Session 4
Issues and Extensions

A. Guidelines to avoid over-fitting
B. Lack of convergence in logistic regression
C. More general CCR models
    1. CCR-Survival/Event history model
    2. Hybrid CCR/Latent Class models
        i. Example: Key Driver Regression on orange juice ratings data
D. Handling missing data

A. Guidelines to Avoid Overfitting

Overfitting data results in incorrect inferences and invalid predictions. It is not possible to avoid overfitting completely, but the expected amount of overfitting can be reduced by applying regularization or structure to models which take into account apriori knowledge about the relationships to be modeled and using cross-validation techniques to determine the most parsimonious model consistent with the data. This will reduce the amount of error variance and improve prediction.

Two important examples of imposing such structure are:

1. Exclude predictors that are known to be irrelevant. If insufficient knowledge exists to rule out irrelevant predictors, utilize statistical methods such as the CCR/ stepdown algorithm that is guided by cross-validation techniques to eliminate irrelevant predictors. Because it is important to retain suppressor variables among the predictors, variable selection methods that select only variables that have a significant (direct) relationship with the dependent variable should be avoided.

2. Include covariates in a model that are known to be important. This can also reduce the amount of irrelevant variance.

Regarding parsimony, in the case where the number of predictors is large relative to the number of cases, it is necessary to impose regularization. The method of regularization that we focused on primarily in this course is to use a K-component model with a relatively small value for K (say 2-6).

Consider the following example application:

Suppose that survey respondents are asked to provide ratings of purchase intent for several products and also rate each product on many product attributes, which tend to be moderately correlated with each other. Some of these attributes may be irrelevant with respect to purchase intent for a given product, and these should therefore be excluded from the model. Moreover, if additional variables (such as demographic or attitudinal variables) are available that are also predictive of purchase intent, they can be included as additional predictors in a CCR analysis. The smaller the sample size, and the larger the number of predictors relative to the sample size, the more important it is to impose regularization to avoid extensive amounts of overfitting.

An example of this type of application is considered later in Example C1 below.

B. Resolving problem of complete or quasi complete separation in logistic regression

Occasionally logistic regression algorithms will not converge due to complete or quasi-complete separation between the 2 dependent variable groups. For example, with a single predictor X such that all cases in group 1 have higher X values than all cases in group 2, any regression coefficient for X will provide perfect separation of the 2 groups, and the maximum likelihood estimate for the X coefficient does not exist (it approaches infinity). Similar convergence problems may occur with near perfect separation.

To prevent such convergence problems, the CCR.logistic algorithm in COREXpress and XLSTAT-CCR contains 2 estimation parameters -- the # iterations and the Ridge parameter. By default, the number of iterations is set to 4, which should be sufficient for most applications. In addition, a ridge regression penalty is included, with default equal to .001. With no penalty (Ridge parameter = 0), the separation problems may cause nonconvergence, in which case increasing the number of iterations will yield larger and larger estimates for at least one regression coefficient. Using a sufficiently large penalty eliminates this non-convergence problem. Typically, the default parameter (.001) will be sufficiently large to prevent the non-convergence problem from occurring.

C. More General CCR Models

Correlated Component Regression (CCR) can be extended in several ways. Thus far, we have considered dichotomous and continuous scale types for the dependent variable. Other scale types are also possible.

For example, time-to-event data can be analyzed to determine predictors associated with an event. The event might be attrition (customers leave), death (survival analysis), or some other occurrence. Such analyses can be viewed as a type of 'conditional' logistic regression (see e.g., Vermunt, 2009), where each case contains multiple records, one record for each time point, the final record being the occurrence of the event (event=1) or nonoccurrence (event=0). An example of this is given below based on the CCR.surv scale type option.

When the logistic regression or LDA option is selected in CORExpress or XLSTAT-CCR, the CV-Accuracy is the primary criterion for determining the number of predictors, and in the case of ties, the CV-AUC is used as a secondary criterion. When the survival scale type is used (not available in XLSTAT-CCR), the priority of these two CV criterion are reversed: for the step-down selection, the primary criterion is CV-AUC. The reason for this change is that unlike logistic regression and LDA where estimation of the intercept (and thus the associated cutpoint) along with the regression coefficients are obtained using a common criterion, for survival analysis there is no criterion for estimating the intercept. Therefore, the CV-AUC, which is intercept invariant, provides a more appropriate criterion than CV-Accuracy.

Another extension consists of a categorical dependent variable with more than 2 categories, or there may be multiple dependent variables. Such extensions are planned to be incorporated in future versions of CORExpress.

Main-effects-only models have been considered thus far. For example, in the case of linear regression, no interaction terms were included in the models. One extension to include interactions is to use a 2-step modeling strategy. In Step 1, a CCR-based linear model is developed, where P* predictors are retained. In Step 2, the predicted scores are included in the model as a single predictor variable, and all possible 2-way interaction variables based on these P* predictors are created as additional candidate predictors. An extended CCR model is then estimated, where the interaction terms are allowed to enter into the model. A variation of this extension is to include each of the P* predictors in the model along with the candidate predictor interactions as a starting point. The P* predictors might be forced to be retained in the final model.

Another alternative is to use COREXpress or XLSTAT-CCR simply to select important variables, and then perform some other analysis using these predictors. For example, if it is suspected that the coefficients in a key driver regression analysis are heterogeneous, the homogeneous CCR model obtained from COREXpress or XLSTAT-CCR can be viewed as an initial model, which simply identifies the important predictors. This model can then be further refined by estimating a latent class regression model based on the selected P* predictor variables.

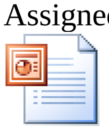
More specifically, it is frequently the case that the effects of a particular driver depend upon different respondent segments. Therefore, latent class models can be used to obtain separate effects for each segment, and to obtain individual level coefficients for each driver for each respondent. LC regression models of this type are often developed based on multiple records per respondent. For example, the dependent variable may be a rating of product satisfaction or purchase intent where respondents rate each of several products.

Alternatively, a latent class cluster model based on one or more dependent variables (indicators) could be estimated initially, and the resulting classes can be predicted as a function of many demographic, attitudinal, or other predictor variables using CCR. In this type of extension, posterior membership probabilities obtained from the latent class analysis may be used as weights (e.g., see Magidson, 2005). The weights themselves may be case weights, so that for example, OLS regression is replaced by weighted least squares (WLS), or replication weights might be used, where the case ID is used as a primary sampling unit.

Example C1: Hybrid CCR/Latent Class models

Key Driver Regression on orange juice ratings data

Assigned reading:


'Session 4 Hybrid.pptx'

Tutorial:

For CORExpress Users:

Tutorial: 'COREtutorial4.pdf'

Data: 'OJtutorial2lc.sav'

For XLSTAT-CCR Users:

Tutorial: 'XLCCRtutorial3.pdf'

Data: 'OJtutorial.xls'

D. Handling Missing Data

In regression analysis, missing data is generally handled using the 'list-wise deletion' option, which eliminates any case that has a missing value on any predictor. This option is used primarily when there are only a few cases (or say 1% of the cases) that have missing values. At the other extreme, missing values on a predictor are replaced by the mean of the predictor. This imputation approach is typically used with continuous predictors, where more than a few cases have missing values. The imputation approach biases coefficients towards zero, which is not a bad idea with high dimensional data, since it is consistent with regularization strategies.

A more sophisticated option for dealing with missing data is multiple imputation, where say 10 data files are created, each with the missing values filled in with different imputed values. The imputed values are determined based on distributional assumptions. For example, a common distributional assumption for continuous predictors is to assume that they follow a multivariate normal distribution. In this case, the imputed values would tend to preserve the means, variances, and correlations among the predictors.

Consider the following Auto Price example (source: Michel Tenenhaus, Professor Emeritus HEC Paris):

- N=24 car models
- Dependent variable: Y = PRICE (car price measured in francs)
- 6 Predictor Variables:
  - X_1 = CYLINDER (engine measured in cubic centimeters)
  - X_2 = POWER (horsepower)
  - X_3 = SPEED (top speed in kilometers/hour):
  - X_4 = WEIGHT (kilograms)
  - X_5 = LENGTH (centimeters)
  - X_6 = WIDTH (centimeters)

Table 1: autoprixb data

Model	Price	Cylinder	Power	Speed	Weight	Length	Width
honda civic	83700	.	90	174	850	369	166
renault 19	83800	1721	.	180	965	415	169
fiat tipo	70100	1580	83	.	970	395	170
peugeot 405	119800	1769	90	180	.	440	169
renault 21	113400	2068	88	180	1135	.	170
citroen bx	106600	1769	90	182	1060	424	.
bmw 530i	226600	.	188	226	1510	472	175
rover 827i	175000	2675	.	222	1365	469	175
renault 25	207200	2548	182	.	1350	471	180
opel omega	118350	1998	122	190	.	473	177
peugeot 405b	114800	1905	125	194	1120	.	171
ford sierra	109700	1993	115	185	1190	451	.
bmw 325ix	205400	.	171	208	1300	432	164
audi 90 quattro	242800	1994	.	214	1220	439	169
ford scorpio	196900	2933	150	.	1345	466	176
renault espace	183000	1995	120	177	.	436	177
nissan vanette	106900	1952	87	144	1430	.	169
vw caravelle	140900	2109	112	149	1320	457	.
ford fiesta	64500	.	50	135	810	371	162
fiat uno	58900	1116	.	145	780	364	155
peugeot 205	73800	1580	80	.	880	370	156
peugeot 205 rallye	71400	1294	103	189	.	370	157
seat ibiza s	71950	1461	100	181	925	.	161
citroen ax sport	66800	1294	95	184	730	350	.

For these data, list-wise deletion is not possible because all cases would be omitted.

In CORExpress or XLSTAT-CCR, if the ‘Include Missing’ option is selected (under ‘Options’ in the Model Control window for CORExpress and in the ‘Missing Data’ tab is XLSTAT-CCR), the mean value of the predictors is automatically imputed when a missing value is observed. If the ‘Include Missing’ option is not selected, the list-wise deletion option is used.

Exercise D1:

For COREXpress Users:

Data: ‘autoprixb.sav’

For XLSTAT-CCR Users:

Data: ‘autoprixb.xls’

Open the Auto Price data and estimate a CCR.lm model using the step-down option (include all 6 predictors) and 10 rounds with 6 folds for tuning parameters. Be sure to check the “Include Missing” box.

1. Is a 1-component or a 2 –component model best?
2. Which are the most important predictors?
3. Are any predictors excluded from the model?

References

Magidson, Jay (2005) “An Extension of the CHAID Tree-based Segmentation Algorithm to Multiple Dependent Variables”, in C. Weihs & W. Gaul, *Classification: The Ubiquitous Challenge*, 176-183. Heidelberg: Springer.
See: http://statisticalinnovations.com/products/8pagearticle.pdf

Vermunt, JK. (2009): Event history analysis. In: R. Millsap and A. Maydeu-Olivares (eds.) *Handbook of Quantitative Methods in Psychology*, 658-674. London: Sage.
See: http://spitswww.uvt.nl/~vermunt/hqmp2007.pdf