Model Selection Criteria Alternatives to R^2-adjusted

Introduction: The coefficient of determination, $\boldsymbol{R}^2$, has a simple interpretation as the proportion of the observed variation in the response variable explained by the variation of the explanatory variable in the data. $\boldsymbol{R}^2$-adjusted is more appropriate when comparing models with different numbers of parameters because it penalizes models for including variables that are not helping to fit the data to the model. However, in using $\boldsymbol{R}^2$-adjusted, we have already traded ease of interpretation in favor of a better statistic for model selection. Inevitably, people have come up with other statistics useful in model selection. I discuss two such below.

AIC - The Akaike Information Criterion, or AIC, involves the maximum likelihood estimate, or MLE. (If you are not familiar with the MLE, you can skip to the conclusion of this section.) The AIC is defined as $\text{AIC} = 2p - 2\ln(L)$, where *p* is the number of estimated parameters (the model is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$ so $p = k+1$) and *L* is the likelihood function under the maximum likelihood estimate of the model. We'll restrict ourselves to the multiple linear regression model considered in class (iid normally distributed errors). Below is a derivation of the log likelihood term in the AIC for this case.

- The likelihood function is $L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \mathbf{e}^{-\frac{1}{2}\frac{(y_i-\hat{\mu}_i)^2}{\hat{\sigma}^2}} = \left(2\pi\hat{\sigma}^2\right)^{-\frac{n}{2}} \mathbf{e}^{-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2}$, where evaluating the $\hat{\mu}_i$ involves using maximum likelihood to estimate the *betas* in $\boldsymbol{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \boldsymbol{\varepsilon}$, and $\hat{\sigma}^2$ is the maximum likelihood estimate of the error variance.
- First, to *maximize* the likelihood *L* we must *minimize* the sum in the exponent, $\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2$, which is the goal of least squares, so the maximum likelihood *betas* are just the least squares *betas* derived in class.
- To estimate the error variance in the model, it is sufficient to maximize the log likelihood,
  $\ln L = -\frac{n}{2}\ln\left(2\pi\hat{\sigma}^2\right) - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}\left(y_i-\hat{\mu}_i\right)^2 = -\frac{n}{2}\ln\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}\left(y_i-\hat{\mu}_i\right)^2 + \text{constant stuff}$
- Differentiating with respect to the variance estimator $\hat{\sigma}^2$, $\frac{\partial}{\partial\hat{\sigma}^2}\ln L = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}\left(y_i-\hat{\mu}_i\right)^2$
- Setting the derivative to zero, $0 = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}\left(y_i-\hat{\mu}_i\right)^2$, or $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{\mu}_i\right)^2$. Note: The MLE of the error variance is *not* the mean square error calculated in regression. Thus, the MLE is biased.
- Substituting $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{\mu}_i\right)^2$ into the exponent of the expression for *L*, $L = \left(2\pi\hat{\sigma}^2\right)^{-\frac{n}{2}} \mathbf{e}^{-\frac{n}{2}}$. Then
  $\ln L = -\frac{n}{2}\ln\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{\mu}_i\right)^2\right] + \text{constant stuff} = -\frac{n}{2}\ln\left[\sum_{i=1}^{n}\left(y_i-\hat{\mu}_i\right)^2\right] + \left[\frac{n}{2}\ln n + \text{constant stuff}\right]$.
- AIC is only a measure of the *relative* fit among the models being compared, so terms which are constant for all models are typically ignored. Then $\text{AIC} = 2p - 2\ln(L)$ simplifies to $\text{AIC} = 2p + n\ln(SSE)$, where *SSE* is the error sum of squares in regression (because the $\hat{\mu}_i$ are the least squares estimates $\hat{Y}_i$).

The Akaike Information Criterion, as its name suggests, is based on information theory. Models with smaller AIC values are judged better. Statgraphics uses $\text{AIC} = 2\frac{p}{n} + 2\,ln\left(RMSE\right)$, where *RMSE* is the *root* mean square error (i.e., the standard error of the estimate ***s***), which is my AIC divided by the sample size *n*. This has the advantage of stressing that the penalty a model pays (the first term) is twice the ratio of the number of coefficients estimated *p* to the number of observations *n*, thus disadvantaging models that overfit the data.

Statgraphics: Follow *Relate* > *Multiple Factors* > *Regression Model Selection* and enter the variables as you would in a multiple regression analysis. Your first option is to enter the maximum number of independent variables you wish to have in your final model. Next, check the *Analysis Summary* and *Best Information Criteria* tables. Looking at the *Analysis Summary*, you'll notice that Statgraphics is using best subsets selection, i.e., fitting all models that have the number of independent variables you specified earlier. The **AIC** appears in the third column of the *Best Information Criteria* table. (Statgraphics uses its own version of the AIC, using the formula derived above divided by *n*.)

Example: Using the 93cars dataset to regress *MPG Highway* on the five independent variables *Engine Size*, *Horsepower*, *Fueltank*, *Wheelbase*, and *Weight* in the *Regression Model Selection* utility, you should see:

Regression Model Selection - MPG Highway

Dependent variable: MPG Highway
Independent variables:
A=Engine Size
B=Horsepower
C=Fueltank
D=Wheelbase
E=Weight

Models with Best Information Criteria

MSE	*Coefficients*	*AIC*	*HQC*	*SBIC*	*Included Variables*
8.34831	5	2.22959	2.28456	2.36575	ACDE
8.63509	4	2.24186	2.28584	2.35078	CDE
8.68082	4	2.24714	2.29112	2.35607	ADE
8.51991	5	2.24993	2.30491	2.38609	BCDE
8.37003	6	2.25369	2.31966	2.41708	ABCDE
8.94952	3	2.25612	2.2891	2.33781	DE
8.91719	4	2.274	2.31798	2.38293	BDE
8.74921	5	2.27649	2.33147	2.41265	ABDE
9.16113	4	2.30099	2.34497	2.40992	ACE
9.4185	3	2.30719	2.34018	2.38889	CE
9.6172	3	2.32807	2.36106	2.40977	AE
9.2125	5	2.32809	2.38307	2.46425	ABCE
9.85293	2	2.33078	2.35277	2.38524	E
9.52299	4	2.33973	2.38371	2.44866	BCE
9.93648	3	2.36073	2.39372	2.44243	BE
10.9827	2	2.43933	2.46132	2.49379	C
10.8958	3	2.4529	2.48588	2.53459	BC
11.0932	5	2.51386	2.56883	2.65002	ABCD
17.4487	2	2.90228	2.92427	2.95674	A
17.7262	2	2.91805	2.94005	2.97252	B
17.856	2	2.92535	2.94734	2.97982	D
28.4273	1	3.36886	3.37985	3.39609	

Notice that the model with the lowest AIC value excludes *Horsepower*, the best three-variable model excludes both *Horsepower* and *Engine Size*, and the two-variable model with only *Wheelbase* and *Weight* also looks pretty good.

As another example, consider the polynomial problem from homework 5. Entering the cubic polynomial model of *MPG Highway* regressed on Horsepower produces the following output in the *Regression Model Selection* utility.

Regression Model Selection - MPG Highway

Dependent variable: MPG Highway
Independent variables:
A=Horsepower
B=Horsepower^2
C=Horsepower^3

Number of complete cases: 93
Number of models fit: 8

Models with Best Information Criteria

MSE	*Coefficients*	*AIC*	*HQC*	*SBIC*	*Included Variables*
13.1749	4	2.66433	2.70831	2.77326	ABC
14.1891	3	2.71699	2.74998	2.79869	AB
14.6722	3	2.75047	2.78346	2.83217	AC
15.7534	3	2.82157	2.85456	2.90327	BC
17.7262	2	2.91805	2.94005	2.97252	A
20.8105	2	3.07847	3.10046	3.13293	B
23.2918	2	3.19111	3.21311	3.24558	C
28.4273	1	3.36886	3.37985	3.39609	

Notice the model selected by the Akaike Information Criterion is the cubic, with the quadratic model coming in second.

Automated model selection methods are usually just a starting point, winnowing possible models down to a manageable few, from which you can choose the one that fits your needs. There are several other selection criteria presented by Statgraphics in the *Best Information Criteria* window (and *Adjusted R-Squared* and *Mallows'* C_p appear in other windows). Although the AIC is popular, some combination of these may prove useful in selecting a final model.

Final Note: Statgraphics is usually pretty good at summarizing what it's doing when you are using a particular utility to analyze data. PDF files for these should have been downloaded onto your computer if you installed the program. Usually the end of the file contains the formulas the program uses when performing computations. (This is where I went to figure out how it calculates AIC.) There are also saved webinars that you can view online, although I haven't watched enough videos to have an opinion about them.

Mean Squared Error - Up until now, we've assumed that the statistics used to estimate parameters were unbiased. We now briefly consider the presence of bias in preparation for a discussion of Mallows $\mathbf{C_p}$ statistic, which provides another criterion for selecting the best model in regression. Below are the relevant definitions.

- Let θ be a parameter to be estimated.
- Let $\hat{\theta}$ be an estimator of θ. **Note:** θ is an unknown *constant*, while $\hat{\theta}$ is a *random variable*.
- $\hat{\theta}$ is an unbiased estimator of θ if $E(\hat{\theta}) = \theta$. Otherwise, $\hat{\theta}$ is said to be biased for θ.
- $E(\hat{\theta}) - \theta$ is the *bias* (which is zero if $\hat{\theta}$ is an unbiased estimator of θ).
- $\hat{\theta} - \theta$ is the *error*, also called the estimation error or the prediction error.
- $E\left(\hat{\theta} - \theta\right)^2$ is the true Mean Squared Error, or *MSE*.

With a view toward investigating the bias, we proceed as follows. (This should look like my Stat 50 notes)

1. Rewrite the *mean squared error*: $E\left(\hat{\theta}-\theta\right)^2 = E\left(\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\theta\right)^2$ We just added $\pm E(\hat{\theta})$
2. Regrouping: $= E\left(\left[\hat{\theta}-E(\hat{\theta})\right]+\left[E(\hat{\theta})-\theta\right]\right)^2$ An old Algebra trick
3. Expanding the right-hand side of the equation and simplifying:
   $= E\left(\hat{\theta}-E(\hat{\theta})\right)^2 + 2\left(E(\hat{\theta})-\theta\right)E\left(\hat{\theta}-E(\hat{\theta})\right) + E\left(E(\hat{\theta})-\theta\right)^2$, where the constant factor in the middle term, $2\left(E(\hat{\theta})-\theta\right)$, factors through the expectation operator $E(\cdot)$
4. Note that a factor in the middle term equals zero: $E\left(\hat{\theta}-E(\hat{\theta})\right) = E(\hat{\theta})-E(\hat{\theta}) = 0$.
5. Note that the last term $E\left(E(\hat{\theta})-\theta\right)^2$ equals $\left(E(\hat{\theta})-\theta\right)^2$. Because $\left(E(\hat{\theta})-\theta\right)^2$ is constant
6. Finally, putting it all together: $E\left(\hat{\theta}-\theta\right)^2 = E\left(\hat{\theta}-E(\hat{\theta})\right)^2 + \left(E(\hat{\theta})-\theta\right)^2$, or $MSE = \sigma^2_{\hat{\theta}} + (Bias)^2$, which says that the mean square error of an estimator equals the sum of its variance and the square of its bias.

Mallows' C_p - We start with a multiple regression model that contains all potentially relevant independent variables. We'll call this the **full** model. Let ***q*** be the number of parameters fit by this model, including the constant. We assume that the full model contains no bias (this means that the expected response vector lies in the column space of the design matrix for the full model). We also assume the random errors ε_i associated with each observation are independent with constant variance. We then consider a smaller model that uses a subset ***p*** of the ***q*** independent variables in the full model. As usual, ***n*** is the number of observations in the sample. Then Mallows' C_p statistic is computed for the ***p*** model and used to evaluate it.

How the Statistic Works: Mallows' C_p investigates the bias in the smaller ***p***-model. Specifically, it considers the total squared bias in estimating $E(\boldsymbol{Y})$ by $\hat{\boldsymbol{Y}}$, i.e., $\sum_{i=1}^{n}\left(E(\hat{\boldsymbol{Y}}_i)-\mu_{\boldsymbol{Y}_i}\right)^2$.

The Formula for the Statistic: Mallows' $\mathbf{C_p} = \dfrac{SSE_p}{MSE_q} - (n-2p)$, where

- SSE_p is the error sum of squares for the smaller ***p***-model.
- MSE_q is the mean square error for the full ***q***-model. Because the full model is assumed unbiased, MSE_q is an unbiased estimator of the common error variance σ^2.

The Decision Rule: Any ***p***-model whose $C_p \approx p$ is thought to be unbiased, and becomes a candidate for selection as the final regression model.

Statgraphics: Follow *Relate > Multiple Factors > Regression Model Selection* and enter the variables as you would in a multiple regression analysis. Your first option is to enter the maximum number of independent variables you wish to have in your final model. Next, check the *Analysis Summary* and *Best C_p* tables. Looking at the *Analysis Summary*, you'll notice that Statgraphics is using best subsets selection, i.e., fitting all models that have the number of independent variables you specified earlier.

Example: Returning to regress *MPG Highway* on the five independent variables *Engine Size*, *Horsepower*, *Fueltank*, *Wheelbase*, and *Weight* in the *Regression Model Selection* utility, you should see:

Regression Model Selection - MPG Highway

Dependent variable: MPG Highway
Independent variables:
A=Engine Size
B=Horsepower
C=Fueltank
D=Wheelbase
E=Weight

Models with Smallest Cp

		Adjusted		*Included*
MSE	*R-Squared*	*R-Squared*	*Cp*	*Variables*
8.34831	71.9096	70.6328	4.77159	ACDE
8.37003	72.1566	70.5564	6.0	ABCDE
8.51991	71.3322	70.0291	6.57572	BCDE
8.63509	70.6145	69.6239	6.81842	CDE
8.68082	70.4589	69.4631	7.30463	ADE
8.74921	70.5607	69.2225	8.98654	ABDE
8.94952	69.2023	68.5179	9.23099	DE
8.91719	69.6545	68.6316	9.81802	BDE
9.16113	68.8243	67.7735	12.4119	ACE
9.2125	69.0018	67.5928	13.8574	ABCE
9.4185	67.5884	66.8681	14.2738	CE
9.52299	67.5929	66.5005	16.2596	BCE
9.6172	66.9046	66.1691	16.4104	AE
9.85293	65.7166	65.3399	18.1222	E
9.93648	65.8059	65.046	19.8434	BE
10.8958	62.5044	61.6712	30.1591	BC
10.9827	61.7857	61.3657	30.405	C
11.0932	62.6737	60.977	33.6303	ABCD
17.4487	39.2871	38.62	100.704	A
17.7262	38.3215	37.6437	103.721	B
17.856	37.8698	37.187	105.133	D

Notice that in this example the Akaike Information Criterion and Mallows' C_p both select the same model (the one that excludes *Horsepower*).

Discussion: Automated model selection criteria, such as $\boldsymbol{R}^2$-adjusted, **AIC**, and Mallows' $\mathbf{C_p}$, are advisory only. They only evaluate the models you feed them. For instance, none of the selection criteria would have chosen the cubic *Horsepower* polynomial model in the second example if I hadn't included the Horsepower^3 term in the input window to the *Regression Model Selection* utility. They also don't run diagnostics on assumptions about the error variable. They are only intended to be one of the tools used to produce a satisfactory model that can be defended and utilized.

Finally, this course has only scratched the surface with respect to model building. If you end up using regression in your professional or academic future, you will undoubtedly have to learn more about both the theory and practice of regression.