

Lecture 5. More Regression in Data Mining

Questions from Previous Classes?

Topics

More on missing values, selection bias, interactions

Cross validation and model validation

Stepwise on ANES 2008

Improving stepwise

Missing Values

Two columns for each original

- One column identifies the missing cases (dummy)
- Other column gets filled in. Interpret the filled in column as the interaction between the unobserved “real” variable and the missing value indicator.
- Instead of having X , I , $X*I$, we observe I and $X*I$.

Implications for stepwise

- No need to try the interaction between I and $X*I$... we already have this one.
- Messy problem in this use of generic tools for stepwise regression.
Tricky to count how many of those variables are “real” and not repeated copies of others in the data.

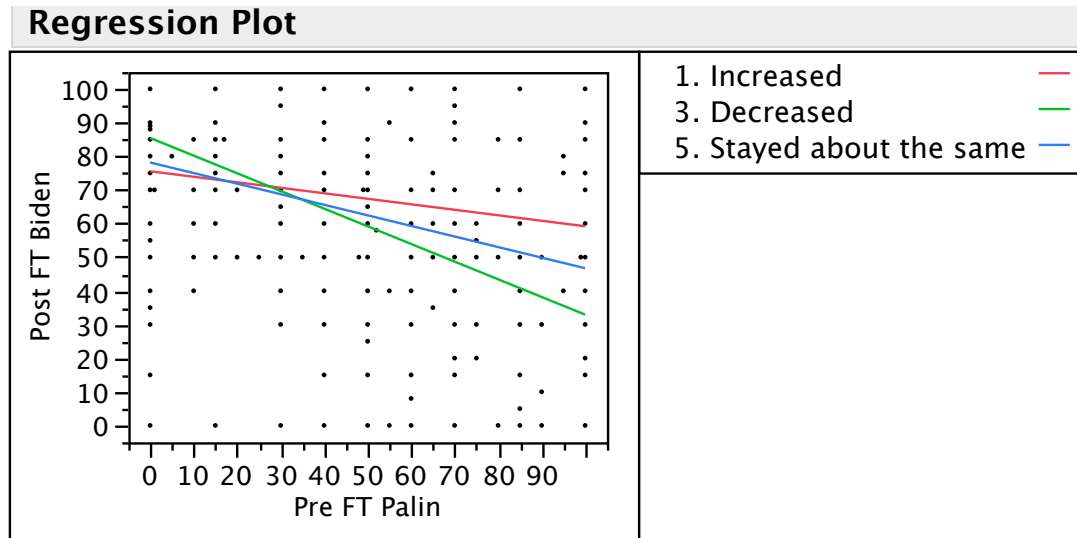
Selection Bias

Regression model

- Regression Post FT on Obama on the Pre FT Obama + State variable
 - Lots of indicator variables: how should we interpret statistical significance?
-

Interactions

Numeric * Categorical



- Response is post election FT for Biden with terrorism explanatory variables

Numeric * Numeric

- Surface plot is interesting and shows the amount of curvature.

Term	Estimate
Intercept	15.587588
Pre FT Obama	0.5180316
(Pre FT Obama-64.2774)*(Post FT Biden-62.0439)	-0.003577
Post FT Biden	0.397857

Comment on Lab Session

Sampling weights

- Need these to get sensible inferences (I've ignored this issue for simplicity)
- Vote analysis in JMP with sampling weights

Level	Count	Prob
1. Barack Obama	1025	0.65537
3. John McCain	514	0.32864
7. Other {SPECIFY}	25	0.01598
Total	1564	1.00000
N Missing	758	
3 Levels		

Level	Count	Prob
1. Barack Obama	82923237	0.53847
3. John McCain	68037440	0.44181
7. Other {SPECIFY}	3038112	0.01973
Total	2e+8	1.00000
N Missing	758	
3 Levels		

Recognition variables

- These in fact indicate whether the subject needed prompting to recognize name.

Negative sign for Biden

- Model for FT post Obama vs all pre FTs
- Rearrange slopes in model to interpret the presence of this negative sign.

Other variables?

- Added sex of R, sampling weights

Cross Validation

Hold-back sample(s)

- Reserve a portion of the data to test the accuracy of the final model
- Randomly divide data into two (or sometimes three subsets)
 - (a) training sample use these data to estimate the model
 - (b) tuning sample Optional: some methods need these to pick model
Stepwise has p-values and so it does not.
 - (c) test sample evaluate performance of model (eg, squared error)
- Repeat the procedure as often as you have time.

Example: 10-fold cross-validation slices the data into 10 subsets, fits/tests on 9/1 of the subsets, repeatedly excluding one of them. Plus, you can do this procedure multiple times, randomly deciding how to subset the data.

Example: split-sample CV randomly divides the data in half, using one half to predict the other half. Then reverse the roles. You can repeat this procedure multiple times as well, each time dividing into subsets at random.
- An algorithm like stepwise regression with the Bonferroni type p-values does not need to reserve a sample to pick the variables or test the model. The R^2 you get is an unbiased estimate of how well the model will perform.
- Even so, editors often demand a comparison to other results, and cross-validation supplies the recognized context.

Catch-22

- How much data to allocate for estimation versus testing?
- A large test sample gives a good estimate of accuracy, but leaves less for picking and estimating the model.
- A large estimation/tuning sample produces a better model, but one whose properties are less well understood.

Recognize over-fitting

- Comparison of the accuracy within the sample to the accuracy when the model is applied to the test sample.
- Over-fitting evident when the model claims to fit better than the test analysis shows.

Cross-validation is optimistic

- Test sample within cross-validation is (by the random selection used to subset the data) a random sample from the same population as the estimation sample.
- That's seldom the case in real modeling (populations drift over time)

Modeling with Stepwise

Modeling Problem

- Naive poly science response: Post FT for Obama
- Post FT for Obama is proxy for likelihood of voting for Obama
- Theory: What variables ought to be in the initial model?

Preliminary steps

- Explore expanded data set (anes2008_for_step)
- Removed cases not in the second wave (n falls to 2101)
- Need to build any other variables, transformations?

Post FT Biden - Pre FT Biden

- Exclude state?

Cross validation planning

- Must use this from the start, or its really not very legitimate.
- Randomly divide data in half (split-sample CV; JMP formula in first column)
- Fit models: saturated and stepwise
- Compare claims from fit to the training sample to accuracy on test sample.

Saturated linear model

- Cannot use all of the cross products... too many so just use all linear terms (also decided to exclude state)
- Lots of singularities since have redundant dummy variables.
- Model summary

212 non-redundant predictors, $R^2 = 0.80$, $\text{adj } R^2 = 0.75$, $\text{RMSE} = 13.3$

Stepwise model

- Selection from linear terms in saturated model, starting with empty model

Initial threshold $0.05/256 = 0.00020$

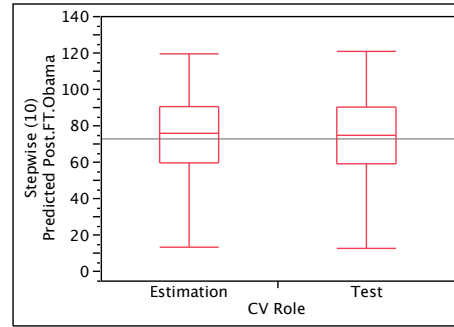
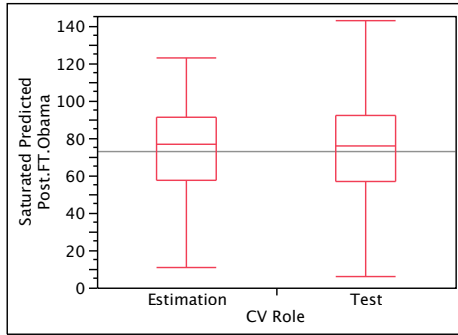
- Model summary

9 predictors, $R^2 = 0.74$, $\text{RMSE} = 13.7$

- Note the agreement between R^2 of the stepwise model and the adjusted R^2 of the saturated model. Stepwise also finds the simpler model that obtains that fit.

Cross validation comparison

- Compare the average squared error in the estimation sample to the average squared error in the test sample.
- Over-fitting is indicated if the error is much larger in the validation sample (we can use tests that compare variances, but usually the figure is enough).



- Neither is over-fit, but stepwise obtains a smaller SD in the validation data since it does not suffer so much from extrapolation of lots and lots of regressors

Bigger stepwise models?

- Response surface of FTs + all indicators

48 numerical variables + squares + interactions = $96 + 48 \cdot 47/2 = 1,224$
 208 categorical variables

$1224 + 208 = 1,432$ possible predictors

$p\text{-to-enter} = 0.05/1432 = 0.000035$

- Center the interactions
- Summary of model: better fit, with two interactions

9 predictors, $R^2 = 0.75$

Biden*Muslim interaction picture

Caution interpreting estimates

Term	Estimate	Std Error
Intercept	22.757	2.486
Pre.FT.Bush	-0.106	0.018
Pre.FT.Obama	0.420	0.022
Post.FT.Biden	0.286	0.025
FT.Fed.Govt	0.094	0.021
FT.Blacks	0.132	0.023
(Pre.FT.Dem.Party-63.1696)*(Post.FT.Biden-62.162)	-0.004	0.001
(Pre.FT.Biden-57.0607)*(FT.Muslims-51.9525)	-0.003	0.001
Party.ID.pre=1. Democrat[0]	-4.413	1.000
Afraid.of.Obama..pre=5. No[0]	-4.778	1.143

