Cross-Validation and Variable Selection Techniques

November 6, 2007

ERSH 8320

Today's Lecture

- Cross-validation.
- Variable selection techniques.

Shrinkage

- You collect data on multiple variables.
- You run a multiple regression.
- You get a high R^2 value.
- There is much rejoicing in your lab.
- But is the value you obtained for R^2 reliable?
- If you used the same regression parameters with a different sample, would you find the same R^2?

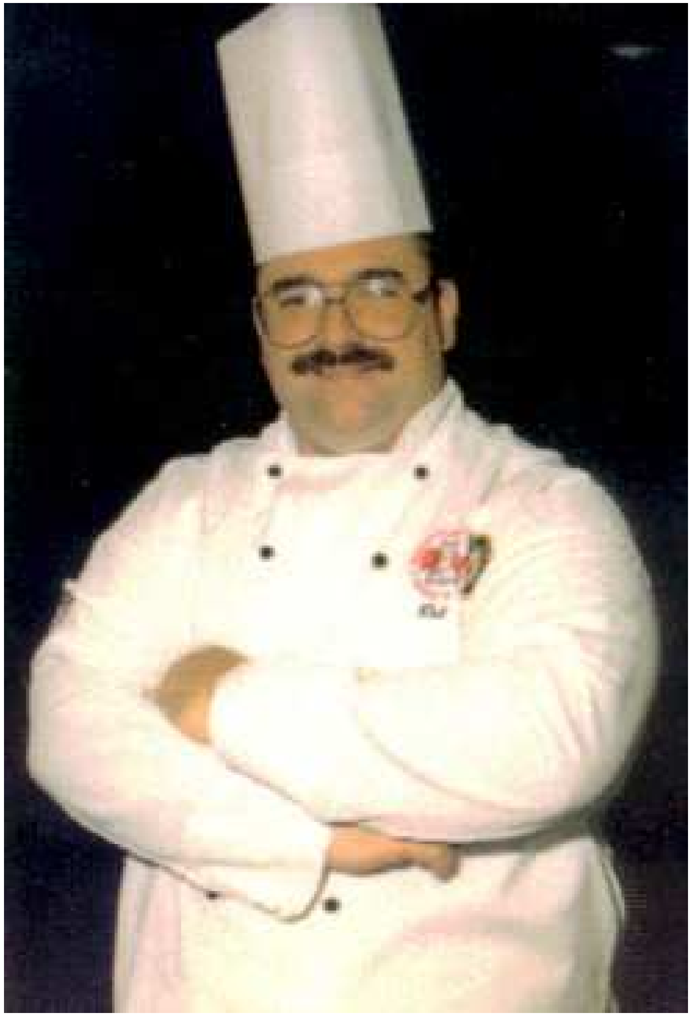
Shrinkage

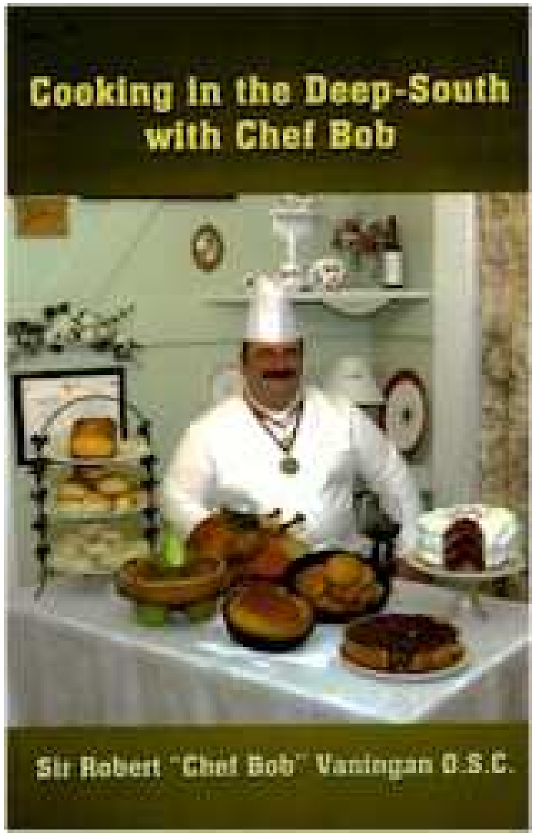
- Enviably, a regression you estimated based on one set of data will not have exactly the same fit on other sets of data.
- The R^2 value computed from an estimated regression line is typically over-estimated.
  - Meaning, R^2 is too large use when fitting the same regression on a different sample.

Shrinkage Example

- Imagine that you are an entrepreneur interested in making a new dessert cake that would be preferred over existing cakes already in the market.
- You have identified two variables that seem to play a primary role in the ratings people assign to other cakes:
  - Moisture content.
  - Sweetness.
- You are interested in determining how moisture content and sweetness play a role in cake preference ratings.
- You collect a random sample of cake eaters here in Athens.
- To be careful about differences in tastes in different locations, you also collect a random sample of cake eaters in the Atlanta area.

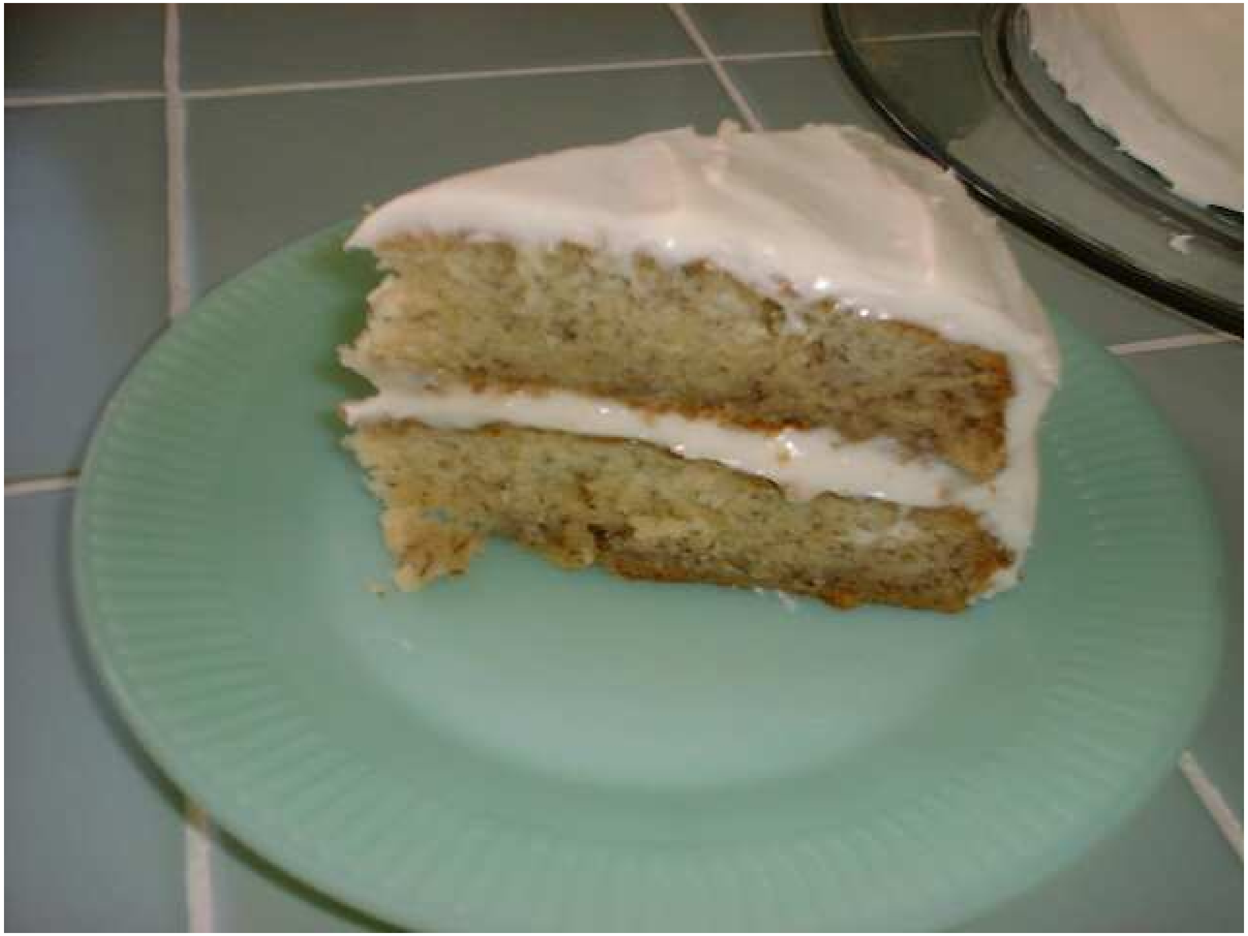
The Cook





Chef Bob.

The Cake



Cake.

The Example

- Using the data from Athens, fit the multiple regression where Y is preference rating, X_1 is moisture content and X_2 is sweetness.
- Obtain the predicted values of Y' for all observations (both in Athens and Atlanta).
- For each group calculate the pearson correlation between Y and Y'.
- For the Athens group, squared correlation of Y and Y' is the R^2 from the regression analysis.
- For the Atlanta group, R^2 is variance in Y explained when using regression parameters from Athens regression.
- Note how different these values are...

Factors that Play a Role in Shrinkage

- Ratio of number of predictors to the size of the sample.
  - Larger the ratio, larger the shrinkage.
  - No real rule of thumb, but power analysis can suggest adequate size of sample.
- Statistical bias in R^2.
  - If true R^2 was zero, expected sample R^2 is $\frac{k}{N-1}$
  - Adjustments can be made in R^2 so that an unbiased version can be ascertained.

Adjusted R^2

- The R^2 estimate can be adjusted using the following formula:

$$\hat{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-k-1}$$

- From our cake preference example, with two variables and eight observations:

$$R^2 = 0.923$$

$$\hat{R}^2 = 1 - (1 - 0.923)\frac{8-1}{8-2-1} = 0.892$$

Cross-Validation

- Instead of using the previous formula to estimate an adjusted R^2, one can instead divide up their sample and create another estimated of the R^2.
- This can be accomplished by splitting the data in half, and estimating a regression on one of the samples.
- Using this regression then compute Y' for the other half of the data.
- The squared correlation between Y and Y' will give an estimate of an adjusted R^2.

Cross-Validation

- In our cake example from previous slides, the procedure used was that of cross-validation.
- We divided our sample into two sets:
  - Athens sample. (Screening)
  - Atlanta sample. (Calibration)
- Using the Athens sample regression parameters, we calculated the R^2 for the Atlanta sample.
- In cross-validation, the parameters of the screening sample are used to compute R^2 in the calibration sample.

Other Cross-Validation Methods

- Double cross-validation.
- Cross-validation formulas.
- Bootstrapping.
- Jack-knifing.

Double Cross-Validation

- Double cross-validation takes cross-validation and applies an additional step:
  - Computation of two R^2 values: one for each sample.
  - R^2 is based on other sample's parameters.
- Double cross-validation steps:
  1. Split data into two samples.
  2. For each sample, estimate regression parameters.
  3. Estimate Y' for each observation, using other sample's regression parameters.
  4. Compute the correlation between Y and Y' for each sample.

Cross-Validation Coefficient

- Much like the adjusted R^2 formula, other formulas have been developed to find R^2 in double cross-validation.

- For times when predictor variables are fixed:

$$\hat{R}^2_{cv} = 1 - \left(\frac{N-1}{N}\right)\left(\frac{N+k+1}{N-k-1}\right)(1-R^2)$$

- For times when predictor variables are random:

$$\hat{R}^2_{cv} = 1 - \left(\frac{N-1}{N-k-1}\right)\left(\frac{N-2}{N-k-2}\right)\left(\frac{N+1}{N}\right)(1-R^2)$$

Cross-Validation Coefficient Example

- Using the cake example, with an R^2 of 0.923:
- For times when predictor variables are fixed:

$$\hat{R}^2_{cv} = 1 - \left(\frac{8-1}{8}\right)\left(\frac{8+2+1}{8-2-1}\right)(1-0.923) = 0.852$$

- For times when predictor variables are random:

$$\hat{R}^2_{cv} = 1 - \left(\frac{8-1}{8-2-1}\right)\left(\frac{8-2}{8-2-2}\right)\left(\frac{8+1}{8}\right)(1-0.923)$$

$$\hat{R}^2_{cv} = 0.818$$

Cross-Validation Via Jackknifing

- Jackknifing is a computationally intense method for producing an estimate of an adjusted R^2 value.
- For a sample of size N, the Jackknife procedure fits a total of N regression models:

1. Remove an observation from the sample.
2. Fit a multiple regression with remaining $N-1$ observations.
3. Estimate the Y' for the removed observation.
4. Repeat steps 1-3 for all N.
5. Calculate new R^2 by using squared correlation between Y and Y'.

Jackknifing Example

- To demonstrate, I used the Athens cake data.
- I ran eight different multiple regression, each with a different observation removed.
- I then computed the squared correlation coefficient between Y and Y'.
- The estimated R^2 from the jackknife procedure was 0.859, which was lower than the original R^2 of 0.923.

Bootstrapping

- Bootstrapping is a computationally intensive procedure designed to produce a confidence interval around R^2 (or any other model parameter).

- The bootstrap method fits a great number of multiple regression models on randomly drawn samples from the original sample:

  1. From the original sample of size N, draw (with replacement) a new sample of size N.
  2. Fit a multiple regression.
  3. Record the R^2 for the model.
  4. Repeat steps 1-3 a large number of times.
  5. Calculate new R^2 and confidence interval from bootstrap of R^2 values.

Predictor Selection

- Often times, one has a lot of predictor variables, many of which do not add to the predictive ability of the regression model.
- In such cases, one may want to make a model comprised of a subset of these variables, so that each remaining variable in the model has adds to the predictive ability of the model.
- Several methods exist to search for such a subset, but be warned: these should only be used for predictive research because of their exploratory nature.
  - All possible regressions.
  - Forward selection.
  - Backward elimination.
  - Stepwise selection.

Erie, Pennsylvania

We have data for 27 houses sold in the mid 1970's in Erie, Pennsylvania:

- X_1: Current taxes (local, school, and county) $\div$ 100 (dollars).
- X_2: Number of bathrooms.
- X_3: Lot size $\div$ 1000 (square feet).
- X_4: Living space $\div$ 1000 (square feet).
- X_5: Number of garage spaces.
- X_6: Number of rooms.
- X_7: Number of bedrooms.
- X_8: Age of house (years).
- X_9: Number of fireplaces.
- Y: Actual sale price $\div$ 1000 (dollars).

Erie, Pennsylvania

Recall what happened when we used all nine X variables to predict Y:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.972[a]	.945	.915	4.1612

a. Predictors: (Constant), number of fireplaces, age of house, number of bedrooms, number of garage spaces, lot size, number of bathrooms, current taxes, living space, number of rooms

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5027.951	9	558.661	32.263	.000[a]
	Residual	294.369	17	17.316		
	Total	5322.320	26			

a. Predictors: (Constant), number of fireplaces, age of house, number of bedrooms, number of garage spaces, lot size, number of bathrooms, current taxes, living space, number of rooms

b. Dependent Variable: sale price in thousands

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	6.076	7.239		.839	.413
	current taxes	1.235	.790	.249	1.564	.136
	living space	13.473	4.611	.526	2.922	.010
	number of bathrooms	7.315	5.875	.217	1.245	.230
	lot size	.190	.562	.032	.339	.739
	number of garage spaces	1.179	1.888	.050	.624	.541
	number of rooms	-.798	2.419	-.065	-.330	.745
	number of bedrooms	-.627	3.631	-.032	-.173	.865
	age of house	-.066	.085	-.065	-.771	.451
	number of fireplaces	2.184	2.416	.073	.904	.379

a. Dependent Variable: sale price in thousands

All Possible Regressions

- The most computationally expensive method for finding the best subset of variables in a predictive sense is the method of all possible regressions.
- For a regression with k possible predictor variables, a total of 2^k possible regressions can be fit.
- In our Erie housing price example, nine variables were used.
- If we were to use all possible combinations, we would have to estimate $2^9 = 512$ different regression equations.
- This method is not too practical unless you can program an iterative algorithm.

Forward Selection

- Forward selection is a variable selection procedure that successively adds a single variable to a regression equation per step.

- For this, consider a data set with k possible predictor variables.

  1. Find the X with the highest squared correlation with Y.
  2. Fit $k - 1$ different regressions using each of the remaining variables separately.
  3. Record the incremental change in R^2 (and the corresponding hypothesis test p-value).
  4. Using some criterion to decide the minimal level of the p-value to allow a new variable into the model, select the variable with the highest R^2 meeting this criterion.
  5. Stop when no variable satisfies this criterion.

Forward Selection Example

Criterion for inclusion: p-value must be less than 0.05.

1. Begin with a model with the highest squared correlation between Y and X:

   All variables:

Variable	ΔR^2	F	p-value
X_1	0.838	129.037	0.000
X_2	0.854	145.675	0.000
X_3	0.528	27.917	0.000
X_4	0.863	157.108	0.000
X_5	0.204	6.407	0.018
X_6	0.599	37.327	0.000
X_7	0.488	23.804	0.000
X_8	0.096	2.663	0.115
X_9	0.232	7.535	0.011

Forward Selection Example

Criterion for inclusion: p-value must be less than 0.05.

1. Begin with a model with the highest squared correlation between Y and X:

   All variables meeting criterion:

Variable	ΔR^2	F	p-value
X_1	0.838	129.037	0.000
X_2	0.854	145.675	0.000
X_3	0.528	27.917	0.000
X_4	0.863	157.108	0.000
X_5	0.204	6.407	0.018
X_6	0.599	37.327	0.000
X_7	0.488	23.804	0.000
X_9	0.232	7.535	0.011

Forward Selection Example

Criterion for inclusion: p-value must be less than 0.05.

1. Begin with a model with the highest squared correlation between Y and X:

   All variables meeting criterion:

Variable	ΔR^2	F	p-value
X_1	0.838	129.037	0.000
X_2	0.854	145.675	0.000
X_3	0.528	27.917	0.000
X_4	0.863	157.108	0.000
X_5	0.204	6.407	0.018
X_6	0.599	37.327	0.000
X_7	0.488	23.804	0.000
X_9	0.232	7.535	0.011

   Here we select X_4 because of all the variables meeting our criterion, it has the highest R^2.

Forward Selection Example

Criterion for inclusion: p-value must be less than 0.05.

2. Now fit all regression models with X_4 and another X.

All variables:

Variables	ΔR^2	F	p-value
X_4 X_1	0.066	22.069	0.000
X_4 X_2	0.040	9.999	0.004
X_4 X_3	0.011	2.004	0.170
X_4 X_5	0.008	1.576	0.221
X_4 X_6	0.000	0.083	0.776
X_4 X_7	0.004	0.724	0.403
X_4 X_8	0.022	4.502	0.044
X_4 X_9	0.021	4.252	0.050

Forward Selection Example

Criterion for inclusion: p-value must be less than 0.05.

2. Now fit all regression models with X_4 and another X.

   All variables meeting criterion:

Variables	ΔR^2	F	p-value
X_4 X_1	0.066	22.069	0.000
X_4 X_2	0.040	9.999	0.004
X_4 X_8	0.022	4.502	0.044

Forward Selection Example

Criterion for inclusion: p-value must be less than 0.05.

2. Now fit all regression models with X_4 and another X.

All variables meeting criterion:

Variables	ΔR^2	F	p-value
X_4 X_1	0.066	22.069	0.000
X_4 X_2	0.040	9.999	0.004
X_4 X_8	0.022	4.502	0.044

Here we select X_1 because of all the variables meeting our criterion, it has the highest R^2.

Forward Selection Example

Criterion for inclusion: p-value must be less than 0.05.

2. Now fit all regression models with X_4 and X_1 and another X.

All variables:

Variables	ΔR^2	F	p-value
X_4 X_1 X_2	0.006	2.132	0.158
X_4 X_1 X_3	0.001	0.478	0.496
X_4 X_1 X_5	0.000	0.014	0.906
X_4 X_1 X_6	0.003	0.947	0.341
X_4 X_1 X_7	0.003	0.877	0.359
X_4 X_1 X_8	0.002	0.717	0.406
X_4 X_1 X_9	0.004	1.252	0.275

Forward Selection Example

Criterion for inclusion: p-value must be less than 0.05.

2. Now fit all regression models with X_4 and X_1 and another X.

All variables meeting criterion:

Variables	ΔR^2 F p-value

Forward Selection Example

Criterion for inclusion: p-value must be less than 0.05.

2. Now fit all regression models with X_4 and X_1 and another X.

All variables meeting criterion:

Variables	ΔR^2 F p-value

Because no other variables meet our criterion, we conclude the forward selection process.

Forward Selection in SPSS

- Go to Analyze...Regression...Linear.
- Select Y and then place all variables into the independents box.
- Select “Forward” from the Method pull-down menu.
- Set p-value or F-value criteria under Options box.

Backward Selection

- Unlike forward selection, backward selection begins by estimating the R^2 for the model with all predictor variables included.

- Using a criterion for removal (as a default, SPSS uses the p-value greater than 0.10), the variable meeting the criterion with the smallest change in R^2 is removed.

- If no variable meets the removal criterion, the backward selection process stops.

Backward Selection in SPSS

- Go to Analyze...Regression...Linear.
- Select Y and then place all variables into the independents box.
- Select "Backward" from the Method pull-down menu.
- Set p-value or F-value criteria under Options box.

Stepwise Selection

- Stepwise selection uses both forward and backward selection to arrive at a subset of variables that best predict Y.

- Each step consists of a forward step, followed by a backward step.

  1. Select the variable meeting the forward selection criterion with the highest R^2.

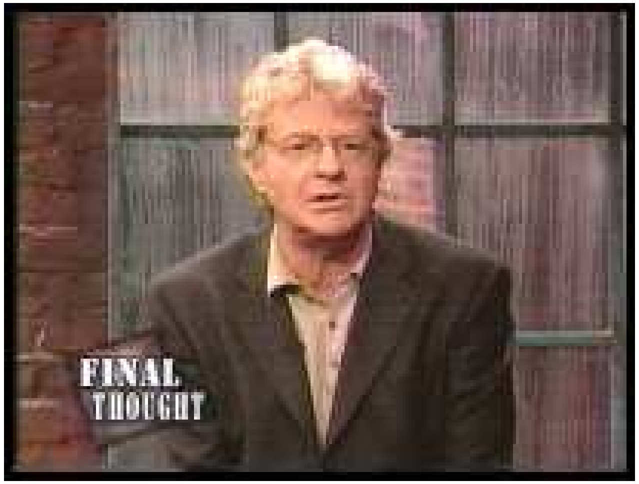
  2. Go through all variables and remove the variable with the smallest change in R^2 that meets the backward selection criterion.

Stepwise Selection in SPSS

- Go to Analyze...Regression...Linear.
- Select Y and then place all variables into the independents box.
- Select "Stepwise" from the Method pull-down menu.
- Set p-value or F-value criteria under Options box.

Final Thought

- R^2 values give information about how well a regression model fits, but are typically inflated.
- Be careful to generalize results



- Be really careful when using variable selection methods, generality is really hard to argue for.

Next Time

- Variance partitioning.
- Chapter 9.