Topic 3. Linear Regression

Case 3: Donor Recapture

using Transaction, Overlay, and Census Data

Reading Assignment

Berry and Linoff (2000)

- Pages 131–168. Data preparation.
- Pages 81–82, 186–191. Lift.

Main Points (aka Takeaways)

- Tool main points
  - Each tool has its own main points that are independent of the cases.
  - Points that relate to a tool will emerge and be illustrated within the analysis of a case.
- Case main points
  - Each case has its own main points that are independent of the tools.
  - Points that relate to a case are itemized at the end of each case subtopic.

Tool Main Points: Regression

- The oldest method (Gauss, 1816).
- Makes efficient use of space and time.
- Works very well when features are well chosen.
- In large samples F-tests and similar measures computed within the learning sample should only be used to rank features, not select them.
- Validation should be used for model selection.

Case Subtopics

- Machine assisted data preparation and cleaning.
- Machine assisted feature selection
- Machine assisted model building
- Model assessment and lift charts.

Machine Assisted

The focus here is "machine assisted."

The usual situation in data mining is a large number of features and cases and the Donor Recapture case is realistic in that respect. One simply must use the machine to automate the analysis.

Regression has a huge advantage for these tasks because it is fast and conserves memory.

Data Preparation

The first subtopic is machine assisted data preparation

This is what needs to be done:

1. Describe Case 3: Donor recapture.
2. Describe the data
3. Describe a strategy for automated reading and cleaning.

Details follow:

Case 3: Donor Recapture

- The data are from a not-for-profit organization, called CTY hereafter.
  - CTY provides programs and services to a specific group with a specific injury.
  - CTY is one of the largest direct mail fund raisers in the country.
  - CTY's in-house database covers 13 million donors.
- The data are proprietary.
  - These data are not to be used outside of this course.

Donor Recapture Data

- The data are lapsed donors from CTY's database.
  - 95412 cases.
  - 481 features describing
    - the donors
    - the solicitations they have received
    - the monetary response to the most recent mail solicitation
- The target is the monetary response to the most recent mailing.
- The model developed will be used to determine which lapsed donors will receive future solicitations.

Project Description

- A complete description of the project is file **cty_doc.txt**
- A complete description of the features is file **cty_dic.txt**
- Recall that you can access these files at

  http://www.econ.duke.edu/∼ arg/datamine

  click on "Cases" then on "Charity"

A Haphazard Sample of the Features

Field	Feature	Type	Definition
1	ODATEDW	Num	Date of donor's first gift
5	ZIP	Char	Seven digit Zipcode
8	DOB	Num	Date of birth
14	MDMAUD	Char	Encodes frequency and amount of giving
15	DOMAIN	Char	Socio-economic status of donor's neighborhood
24	NUMCHLD	Num	Number of children
25	INCOME	Num	Household income
27	WEALTH1	Num	Wealth rating
28	HIT	Num	Responses to mail offers other than CTY's
35	MAGMALE	Num	Buys sports magazines
43	DATASRCE	Char	Source of overlay data (MetroMail,Polk,Both)
49	STATEGOV	Num	Employed by State Gov
53	MAJOR	Char	Major donor flag
58	BIBLE	Char	Interest in bible reading
64	PCOWNERS	Char	PC owner
71	KIDSTUFF	Char	Buys children's products
85	ETH2	Num	Percent Black in donor's neighborhood
327	ANC15	Num	Percent Ukranian ancestry in neighborhood
362	ADATE_2	Num	Date the 97NK promotion was mailed
363	ADATE_3	Num	Date the 96NK promotion was mailed
385	RFA_2	Char	Recency/frequency/amount status as of 97NK
413	RDATE_3	Num	Date gift received for 96NK
435	RAMNT_3	Num	Dollar amount of the gift for 96NK
471	TARGET_B	Num	Responded to 97NK mailing
472	TARGET_D	Num	Donation amount associated with 97NK mailing

Comments on cty_doc.txt

It is common in data mining, as in advertising, to choose a vendor by a competition. The vendors are given two sets of randomly selected data, in one the target is missing. They build a model from the data that has a target and predict the target in the one that doesn't. The winner is the vendor who best predicts the missing target.

Marketers often use RFA features – recency, frequency, and amount – as predictors. These are in the data as are many RFA features derived from them.

The task is to maximize revenue. A common approach is to predict the a binary response target and then fit another model within respondents to predict the amount. There is a warning in the document that this will not work for these data.

Comments on cty_doc.txt

There is also advice to subtract ADATE_2, the date of the mailing, from all dates. But this date is June 1997 for all but 0.07% of the cases and for these it was May 1997. I didn't bother.

The really bad news is the response rate: 5.1%
The lower the response rate, the harder it is to build a good model. As we shall see, this is a really hard problem.

Only the Paranoid Survive

Andrew S. Grove, 1996

- The Documentation:
  The data are in comma delimited format. The first row of the data set contains the field names. Blanks in the character variables and periods in the numeric variables correspond to missing values.

- The Actuality:
  Blanks in the character variables sometimes do not represent missing values but are actually valid data. There are no periods in the numeric variables; there are null fields instead. Feature 481 (GEOCODE2) contains non-printable characters (Ctrl-@) and Feature 5 (ZIP) sometimes contains a trailing minus sign.

Flat Files

The data are supplied as a flat file with a data dictionary – **cty_dic.txt**. "Flat file" means a two by two table with features as columns and cases as rows.

This is standard. Nearly all tools expect a flat file. If the data does not come as flat file, the first task of the data mining team is to create one.

The Bad News

These data are hard to read with a lot of ambiguity as to what is or isn't a missing value.

The data set is so large that it will overwhelm most software; SAS is the main exception.

Unfortunately, the way SAS handles missing values is exasperating, especially because, in these data, what is and is not a missing value is ambiguous.

I, at least, do not have the patience to deal them within SAS.

What to Do?

Divide and conquer: Break the data up into one file for each feature using a strategy that is memory efficient.

Reassemble only the features necessary to an analysis into flat files and deal with the missing value problem at the time of reassembly.

For disassembly I used the scripting features of the Borne shell in Unix, a few Unix utilities, and the C++ programming language in combination.

Berry and Linoff p. 57: "... the full power of a programming language is often needed ... ".

Some Comments on C++

- It is a well documented, standardized language (ISO 14882). One knows what the code will do in any circumstance, which is not true of statistical packages.
- Its library must contain implementations of the standard data structures from computer science, associative maps in particular.
- The language allows one to get as close to the machine architecture as circumstances require.
- C++ is available on all platforms: Windows, Linux, Sun OS, HP UX, IBM Aix, etc.
- This is my standard way of merging and cleaning data.

Step One: Partition the Data

The first step is to randomly divide the data into learning, validation, and test data sets.

It is not actually necessary to partition first, but it gives one smaller data sets to deal with, is easy to do, and puts off the evil hour when I really have to face up to the task of reading the data.

For this I used C++. When finished I have three new files – **cty_lrn.dat**, **cty_val.dat**, and **cty_tst.dat** – each identical in format to the original data. They are 70%, 20%, and 10% of the original data, respectively.

File cty_lrn.dat

cty_lrn.dat looks like this

```
ODATEDW,OSOURCE,TCODE,STATE,ZIP,MAILCODE,CTYSTATE,DOB,NOEXCH,RECINHSE, etc
8901,GRI,0,IL,61081, , ,3712,0, etc.
9401,BOA,1,CA,91326, , ,5202,0, etc.
9001,AMH,1,NC,27017, , ,0,0, etc.
8701,BRY,0,CA,95953, , ,2801,0, etc.
8601, ,0,FL,33176, , ,2001,0,X, etc.
9401,CWR,0,AL,35603, , ,0,0, etc.
8701,DRK,0,IN,46755, , ,6001,0, etc.
9401,NWN,0,LA,70611, , ,0,0, etc.
8801,LIS,1,IA,51033, , ,0,0, etc.
9401,MSD,1,TN,37127-, , ,3211,0, etc.
9601,AGR,0,KS,67335, , ,0,0, etc.
8901,ENQ,0,MN,56475, , ,2603,0, etc.
9201,HCC,1,LA,70791, , ,0,0,X, etc.
9301,USB,1,UT,84720, , ,2709,0, etc.
9401,RKB,0,MI,48067, , ,5401,0, etc.
8801,PCH,2,IL,62376, , ,5201,0, etc.
etc.
```

cty_lrn.dat was constructed using C++ from **cty_raw.dat**

Step Two, Divide and Conquer

The data are in comma delimited format, which make it very easy to peal off the features column by column using Unix utilities.

There is a convenient listing of the features in the middle of **cty_doc.txt**.

While looping over the features some summary files are produced for each feature to try and help make sense of the data, especially to try to figure out how much data are missing in each feature.

Here is a sample set of output files.

Output File chr/15.dat

Column 15, which is DOMAIN, generates file **15.dat**, which looks like this:

```
DOMAIN
T2
S1
R2
R2
S2
T2
T2
T2
R2
T1
R3
etc.
```

That is, **chr/15.dat** has the name of the feature as the first row and the data in succeeding rows.

Output File chr/15.frq

Column 15, which is DOMAIN, generates file **15.frq**, which looks like this:

```
DOMAIN
    1638
C1  4271
C2  5791
C3  3680
R1  962
R2  9541
R3  3459
S1  8033
S2  6006
S3  1327
T1  3461
etc.
```

That is, **chr/15.frq** has the name of the feature as the first row and a frequency count of every string that occurs in column 15.

Output File chr/15.max

Column 15, which is DOMAIN, generates file **15.max**, which looks like this:

```
DOMAIN
Null:        0
Missing:     1638
Minimum: R1  962
Maximum: R2  9541
```

That is, **chr/15.max** has the name of the feature as the first row and a frequency count of null strings, blank strings, the string that occurred the minimum number of times, and the string that occurred the maximum number of times.

Step Two: Continued

We do exactly the same for the numeric fields.

As yet I have not taken any chances because the numeric data have been manipulated as character data.

Data for Student Use

A randomly chosen subset of the data of a size small enough for use with XL-Miner is in directory (folder) **sub/num** for numeric features and **sub/chr** for categorical features.

A file such as **sub/num/25.dat** is the data for feature 25, INCOME, of the preceding list of features.

The corresponding file **sub/num/25.frq** lists the values taken on by the feature and their frequency. The file **sub/num/25.max** is the same but only for the values maximum, minimum, blank, null, and missing.

These are clean data. They are also over sampled. The frequency of a positive target is 0.0507588 in the raw data and 0.1886 in these data.

Recall

- A complete description of the project is file **cty_doc.txt**
- A complete description of the features is file **cty_dic.txt**
- You can access these files at

  http://www.econ.duke.edu/∼ arg/datamine

  click on "Cases" then on "Charity"

Subtopic Main Points
Machine Assisted Data Preparation

Data preparation and cleaning is a one-off *ad hoc* process. Each project presents its own idiosyncratic difficulties.

Machine assistance is usually essential.

Statistical packages are often unable to deal with data as it comes.

We resorted to low level machine code to deal with the task, which is not atypical,

In our case, the machine assistance was remarkably effective: Compact code dealt efficiently with a potentially very messy task.

Feature Selection

The next subtopic is machine assisted feature selection.

We shall consider feature selection in this order:

1. Numeric features

2. Character features

There is no significance to the ordering; details follow.

Which Features are Important?

As we have seen previously, selection of the right (derived) features is the key to successful data mining: Even simple linear methods work well with good features.

Domain knowledge and experience are extremely useful. For these data, it seems everything relevant is present.

Graphs and multi-way tables (slice and dice) can be helpful. I invite you to play with the data using the graphical and slice and dice tools in Excel to see if you can find derived features I missed. There are two data sets of interest to you: **charity/sub** and **charity/don**. The first is the oversampled subset of the data described earlier. The second are all cases that donated. Both are small enough to fit in XLMiner.

Automatic Feature Selection

There are limits to what graphics and slice and dice can accomplish, especially if one is overwhelmed by features, as we are here.

One common approach is to use decision trees to select features. I do not like that approach because it is too much of a *fait accompli* and is too sensitive to tuning parameters.

I prefer something that gives a menu of reasonable suggestions. Regression does this and is probably the most commonly used tool for automatic feature selection.

Regression Feature Detection

One approach is to regress the target on each feature in the data and set aside for further consideration those features that have a highly significant F-statistic.

This might not ordinarily be the best approach but the data sets involved have different sizes due to missing values. The F-statistic can cope with comparisons among data sets of different sizes better than R^2, MSE, or other measures one might think of.

I actually did try learning MSE and validation MSE without much success: I had trouble interpreting the results.

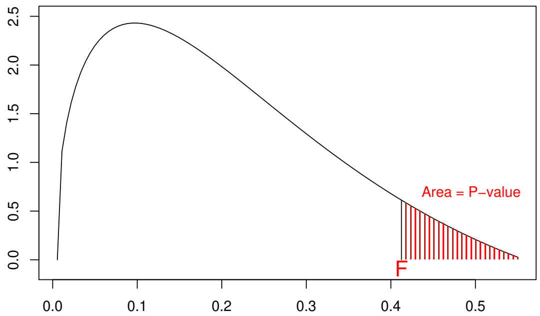
We explore the regression approach next.

The F-statistic in Regression Analysis

- Minimize: $\text{RSS}(b) = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip})^2$
- Solution: $\hat{b} = (\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_p)$
- F-statistic: $\hat{F} = \frac{\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 - \text{RSS}(\hat{b})\right]/(p-1)}{\text{RSS}(\hat{b})/(n-p-1)}$

where $\bar{y}$ is the average value of the target.

Fig 51. Interpretation of the F-statistic



F is the computed F-statistic. The black curve is the density of F assuming that there is no regression, i.e. $b_1 = \ldots = b_p = 0$. The probability of a value as large or larger than F is the shaded area, called the P-value. If there is no regression but P is small, then an unusual event has occurred. If there is a regression, then the density shifts to the right and the event is not unusual. Thus, the smaller is P the stronger the evidence that there is a regression. F-statistics with different p or n cannot be compared, but their P-values can.

Caveats Regarding the F-statistic

- The interpretation of the F-statistic depends on assumptions, the most critical of which is that the model is exactly correct.
- In large data sets, violations of this assumption can lead to a conclusion that there is a regression when that conclusion is misleading.
- Because data sets in data mining are large by definition, the F-statistic is unreliable as a model selection device.
- This is one of the reasons for using validation samples in data mining rather than formal statistical tests.
- Nonetheless, the P-value associated with the F-statistic is a useful device for a preliminary ranking of regression models.
- To summarize, P-values are a useful preliminary model ranking device but a poor model selection device.

Some Other Methods

Mallow's C_p, AIC, BIC, etc. try to take model complexity into account.

Most interesting is Mallow's C_p which tries to estimate σ_{ef} and add it back in to the computed MSE.

Other than the sample size issue mentioned previously, it doesn't matter much if these other measures are used for feature detection if we only rely on them to rank features. Their ranking will not differ too much from an F-test ranking.

Like the F-test, these methods are not robust to violations of the assumptions that justify them and should not be used for model selection. Validation samples should be used instead.

More Intensive Methods

Some tools also try all possible regressions rather than look at features one at at time as we will. When feasible, this approach has a slight chance of beating a strategy of adding variables one at a time according to incremental merit that we shall use later.

A more interesting approach would be some attempt to find derived features automatically. An example would be to try all possible pairs with squares and cross products added in. As far as I know, no tools do this.

Numeric Variables

Running a regression for a truly numerical feature is straightforward. Simply fit the model

$$y_i = b_0 + b_1 x_i$$

where x_i is the numerical value of the feature for Case i.

Three approaches to missing data:

1. Delete cases with missing values.
2. Treat x_i as a target and try to fill in its values as a prediction from a regression on other features.
3. Fill in the value in some other reasonable way; i.e. the average value of the feature or a value smaller than all other values of the feature.

We shall use Option 1, deletion. Results follow.

Numeric Variables: Some Significant Results

charity/lrn/num/num_reg.txt

File	Feature	Type	Levels	Missing	P-Value	
146	HV1	num	4181	0	2.2e-16 ***	Median home value
147	HV2	num	4356	0	2.2e-16 ***	Average home value
148	HV3	num	13	0	2.524e-16 ***	Median rent
149	HV4	num	13	0	2.2e-16 ***	Average rent
199	IC1	num	1087	0	2.2e-16 ***	Med hshld income
200	IC2	num	1166	0	2.2e-16 ***	Med family income
201	IC3	num	1054	0	2.2e-16 ***	Ave hshld income
202	IC4	num	1121	0	2.2e-16 ***	Ave family income
203	IC5	num	19511	0	2.2e-16 ***	Per capita income
224	HHAS3	num	98	0	9.136e-15 ***	% Int,div,rent inco
297	EC8	num	69	0	2.2e-16 ***	% 25+ /w grad deg
457	RAMNTALL	num	1761	0	2.2e-16 ***	$ tot all gifts
462	MAXRAMNT	num	224	0	2.2e-16 ***	$ largest gift
469	AVGGIFT	num	6464	0	2.2e-16 ***	$ ave all gifts

Numeric Variables: Conclusions

A careful study of file **num_reg.txt** leads to the following conclusions:

- Wealth and income measures are the best predictors of the target.
- Wealth and income measures from the census have no missing values.
- Measures of previous amount and frequency of previous gifting from the transactions data are good predictors.
- Some of these are aggregates with no missing values and seem to be as good as disaggregated measures that have missing values.

Character Variables

Character variables are more trouble. A character variable is a categorical feature and the categories must be converted to a set of derived numerical features called dummy variables.

There are several problems with this approach:

- If the number of categories is large, the number of dummy variables required is so large as to exceed memory limits. This can often be fixed by combining categories in a reasonable way. E.g. combine states into Atlantic, Mid-Atlantic, South, Midwest, etc.
- Are missing values truly missing or are they valid data requiring their own dummy variable? Our solution will be to run the regression both ways and thereby try to get the machine to provide the answer.
- Dummy variables cause exact multicolinearity. SAS, R, and many other statistical packages cope with this gracefully. Others go berserk which requires the user to hand code dummies in an unnatural fashion.

We will illustrate these issues by example.

RFA: Recency, Frequency, Amount

476.frq

```
RFA_2A
D  5138
E  15275
F  32863
G  13624
```

RFA_2A is the third byte of RFA_2 which is status as of the 97NK mailing

A=\$0.01-\$1.99, B=\$2.00-\$2.99, C=\$3.00-\$4.99, D=\$5.00-\$9.99, E=\$10.00-\$14.99, F=\$15.00-\$24.99, G=\$25.00 and above.

There is a problem here. In 66,900 cases, codes A, B, and C did not occur. We must build our model without them but if they occur in the future we will have to do something sensible. Most reasonable would probably be to use the prediction for D.

Table 7. RFA_2A Coded as Dummy Variables

476.dat	x_1	x_2	x_3	x_4	x_5	x_6	x_7
RFA_2A	A	B	C	D	E	F	G
E	0	0	0	0	1	0	0
G	0	0	0	0	0	0	1
E	0	0	0	0	1	0	0
E	0	0	0	0	1	0	0
F	0	0	0	0	0	1	0
F	0	0	0	0	0	1	0
E	0	0	0	0	1	0	0
E	0	0	0	0	1	0	0
F	0	0	0	0	0	1	0
F	0	0	0	0	0	1	0
F	0	0	0	0	0	1	0
D	0	0	0	1	0	0	0
G	0	0	0	0	0	0	1
F	0	0	0	0	0	1	0
F	0	0	0	0	0	1	0
G	0	0	0	0	0	0	1
F	0	0	0	0	0	1	0
F	0	0	0	0	0	1	0
D	0	0	0	1	0	0	0
etc.							

RFA_2A Regression

With dummy variables hand coded as in Table 7 one would run the regression

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + b_5 x_{i5} + b_6 x_{i6} + b_7 x_{i7}$$

This regression has two technical problems: The variables x_1, x_2, and x_3 are identically zero and the variables x_4, x_5, x_6, and x_7 sum to one identically. I.e. there are three null variables and the last four variables are colinear with the constant term, which is the implied variable $x_0 \equiv 1$ that multiplies b_0.

Much statistical software will handle this gracefully. What the program will do automatically is run the regression

$$y_i = b_0 + b_4 x_{4i} + b_5 x_{5i} + b_6 x_{6i}$$

either explicitly or by doing something that is mathematically equivalent.

If the software doesn't, you'll have to do it yourself. That is, one deletes x_1, x_2, and x_3 to get rid of null variables and deletes x_7 to remove the exact multicolinearity.

Comment on Multicolinearity

Multicolinearity causes one fits in ordinary regression analysis where the coefficients have an interpretation and are of interest in and of themselves, e.g. a price elasticity.

Multicolinearity is the situation where one independent variable in a regression can be nearly perfectly predicted in a regression of it on all the other independent variables.

In data mining, multicolinearity is not a problem as long as it does not destabilize the algorithm used to solve the least squares minimization problem.

This is because the coefficients themselves have no meaning. It is only the prediction that matters. The prediction is little influenced by multicolinearity.

Character Variable Regressions

Dummy variable regressions were run subject to these rules:

- Run with missing values deleted. The consequence may be that the regression cannot be run because only one level remains.
- Run again with missing values treated as a level.
- If there are more than 25 levels, randomly delete cases so that only 15,000 remain to preserve memory. (A defect of R; SAS would not have this problem.)
- If there are more than 100 levels, do not run the regression. (Even SAS would be hard pressed at this point; it would probably run but take forever.)

Character Variables: Significant Results

charity/lrn/chr/chr_reg.txt

```
File   Feature    Type  Levels  Missing  P-Value
 12    RECPGVG     chr  2       0        2.2e-16 ***    planned giver
 26    GENDER      chr  6       2096     2.2e-16 ***    m,f,u,j
 26    GENDER      chr  7       0        2.2e-16 ***    m,f,u,j
 53    MAJOR       chr  1       66703    too few levels major donor
 53    MAJOR       chr  2       0        2.2e-16 ***    major donor
385    RFA_2       chr  14      0        2.2e-16 ***    status as of
385    RFA_2       chr  14      0        2.2e-16 ***    97NK mailing
476    RFA_2A      chr  4       0        2.2e-16 ***    amount of last gift
476    RFA_2A      chr  4       0        2.2e-16 ***    an rfa_2 field
477    MDMAUD_R    chr  5       0        2.2e-16 ***    recency code
477    MDMAUD_R    chr  5       0        2.2e-16 ***    a mdmaud field
478    MDMAUD_F    chr  4       0        2.2e-16 ***    frequency code
478    MDMAUD_F    chr  4       0        2.2e-16 ***    a mdmaud field
479    MDMAUD_A    chr  5       0        2.2e-16 ***    donation amount
479    MDMAUD_A    chr  5       0        2.2e-16 ***    a mdmaud field
```


Character Variables: Conclusions

A careful study of file **chr_reg.txt** leads to the following conclusions:

- Categorical measures of previous amount and frequency of previous gifting from the transactions data are good predictors.
- In many instances missing values actually convey information.
- There are derived features in the data that are useful predictors with no missing values.

Candidate Variables List

The variables identified as being good candidates from the files **num_reg.txt** and **chr_reg.txt** are collected into the file **cty_sel.txt** which looks like this

4	STATE	chr	56	0	0.0001086 ***	can be collapsed
6	MAILCODE	chr	2	0	0.005963 **	bl=good adr
8	DOB	num	923	0	0.5948	birth date sig as quad
10	RECINHSE	chr	2	0	3.621e-13 ***	bl=not IH
11	RECP3	chr	2	0	1.807e-08 ***	bl=not P3
12	RECPGVG	chr	2	0	2.2e-16 ***	bl=not plan giver
14	MDMAUD	chr	26	0	2.2e-16 ***	major donor code
15	DOMAIN	chr	16	1638	1.636e-11 ***	bl=miss urbanicity
26	GENDER	chr	6	2096	2.2e-16 ***	bl=miss m,f,u,j
35	MAGMALE	num	5	37083	4.338e-08 ***	bl=miss
52	SOLIH	chr	8	0	1.114e-09 ***	bl=unlim mail
53	MAJOR	chr	2	0	2.2e-16 ***	bl=not major donor
75	PEPSTRFL	chr	2	0	0.0004269 ***	bl=not
146	HV1	num	4181	0	2.2e-16 ***	Median home value
147	HV2	num	4356	0	2.2e-16 ***	Average home value
148	HV3	num	13	0	2.524e-16 ***	Median rent
149	HV4	num	13	0	2.2e-16 ***	Average rent
173	HVP1	num	99	0	2.2e-16 ***	% HV >= $200,000
174	HVP2	num	99	0	2.2e-16 ***	% HV >= $150,000
175	HVP3	num	99	0	2.2e-16 ***	% HV >= $100,000
176	HVP4	num	99	0	2.2e-16 ***	% HV >= $75,000
177	HVP5	num	99	0	2.2e-16 ***	% HV >= $50,000

etc.

Subtopic Main Points

Machine Assisted Feature Detection:

Upward F-test regression is an effective device to screen features for candidacy and uses machine resources efficiently.

The features selected in this application are reasonable: features derived from RFA and wealth features seem most important, which is consistent with domain knowledge.

We shall not have to deal with missing values: either they are informative or an equivalent feature without missing values is available for every useful feature that has them. Neither deletion nor imputation will be required later.

Machine Assisted Model Building

The third subtopic is machine assisted model building

Here is what we need to do:

1. Discuss the fitting method.
2. Discuss the evaluation criterion.
3. Discuss the algorithm.

Fitting Method and Evaluation Criterion

Validation MSE is computed as

$$\text{MSE} = \frac{1}{N_v}\sum_{i=1}^{N_v}\left(y_i - \hat{b}_0 - \hat{b}_1 x_{i1} - \cdots - \hat{b}_p x_{ip}\right)^2$$

where the average is taken over the validation sample and $\hat{b}$ is computed from the learning sample.

Repeat: The average is over the validation sample, $\hat{b}$ is the regression estimate computed from the learning sample.

Selection Using the Validation Sample

Algorithm:

1. Fit every variable in **cty_sel.txt**; select the variable for which validation MSE is smallest; delete that variable from **cty_sel.txt**.

2. Fit every variable in **cty_sel.txt** as an additional variable to the previous regression; select the variable for which MSE is smallest; delete that variable from **cty_sel.txt**.

3. Repeat 2 if validation MSE declines; terminate if validation MSE is the same or larger.

Additional Details

Before proceeding, we need to discuss (next slide) two derived features:

- One uses common sense to conserve memory and time.

- The other I stumbled across in an (unsuccessful) data cleaning attempt.

Two Derived Features

STATE has too many levels. They can be reduced by combining states into groups. We shall group states whose dummy coefficients are insignificant and/or nearly identical.

```
S1: AS, DC, DE, MA, ME, OH, RI, VI, WV
S2: AA, AE, AP, CT, GU, MD, NJ, NY, PA, VA, VT, WY
S3: AK, UT, MS
S4: NE, ND
S5: SD, SC
```

DOB is not significant as is, but all coefficients in the regression

$$y = b_0 + b_1(\text{DOB}) + b_2(\text{DOB})^2$$

are significant. DOB also has several coding errors. All my attempts to correct them automatically only made the situation worse; manual repair is not in the spirit of machine learning. One can also add a dummy variable for the missing values in DOB to further improve the fit.

Variables Selected

charity/cty_mod.txt

```
464     LASTGIFT   num
 75     PEPSTRFL   chr
  4     STATE      chr
 11     RECP3      chr
  8     DOB        num
  6     MAILCODE   chr
359     MHUC2      num
465     LASTDATE   num
460     MINRAMNT   num
```

Table 8. Definitions of the Selected Features

File	Feature	Type	Definition
464	LASTGIFT	num	Dollar amount of most recent gift
75	PEPSTRFL	chr	Has given to three consecutive card mailings
4	STATE	chr	State of residence
11	RECP3	chr	Has given to CTY's P3 program
8	DOB	num	Date of birth
6	MAILCODE	chr	Mailing address is correct
359	MHUC2	num	Census tract homeowner cost w/out mortgage
465	LASTDATE	num	Date associated with the most recent gift
460	MINRAMNT	num	Dollar amount of smallest gift to date

Subtopic Main Point

Machine Assisted Model Building

- Use the learning sample to estimate model parameters.
- Use validation sample to select the best model.

Stated more succinctly:

- **Learn in the learning sample.**
- **Validate in the validation sample.**

The most important point of the entire course.

Model Assessment

The fourth and last subtopic of linear regression is model assessment.

Two items need discussion:

1. The disconnect.
2. Lift charts.

The Disconnect

We have discussed loss functions.

Tools train by implicitly or explicitly minimizing loss in the learning sample.

Tools are evaluated by computing loss in the validation sample.

Models are assessed by means of lift charts, profits, revenues, or otherwise. These too are loss computations, either implicit or explicit.

The disconnect is that often these three measures of loss are not the same.

What to Do?

Mostly nothing except be aware of the disconnect because in most instances the choice of the learning and validation loss functions are beyond your control.

If you have some control, think about what loss is appropriate to your application and try to make loss the same in learning, validation, and assessment. This is usually done with weights or oversampling as discussed earlier.

Lift charts, next, are an example. They are oriented toward profit maximization, but we used MSE for learning and validation, which is not directed toward that goal.

Later in the course we will see how much effect weighting by over sampling has on the lift chart performance of various tools.

Lift Chart

A lift chart shows the result of using model predictions to maximize revenue.

The formula for a prediction is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p$$

where $\hat{\beta}_0, \ldots, \hat{\beta}_p$ are the coefficients estimated in the learning sample and $x_1, \ldots, x_p$ are the selected features, which are those listed in Table 8 for the case under consideration.

The Rule: Compute $\hat{y}_i$ for each case i in a data set and solicit in order of decreasing $\hat{y}_i$.

When the sample contains targets y_i one can see what revenues would have been achieved had this rule been followed.

A plot of these revenues is called a lift chart (or profit curve if net rather than gross revenue by Berry and Linoff, see p. 85.)

Construction of Lift Charts

In Excel or something similar:

1. Put the features in columns labeled X1,...,Xp and the target in a column labeled TARGET; this will fill rows 1,...,n.

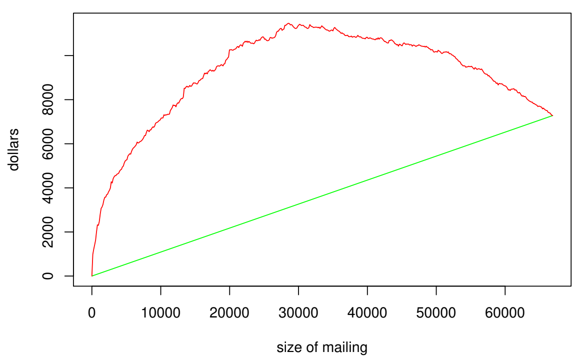
2. Use the features to predict the target and put the prediction in a column labeled YHAT.

3. Sort the entire spread sheet on column YHAT in decreasing order.

4. Put the numbers 1,....,n in a column labeled XAXIS. (Multiply by 100/n if you want results in percent).

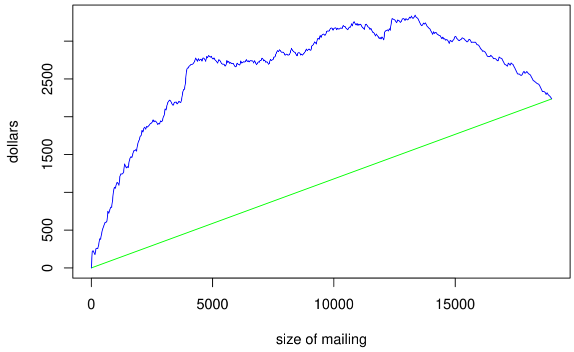
5. Plot the cumulative sum of column TARGET against column XAXIS.

Fig 52. Learning Sample Lift Chart



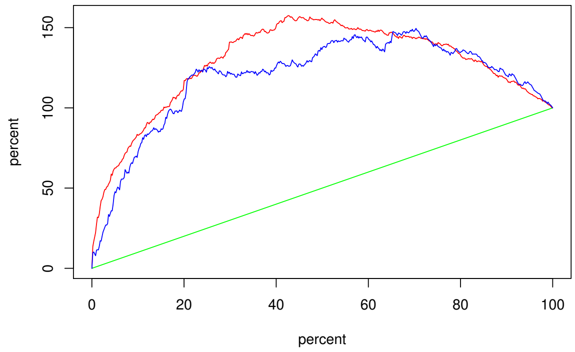
The green curve shows net revenue if persons in the learning sample were mailed solicitations in random order. The red curve shows net revenue if persons are sorted by their predicted gift and mailed solicitations in sorted order, highest first. Net revenue is the gift less a mailing cost of $0.68.

Fig 53. Validation Sample Lift Chart



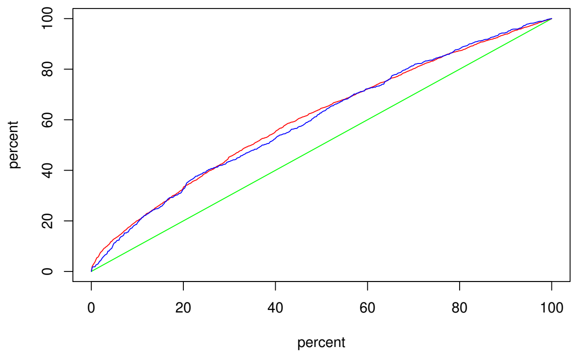
The green curve shows net revenue if persons in the learning sample were mailed solicitations in random order. The blue curve shows net revenue if persons are sorted by their predicted gift and mailed solicitations in sorted order, highest first. Net revenue is the gift less a mailing cost of $0.68.

Fig 54. Both Lift Charts



Same as Figs 52 and 53 but with endpoints normalized to plot at (100,100). Validation maximum net revenue occurs at 70% of the sample whereas learning is 45%. The difference is a measure of generalization discrepancy.

Fig 55. Conventional Lift Chart



Same as Figs 54 but with gross revenue rather net revenue. The chart shows, e.g., that the top 20% of the solicitations will account for 30% of the total donations. Unlike Figs 54, one cannot easily determine the point of maximum profitability from the chart.

Subtopic Main Point

Model Assessment

Lift charts are a common way to summarize data mining findings.

They give a visual impression of the possible gains to use of model output.

They give a visual impression of the generalizability of a model.

Charts in terms of net revenue convey more information.

Regression Main Points

- The oldest method (Gauss, 1816).
- Efficient use of space and time.
- Works very well when features are well chosen.
- F-tests should only be used to rank features in large samples.
- Validation should be used for model selection.

Age Bias

Don't be fooled: There is a bias toward the newer tools because "Who needs a consultant if all they can recommend is an idea that is two hundred years old."

As we shall see as the case unfolds, regression is a powerful tool that holds its own quite well in competition with newer methods.

Its drawback is that it relies heavily on correct choice of features and derived features. It has no built in mechanism for deriving features itself as do the newer tools.

Nonetheless, as we shall see, human intelligence can, at times, be better than machine intelligence.

Blank page

Blank page