Stat 102

Introduction to Business Statistics
Class 19

Richard P. Waterman

Wharton

Table of contents I

1. Today's module

2. Last time

3. Stepwise Regression

4. Review

5. JMP tasks

6. Next time

Today's module

Topics to be covered in this module:

- Last time
- Introduction to stepwise regression.
- Illustrate the JMP platform.
- Identify the BIG problem – overfitting.
- Alternative stopping rules (in addition to p-values)
  1. RMSE
  2. Adjusted R^2
  3. Mallows' C_p
  4. Akaike Information Criterion (AIC)
  5. Bayesian Information Criterion (BIC)
- Comparing the RMSE in and out-of-sample.
- Summary
- Next time

Last time

- Ways of correcting for multiplicity:
  1. Tukey's HSD
  2. Bonferroni
  3. The False Discovery Rate, adjusting the p-values

The Apple data set

- The goal: based on data available today, predict Apple's return tomorrow.
- We have 106 data points, 42 are *held out*, October onward.
- Variables: 10 other stocks `GOOG, INTC` *etc*.
- Features provide 41 variables to choose from: price, volume, number of trades and returns
- Even just considering main effects (no interactions or squares) there are 2^{41} possible models. That is 2,199,023,255,552 (two trillion) different models to explore.
- Including all possible interactions and squares there are
  $$\underbrace{41}_{maineffects} + \underbrace{41}_{squares} + \underbrace{\frac{1}{2} \times 40 \times 41}_{interactions} = 902 \text{ terms and}$$
  $2^{902} = 3.381 \times 10^{271}$ possible models[1].

[1] The number of atoms in the universe is estimated at 10^{80} atoms.

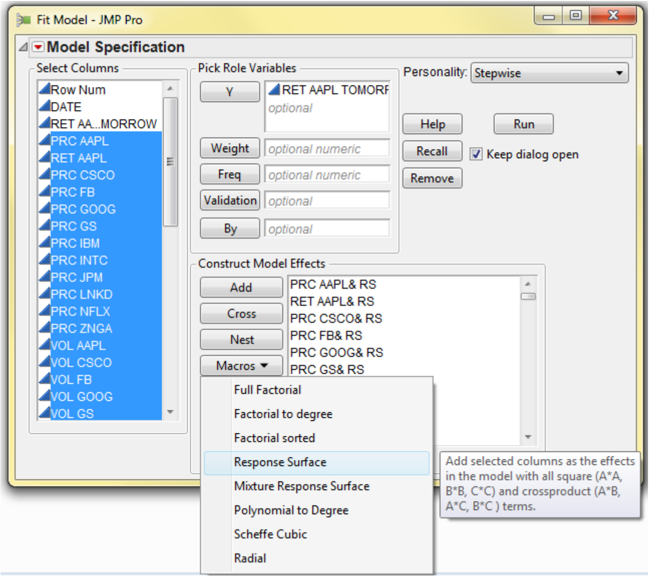
Heuristic

- So it is a big problem. A big space of models to explore – the so called *curse of dimensionality*.
- We bring a *heuristic* to the problem.
- *Merriam Webster: involving or serving as an aid to learning, discovery, or problem-solving by experimental and especially trial-and-error methods*
- Then follow our natural instinct: iterative model fitting. Find the single best variable. Given this variable, find the second best. Given these two, find the third best and so on.
- The essence of stepwise regression, the original automated model selection tool.

Elements of stepwise

- Choose a direction to make a path through the big model space
  1. Forward selection.
  2. Backwards elimination.
  3. Forwards and backwards = mixed.
  4. When there is collinearity these will not necessarily identify the same model.
- A rule for variable selection.
  1. Add step: the one with the lowest p-value (or R^2 up the most).
  2. Removal: the one with the highest p-value (or R^2 down the most).
- A rule for stopping.
  1. P-Value Threshold
  2. AICc
  3. BIC
- Details: rules for treating categorical variables and interactions.

Choosing the variables to offer to stepwise



Fit Model - JMP Pro

Model Specification

Select Columns

Row Num
DATE
RET AA...MORROW
PRC AAPL
RET AAPL
PRC CSCO
PRC FB
PRC GOOG
PRC GS
PRC IBM
PRC INTC
PRC JPM
PRC LNKD
PRC NFLX
PRC ZNGA
VOL AAPL
VOL CSCO
VOL FB
VOL GOOG
VOL GS

Pick Role Variables

Y: RET AAPL TOMORF

Y: optional

Weight: optional numeric

Freq: optional numeric

Validation: optional

By: optional

Personality: Stepwise

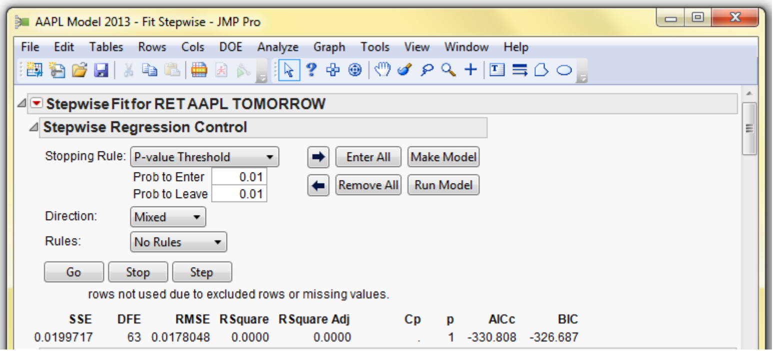
☑ Keep dialog open

Construct Model Effects

PRC AAPL& RS
RET AAPL& RS
PRC CSCO& RS
PRC FB& RS
PRC GOOG& RS
PRC GS& RS

Add selected columns as the effects in the model with all square (A*A, B*B, C*C) and crossproduct (A*B, A*C, B*C) terms.

The stepwise dialog



AAPL Model 2013 - Fit Stepwise - JMP Pro
File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help
Enter All
Make Model
Remove All
Run Model
Go
Stop
Step

Stepwise Fit for RET AAPL TOMORROW

Stepwise Regression Control

Stopping Rule: P-value Threshold

Prob to Enter 0.01

Prob to Leave 0.01

Direction: Mixed

Rules: No Rules

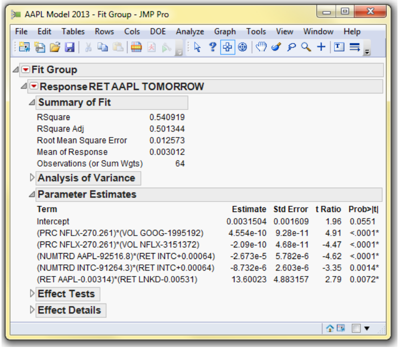
rows not used due to excluded rows or missing values.

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
0.0199717	63	0.0178048	0.0000	0.0000	.	1	-330.808	-326.687

The stepwise elements

- Variable selection: which one makes R^2 go *up or down the most*.
- Direction: *mixed*
- Stopping rule: *P-Value threshold*
- Rules: *No rules*!

Output from the chosen model



Fit Group

Response RET AAPL TOMORROW

Summary of Fit

RSquare	0.540919
RSquare Adj	0.501344
Root Mean Square Error	0.012573
Mean of Response	0.003012
Observations (or Sum Wgts)	64

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0031504	0.001609	1.96	0.0551
(PRC NFLX-270.261)*(VOL GOOG-1995192)	4.554e-10	9.28e-11	4.91	<.0001*
(PRC NFLX-270.261)*(VOL NFLX-3151372)	-2.09e-10	4.68e-11	-4.47	<.0001*
(NUMTRD AAPL-92516.8)*(RET INTC+0.00064)	-2.673e-5	5.782e-6	-4.62	<.0001*
(NUMTRD INTC-91264.3)*(RET INTC+0.00064)	-8.732e-6	2.603e-6	-3.35	0.0014*
(RET AAPL-0.00314)*(RET LNKD-0.00531)	13.60023	4.883157	2.79	0.0072*

Effect Tests

Effect Details

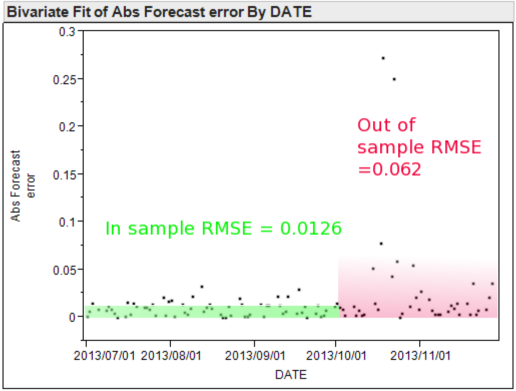
All variables highly significant. $R^2 = 54\%$. RMSE $= 0.012573$. The initial raw standard deviation of the returns was 0.0178048. So RMSE is

$$\frac{0.012573}{0.0178048} = 70.616\%$$

of the initial unexplained variation. The model looks OK.

Out-of-sample prediction

Disaster strikes! A plot of the absolute forecast error, both in and out-of-sample.



Overfitting and the in and out-of-sample RMSE

- This is overfitting, the big danger of greedy algorithms run amok. The model actually performs much worse out-of-sample than the in-sample summaries suggest.
- In-sample RMSE $= 0.012573$.
- Out-of-sample RMSE $= 0.0623$.
- A 500% inflation factor.

Comments on stepwise regression

- Can use stepwise after a hand-crafted model has been made to make sure nothing has been overlooked.
- Stepwise can't find variables unless you offer them to it!
- Stepwise can't think about transformations and normalization.
- Stepwise can't help in interpretation.
- Stepwise looks one step ahead. It is a greedy algorithm; that is one that makes locally optimal decisions in the hope that it comes close to a globally optimal one. You could imagine looking over pairs of variables or triplets, rather than one at a time. Kasparov looked 3-5 moves ahead in chess and sometimes as many as 12. Stepwise looks one step ahead!
- Use *center polynomials* to reduce collinearity.

Use stepwise as a validation/exploratory tool, not as the only approach.

Stopping criteria other than the p-value cut-off

K.I.S.S = Occam's razor = Parsimony

Among competing theories that equally well explain the observations, choose the one that is simplest.

- Comparing R^2 (the same as minimizing Sums of Squared Error [SSE]) across models doesn't capture the idea of simplicity.
- The sums of squared error:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2,$$

the sum of the squares of the residuals.

Stopping criteria other than the p-value cut-off

- RMSE doesn't capture simplicity either. Two models can have the same RMSE but that doesn't distinguish the complexity of the models.
- Unlike regular R^2, Adjusted R^2 doesn't have to increase with additional variables so looks like a better choice, but

$$\text{Adjusted} R^2 = \left(1 - \frac{RMSE^2}{s_y^2}\right),$$

so maximizing Adjusted R^2 is equivalent to minimizing RMSE.

We need a new idea.

Explicitly incorporating complexity in the model selection criterion

Rather than choosing the model with the smallest sums of squares error (SSE), you can penalize more complex models directly through the number of variables included (k is the number of variables in the model and $\hat{\sigma}^2$ is an estimate of the variance of the ϵ_i).

1. Mallows' $C_p = \frac{SSE_k}{\hat{\sigma}^2} - n + 2k$.
2. Akaike Information Criterion: $AIC(k) \propto \frac{SSE_k}{\hat{\sigma}^2} + 2k$.
3. Bayesian Information Criterion: $BIC(k) \propto \frac{SSE_k}{\hat{\sigma}^2} + \log(n)k$.

Notice that as k goes up, the penalty term increases. With normal error terms C_p and AIC are equivalent.

BIC penalizes complexity more than AIC (when $\log(n) > 2$) so prefers smaller models.

These are the other Stopping Rules in the stepwise dialog.

These stopping rules are appropriate when the goal is model selection for prediction.

Interpreting the model selection criteria through t-stats

We add a variable to the model if the increased complexity (k goes up) is appropriately offset with a smaller SSE.
One can show that a variable is added to an existing model if:

Criterion	Approx $\|t\|$ cut-off	Equiv. p-value	Goal
Adjusted R^2	$\|t\| > 1$	0.33	Minimize RMSE
C_p / AIC	$\|t\| > \sqrt{2}$	0.16	Achieve an unbiased estimate of prediction accuracy
BIC	$\|t\| > \frac{1}{2}\log(n)$	Depends on n	Something Bayesian!

Recall that in standard hypothesis testing a significant t is one such that $|t| > 2$.

The reason to adjust for complexity

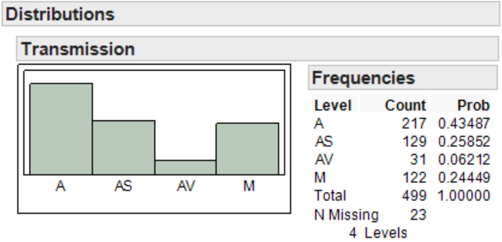
- Statistics estimates parameters through optimization – typically by making something as small as possible.
- In particular, in regression, by making SSE as small as possible.
- Hence it provides an over-optimistic view of what will happen in practice.
- Penalizing by the number of parameters in the model is a way to mitigate this over-optimism.
- Recommendation: use AIC when you are looking for a predictive model through stepwise. Choose the model with the lowest AIC.

Categorical variables in stepwise

- JMP creates two-level contrasts of the categorical variables.
- Example: After fitting a stepwise model with a categorical variable (`Transmission`), run the model and then look in the datatable of the `Car_08` dataset to see the coding.
- If a categorical is selected by stepwise, you could choose to put the entire variable into the final model, or you could create an interpretable recoding or even use the contrasts chosen by stepwise itself.

Illustration of the coding for the `Transmission` variable

Remember that JMP will use a {+1,-1,0} coding by default, and not the {+1,0} dummy variable coding scheme we saw earlier.



Distributions

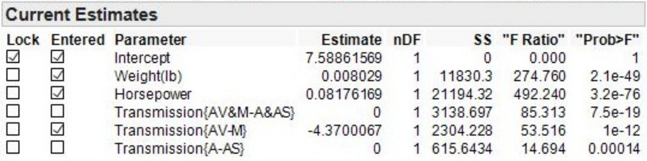
Transmission

Frequencies

Level	Count	Prob
A	217	0.43487
AS	129	0.25852
AV	31	0.06212
M	122	0.24449
Total	499	1.00000
N Missing	23	

4 Levels

`Transmission` is a four level categorical with levels {A, AS, AV, M}.



Current Estimates

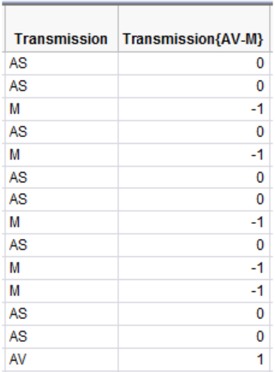
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	7.58861569	1	0	0.000	1
□	☑	Weight(lb)	0.008029	1	11830.3	274.760	2.1e-49
□	☑	Horsepower	0.08176169	1	21194.32	492.240	3.2e-76
□	□	Transmission{AV&M-A&AS}	0	1	3138.697	85.313	7.5e-19
□	☑	Transmission{AV-M}	-4.3700067	1	2304.228	53.516	1e-12
□	□	Transmission{A-AS}	0	1	615.6434	14.694	0.00014

Interpretation of the categorical variable parameters

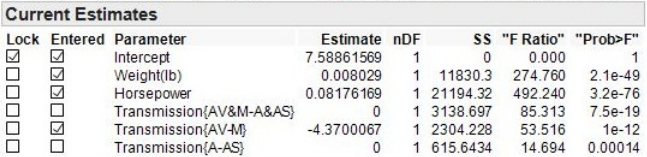
JMP notation: "&" means to combine the categories and "-" means to compare/contrast them.

Contrast	A	AS	AV	M
AV&M - A&AS	-1	-1	+1	+1
AV - M	0	0	+1	-1
A - AS	+1	-1	0	0

Repesentation of the categorical variable in the data table

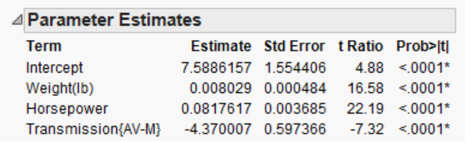


Transmission	Transmission{AV-M}
AS	0
AS	0
M	-1
AS	0
M	-1
AS	0
AS	0
M	-1
AS	0
M	-1
M	-1
AS	0
AS	0
AV	1



Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	7.58861569	1	0	0.000	1
□	☑	Weight(lb)	0.008029	1	11830.3	274.760	2.1e-49
□	☑	Horsepower	0.08176169	1	21194.32	492.240	3.2e-76
□	□	Transmission{AV&M-A&AS}	0	1	3138.697	85.313	7.5e-19
□	☑	Transmission{AV-M}	-4.3700067	1	2304.228	53.516	1e-12
□	□	Transmission{A-AS}	0	1	615.6434	14.694	0.00014



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	7.5886157	1.554406	4.88	<.0001*
Weight(lb)	0.008029	0.000484	16.58	<.0001*
Horsepower	0.0817617	0.003685	22.19	<.0001*
Transmission{AV-M}	-4.370007	0.597366	-7.32	<.0001*

Adding the variable Transmission{AV - M} to the model adds a parameter estimate. Its value is -4.37. So the AV's forecast changes by -4.37, the M's changes by +4.37 and the A's and AS's change by 0.

Summary

- The need for tools like stepwise – the space of all models is typically too big to exhaustively explore.
- The mechanics of stepwise; stopping rules, variable selection criterion.
- The big issue with stepwise: over-fitting.
- The Information criteria that penalize complexity.
- How JMP treats categoricals in stepwise.

JMP tasks

Make sure you can:

1. Use the model dialog to include all interactions and squares as potential model effects.
2. Run the stepwise tool.

Creating all interactions and squared terms

In the fit model dialog
Select the X-variables of interest
In model effects, go to Macros
Choose Response Surface.

Stepwise dialog

JMP stepwise using the `p-value threshold rule`:

```
After having selected the X-variables
in the Fit Model dialog
In the Personality choose Stepwise
Check Keep dialog open and click Run
For Stopping Rule choose P-value Threshold
Enter Prob To Enter and Prob to Leave
For Direction choose Mixed
For Rules choose No rules
Click Step to step through the variables.
```

Next time

- The train and test paradigm