "test" of theory here, we applied BIC to the fat data refered to in Supplement B. The minimum of the analogous BIC variable importance values occurs at x_5 and is 0.066373. The theoretical minimum at n = 252 is 0.05926, which is less than 0.066373.

So far, everything here is for the case of just one predictor, hence one parameter. We can consider the joint importance of *m* predictors. Hence, let a set of *m* parameters be considered as a single multivariate parameter. These *m* are either always in or out of the two subsets of

models. Then for AIC the formula for OR and $w_{c+}(j)$ is structurally the same except for the one factor e^{-1}. That one factor becomes e^{-m}, hence the lower bound on $w_{c+}(j)$ is

$$\frac{1}{1+e^{m}} \cdot$$

There is often an issue about "pretending" variables (see Anderson 2009, pp 65-66), i.e., truely unimportant variables. It is hard to rule out one such variable considered by itself ($m = 1$). But as a set of m such variables the weight of evidence can be very small. For example, at m of 2, 3, or 4, respectively, the minimum possible importance probabilities for the set of variables are 0.12, 0.047, and 0.018).

Supplement B

This is a numerical example using the fat data (Burnham and Anderson 2002, pp 268-273); AIC_c was used. Model averaged estimates are denoted by *ma*; also, $t_j = |\hat{\beta}_j / \widehat{se}(\hat{\beta}_j)|$, and $t_{ma} = |\hat{\beta}_{ma} / \widehat{se}(\hat{\beta}_{ma})|$:

	from fitted full model			from multimodel inference results			
xj variable	$\hat{\beta}_j$	$\widehat{se}(\hat{\beta}_j)$	t_j	$w_{c+}(j)$	$\hat{\beta}_{ma}$	$\widehat{se}(\hat{\beta}_{ma})$	t_{ma}
x1 age	.000109	.000067	1.63	0.495	.0000466	.0000673	0.69
x2 weight	-.000215	.000128	1.68	0.933	-.0002714	.0001197	2.27
x3 height	-.000163	.000370	0.44	0.314	-.0000716	.0002404	0.30
x4 neck	-.000971	.000488	1.99	0.652	-.0005817	.0005848	0.99
x5 chest	-.000106	.000214	0.50	0.283	-.0000258	.0001216	0.21
x6 abdomen	.002036	.000187	10.89	1.000	.0020362	.0001717	11.86
x7 hips	-.000432	.000300	1.44	0.445	-.0001724	.0002891	0.60
x8 thigh	.000525	.000303	1.73	0.588	.0002723	.0003189	0.85
x9 knee	.000024	.000513	0.05	0.293	.0000746	.0003015	0.25
x10 ankle	.000571	.000461	1.24	0.448	.0002666	.0004282	0.62
x11 biceps	.000492	.000357	1.38	0.600	.0003684	.0004138	0.89
x12 forearm	.000923	.000413	2.23	0.828	.0007818	.0005196	1.50
x13 wrist	-.003649	.001100	3.32	0.976	-.0032857	.0012253	2.68

The variables t, $w_{c+}(j)$, and t_{ma} were replaced by their ranks in order to better meet some assumptions of the standard Pearson correlation coefficient, which was then used on the rank data. Those ranks are given below (1 is least important, 13 most important):

variable	t	$w_{c+}(j)$	t_{ma}
age	7	6	6
weight	8	11	11
height	2	3	3
neck	10	9	9
chest	3	1	1
abdomen	13	13	13
hips	6	4	4
thigh	9	7	7
knee	1	2	2
ankle	4	5	5
biceps	5	8	8
forearm	11	10	10
wrist	12	12	12

The rank order correlation of $t = |\hat{\beta} / \widehat{se}(\hat{\beta})|$ with $w_{c+}(j)$ is $r = 0.9011$ ($P < 0.0001$; we do not have an easy way to get the evidence ratio here for the null hypthesis $\rho = 0$). To check the parametric result, we simulated independent sets of "x, y" ranks ($n = 13$). With several cases of 1,000 repititions *r* never exceeded 0.9. So we did 10,000 independent samples and exactly 1 of the 10,000 *r*'s was >0.9011; the other 9,999 *r* values were < 0.9011. Thus, estimated *P* is 1/10,000 = 0.0001. Clearly, for this example there is a strong positive correlation of these two

measures of variable importance of these 13 predictors. This corroborates that $w_{c+}(j)$ measures importance, hence also relative importance, of predictor variables.

The rank order correlation of $w_{c+}(j)$ and t_{ma} is $r = 1.0$. This is clearly statistically significant. A simple permutation evaluation is P = 1/13!; this is roughly 10^{-10}. This perfect rank order correlation result was unexpected; we doubt it would always occur.

It is preferable to relate $w_{c+}(j)$ to $\left|\hat{\beta}_{ma}/\hat{\text{se}}(\hat{\beta}_{ma})\right|$. Whereas $\hat{\beta}_{ma}$ has been obtained easily from SAS. The SAS feature needed (in PROC REG) to allow computing $\hat{\text{se}}(\hat{\beta}_{ma})$ did not work. The R-package AICcmodavg did work.

The correlation of $\hat{\beta}$ from the full model with $\hat{\beta}_{ma}$ is r = 0.991. Regressing $\hat{\beta}_{ma}$ on $\hat{\beta}$ gives

$$\hat{\beta}_{ma} = 0.00001314 + 0.8863\hat{\beta} \qquad \text{(se on slope is 0.0365);}$$

with no intercept,

$$\hat{\beta}_{ma} = 0.8857\hat{\beta} \qquad \text{(se on slope is 0.0351).}$$

Here, the model averged estimates are like shrinkage estimates from the full model - about 11.4% shrinkage. In this body-fat example one could use the full model for prediction without much concern for model selection. But that is because there are only 13 predictors with a sample size of 252. Even so, model averaging performs better (for prediction) than the full model and better than the AICc selected best model (see Burnham and Anderson 2004). Moreover, one purpose of model selection is still of interest in this example: identify a smaller, hence perhaps less expensive, subset of predictors to use in a predictive model with operational (i.e., new) data. This approach is useful without using a null hypothesis testing approach with its issues of alpha levels and multiple testing concerns. With the IT approach, including variable importance weights, we get evidence about predictor variable relative and absolute importance (e.g., we could choose to say x_j is important when $w_+(j) \geq 0.5$).

Supplement C

Simulated (i.e., not real) data about predicting Freshman-year college gpa is given in Graybill and Iyer (1994, Regression analysis: concepts and applications, Duxbury Press). The sample size of $n = 20$ is too small for a realistic example, but is conveniently small for an illustrative example. Therefore, Burnham and Anderson (2002, pp 225-238) used this as an example. The four nominal predictors are math and verbal (i.e., english) SAT results and high school gpa in these subject areas (denoted x_1, x_2, x_3 and x_4 in Burnham and Anderson 2002). The relative variable importance weights are given in Burnham and Anderson (2002, p 227), but not the full model averaged regression parameters and their unconditional standard errors. We give that information below along with the metric $t = \left|\hat{\beta}/\widehat{\text{se}}(\hat{\beta})\right|$ which is a more sensitive measure of relative variable importance and is an actual estimated effect size. The model set is "all subsets:" 16 different models. The simple full model has all four predictors in it. The model averaged $\hat{\beta}_{ma}$ also define a "full" model, i.e. all four predictors are being used.

	Multmodel results				Full model		
x	w_+	$\hat{\beta}_{ma}$	$\widehat{\text{se}}(\hat{\beta}_{ma})$	t_{ma}	$\hat{\beta}$	$\widehat{\text{se}}(\hat{\beta})$	t
satm	0.997	0.00236	0.000558	4.23	0.00201	0.000584	3.44
satv	0.834	0.00117	0.000729	1.61	0.00125	0.000552	2.27
hsm	0.654	0.12620	0.119120	1.06	0.18944	0.091870	2.06
hse	0.147	0.01317	0.083860	0.16	0.08756	0.176500	0.63
intercept		0.28971			0.16155		

Note: $n = 20$, so AIC_c was used, hence theoretical minimum w_+ is unknown. We re-ran it with AIC, then hse ($= x_4$) has $w_+ = 0.304$.

The key point is that here the relative variable importance ranking given by w_+ and by t_{ma} (and for that matter, by t from the full model) are identical. If we compute the Pearson correlation on the ranks we get $r = 1$. A second point: the model averaged inportance of predictor variable x_1 (= *satm*), based on t, is increased compared to the result from the simple saturated model, while importance of *satv*, *hsm*, and *hse* decrease.