

NOTES ON STEPWISE REGRESSION

In multiple regression problems one often has available a large number of potential explanatory variables.

In predicting the price of a house, for example, there are generally a multitude of housing attributes which could influence price. While it is in principle possible to run all possible regressions and compare adjusted R^2 , this is practically impossible in most cases (10 variables allows $2^{10} = 1024$ regressions). A practical procedure for 'searching' the space of possible regressions is Stepwise Regression:

This method can be illustrated by the housing data in `housing-s.jmp`, in which price (PR), is regressed against no. of bedrooms (BDR), stormwindows (ST), garage (GAR), location (LOC) and construction type (CSTR). [This is a variation of problem 7 in Problem Set 3.]. If we open 'Fit Model' with all variables entered, and use option 'Stepwise' [rather than 'Standard Least Squares'] then the following menu appears:

Stepwise Regression Control							
Prob to Enter	0.250	Enter All					
Prob to Leave	0.100	Remove All					
Direction	Forward						
Go	Stop	Step	Make Model				

Current Estimates							
SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC	
3957.3846	25	158.2954	0.0000	0.0000	32.12125	132.6563	
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	56.1538462	1	0	0.000	1.0000
<input type="checkbox"/>	<input type="checkbox"/>	BDR	?	1	1227.877	10.796	0.0031
<input type="checkbox"/>	<input type="checkbox"/>	ST	?	1	1512.035	14.840	0.0008
<input type="checkbox"/>	<input type="checkbox"/>	GAR	?	1	1171.047	10.087	0.0041
<input type="checkbox"/>	<input type="checkbox"/>	LOC	?	1	0.009615	0.000	0.9940
<input type="checkbox"/>	<input type="checkbox"/>	CSTR	?	1	29.54219	0.181	0.6747

(2)

The simplest way to understand this procedure is to press 'step'. The new tableau is shown below (on the left):

Current Estimates								Summary of Fit		
		SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC	RSquare	0.382079
		2445.35	24	101.8896	0.3821	0.3563	12.67848	122.14	RSquare Adj	0.356333
									Root Mean Square Error	10.09404
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"		Mean of Response	56.15385
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	50.125	1	0	0.000	1.0000		Observations (or Sum Wgts)	26
<input type="checkbox"/>	<input checked="" type="checkbox"/>	BDR	?	1	574.0347	7.055	0.0141			
<input type="checkbox"/>	<input checked="" type="checkbox"/>	ST	15.675	1	1512.035	14.840	0.0008			
<input type="checkbox"/>	<input type="checkbox"/>	GAR	?	1	419.9147	4.768	0.0394			
<input type="checkbox"/>	<input type="checkbox"/>	LOC	?	1	40.01667	0.383	0.5423			
<input type="checkbox"/>	<input type="checkbox"/>	CSTR	?	1	16.01667	0.152	0.7006			
Parameter Estimates										
Term	Estimate	Std Error	t Ratio	Prob> t						
Intercept	50.125	2.523509	19.86	<.0001						
ST	15.675	4.069036	3.85	0.0008						

Notice that the variable ST has been entered. By comparing this tableau with the initial one, you will see that all numbers in the ST-row have not changed, but all others have changed. To see what is happening, look at the results of regressing PR on ST, shown above on the right. Notice that the p-value, .0008, is the same as the value of "Prob > F" in the stepwise tableau. (Note also that the "F Ratio", 14.84, is simply the square of the regression t-value, 3.85. Here t^2 is the simplest instance of an F-statistic, which we shall not discuss). More generally, all "Prob > F" entries in the initial tableau are simply the p-values for the $\hat{\beta}_i$ that would result from regressing each individual variable i against PR. Hence the variable entered by the stepwise regression is always the one with the lowest p-value (i.e. with the most significant slope coefficient). Notice also that the adjusted R^2 , .356, achieved by entering ST is higher than for any other single variable. [Here Mallows' "Cp" and Akaike's "AIC" are alternative measures of 'goodness of fit', which we shall not discuss.]

③

Next click 'Step' again and observe that BDR is now entered. Again we see from the tableau below that the values in the row entered remain unchanged from the previous tableau. Moreover, the "Prob > F" value for BDR, .0141, is again seen to agree with the p-value for the slope coefficient of BDR in the two-variable regression on the right (and again the "F Ratio" is simply t^2).

Current Estimates							
	SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
	1871.3153	23	81.36153	0.5271	0.4860	6.537867	117.1838
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	37.4734134	1	0	0.000	1.0000
<input type="checkbox"/>	<input checked="" type="checkbox"/>	BDR	3.96912521	1	574.0347	7.055	0.0141
<input type="checkbox"/>	<input checked="" type="checkbox"/>	ST	12.4500858	1	858.1928	10.548	0.0035
<input type="checkbox"/>	<input type="checkbox"/>	GAR	?	1	298.566	4.176	0.0531
<input type="checkbox"/>	<input type="checkbox"/>	LOC	?	1	25.52341	0.304	0.5868
<input type="checkbox"/>	<input type="checkbox"/>	CSTR	?	1	73.73193	0.902	0.3525

Summary of Fit	
RSquare	0.527133
RSquare Adj	0.486014
Root Mean Square Error	9.020063
Mean of Response	56.15385
Observations (or Sum Wgts)	26

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	37.473413	5.269893	7.11	<.0001
ST	12.450086	3.833447	3.25	0.0035
BDR	3.9691252	1.494291	2.66	0.0141

Hence it should now be clear that all "Prob > F" values in the previous tableau (other than the variable, ST, already entered) correspond to the p-values for the appropriate slope coefficient in each possible two-variable regression (given ST). The variable entered, BDR, is thus the one with the lowest p-value (most significant slope coefficient) in the corresponding two-variable regression. Notice again that the adjusted R² has increased to .486 (and is higher than that of each other possible two-variable regression given ST).

By clicking 'Step' again, we see that GAR is now entered, and that the same pattern holds true. Namely, all "Prob > F" values for unentered variables in the tableau above correspond to p-values for the appropriate slope coefficients

(4)

in three-variable regressions. Clearly, GAR has the lowest such p -value, .0531. Again the new table and corresponding regression shown below yield a higher Adjusted R^2 value, .548.

Current Estimates							
	SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
	1572.7493	22	71.4896	0.6026	0.5484	4.303783	114.6646
Lock	Entered	Parameter	Estimate	nDF	SS	F Ratio	Prob>F
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	36.0291871	1	0	0.000	1.0000
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	BDR	3.56045679	1	452.6861	6.332	0.0196
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	ST	9.74612914	1	463.1123	6.478	0.0184
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	GAR	4.62628529	1	298.566	4.176	0.0531
<input type="checkbox"/>	<input type="checkbox"/>	LOC	?	1	16.62909	0.224	0.6406
<input type="checkbox"/>	<input type="checkbox"/>	CSTR	?	1	93.60663	1.329	0.2619

Summary of Fit	
RSquare	0.602579
RSquare Adj	0.548385
Root Mean Square Error	8.455093
Mean of Response	56.15385
Observations (or Sum Wgts)	26

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	36.029187	4.990109	7.22	<.0001
ST	9.7461291	3.829194	2.55	0.0184
BDR	3.5604568	1.414899	2.52	0.0196
GAR	4.6262853	2.263761	2.04	0.0531

Note finally that by clicking "Step" once again, nothing happens! Neither of the remaining two variables (LOC, CSTR) are entered. To see why, return to the original tab and notice at the top that "Prob to Enter = .25". This setting precludes any variable from being entered with a p -value $> .25$ (such as the p -values, .6406 and .2619, for LOC and CSTR, respectively.) This is simply a default setting which can easily be changed by clicking on the value and typing a new one.

This example also shows that the present 'minimum P -value' criteria need not agree with increases (or decreases) in Adjusted R^2 . By running a four-variable regression including (BDR, ST, GAR) together with each omitted variable, you will see that Adjusted R^2 decreases to .5312 by entering LOC, but increases to .5550 by adding CSTR. Hence the focus here is on significance of the individual variables rather than on degree of overall explanation.

(5)

So far we have only illustrated the 'Forward' option in Stepwise Regression. There is also a 'Backward' option in which one starts with all variables and then successively deletes those values with largest P-values until no more can be removed under the 'Prob to Leave' criteria which is specified.

By far the most important option is the 'Mixed' option which successively adds and deletes variables according to the two criteria. The default settings 'Prob to Enter' = .25 and 'Prob to Leave' = .10 can be changed. You should experiment a bit to see what happens when these are changed. (I prefer 'Prob to Enter' = .15).

→ When doing Stepwise Regression, always use the Mixed Option.

ALL-MODELS OPTION IN STEPWISE REGRESSION

The more recent versions of JMP offer a practical implementation of the “all-models” option in which it is possible to compare *all possible* regression models in terms of goodness of fit. The key shortcoming of this procedure is that it uses R-square comparisons rather than *adjusted* R-square, which (as I pointed out in class) will automatically favor models with more explanatory variables. Hence the purpose of these notes is to outline the JMP procedure for *All Possible Models* together with an extension of this procedure that allows adjusted R-square (R_{adj}^2) comparisons.

Here we continue with the **Housing_s.jmp** example. Assuming that the **Stepwise** tableau is open, you can access the *All Possible Models* option by right clicking on the top bar, labeled **Stepwise Fit**. Here you will be asked to specify the maximum number of explanatory variables, k , in the models to be considered, and the maximum number of “best models” (in terms of R-square) to be displayed for each model size, k . The default options are a good place to start. [Remember that if you have say 10 explanatory variables and try to display *all* models, you will be looking at $2^{10} - 1 = 1023$ models!]

If you then click **OK**, all models will be displayed at the bottom of the tableau, organized by increasing values of k . If you then select any of these models (by the buttons on the right) you can see the full tableau for this model – including R_{adj}^2 . But if you want to order all models in terms of R_{adj}^2 , then there is no automatic method for doing so in JMP. However, if you open the companion data set, **Housing_s_All_Models.jmp**, in the class directory then you will see an example of how to display all values of R_{adj}^2 automatically.

To reproduce this table, first right click anywhere on the model display at the bottom of the Stepwise tableau, and select “Make into Data Table”. This will yield a file that looks very much like **Housing_s_All_Models.jmp**. But some changes must be made.

- (i) First, change the column label, **Number**, to **k**. This will be used as an argument in the construction of R_{adj}^2 .
- (ii) Next click **Col** → **Add Multiple Columns** and add two columns. Label the first as **n** (for sample size). Open the formula window for this column and insert the number of samples, which in this housing data set is **26**.
- (iii) Finally label the last column as **Adj Rsquare** and open the formula for this column. The desired formula [which can be found in Devore (7th Edition) p.522] is given by

$$\frac{(n-1)R_{\text{square}}-k}{n-1-k}$$

- (iv) Alternatively, you can simply open the corresponding formula in the file **Housing_s_All_Models.jmp** and copy-paste this into your formula window.
- (v) You should then obtain a list of all R_{adj}^2 values in this last column. To sort these values by size, click **Tables** → **Sort** and then sort by **Adj Rsquare**. In doing so, you can sort in *descending order* by clicking on the “triangle” icon to the left of **Adj Rsquare** and then replacing it with the “inverted triangle” icon below (by clicking on this icon).

If you have done this correctly, a new copy of the file will open with model **(BDR,ST,GAR,CSTR)** in the first row, having the maximum value of R_{adj}^2 , namely **0.55503843**. You can now compare models in terms of their overall goodness of fit.

Finally, remember that this comparison focuses only on R_{adj}^2 , and does not indicate which of these model variables is significant in terms of *p-values*. One can check these by running the individual regressions. Also, it is much safer to run the regressions (by clicking “Make Model” in the **Stepwise Regression Control** at the top of the tableau). If you simply rely on the p-values in the stepwise tableau, then you will find that sometimes these will *change* in the final regression, especially when you have some missing data entries.