Model Validity Checks In Data Mining: A Luxury or A Necessity?

Mike Speed & Simon Sheather
Department of Statistics
Texas A&M University

Abstract

It has been stated that, because of vast amounts of data, the data miner does not have the luxury of verifying assumptions and hence checking for model validity. This paper uses two databases to investigate implications of not verifying assumptions and hence the validity of models. The first one is the Dominick’s Finer Foods database (James M. Kilts Center, GSB, University of Chicago). This database consists of same store sales of 58 stores over 268 weeks. The second one is The Direct Mail Fundraising database. This paper demonstrates that ensuring that the fitted models for either continuous or binary outcomes are valid improves predictions, sometimes dramatically so. In these real examples considered in this talk, using valid models, greatly improved the results. Thus, the data miner may want to consider that the checking of assumptions and validity is not a luxury, but in fact a necessity.

KEYWORDS: data mining  model validity covariance structure
Database Number 1

- Dominick’s: Customer Count Files

- James M. Kilts Center, GSB, University of Chicago

- [http://research.chicagogsb.edu/marketing/databases/dominicks/index.aspx](http://research.chicagogsb.edu/marketing/databases/dominicks/index.aspx)
Dominick’s Finer Foods

- Dominick's: Customer Count Files
- James M. Kilts Center, GSB, University of Chicago
- http://research.chicagogsb.edu/marketing/databases/dominicks/index.aspx

“The customer count file includes information about in-store traffic. The data is store specific and on a daily basis. The customer count data refers to the number of customers visiting the store and purchasing something. Also in the customer count file is a total dollar sales and total coupons redeemed figure, by DFF defined department. These figures are compiled daily from the register/scanner receipts.”

Data

- Weekly averages were computed from the daily sales
- No coupon data was used
- Variables used: GROCERY_mean (Dependent)
  DAIRY_mean; FROZEN_mean; MEAT_mean;
  MEATFROZ_mean; FISH_mean :PRODUCE_mean;
  BULK_mean; SALADBAR_mean; FLORAL_mean;
  DELI_mean; CHEESE_mean; BAKERY_mean; GM_mean;
  HABA_mean; BEER_mean; WINE_mean; SPIRITS_mean;
  CUSTCOUN_mean
- 58 stores
- 268 weeks
- 13998 observations
### Variable Description

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROCERY_mean</td>
<td>Grocery Sales in Dollars</td>
</tr>
<tr>
<td>Week</td>
<td>Week Number</td>
</tr>
<tr>
<td>Store</td>
<td>Store Code</td>
</tr>
<tr>
<td>BAKERY_mean</td>
<td>Bakery Sales in Dollars</td>
</tr>
<tr>
<td>BEER_mean</td>
<td>Beer Sales in Dollars</td>
</tr>
<tr>
<td>BULK_mean</td>
<td>Bulk Sales in Dollars</td>
</tr>
<tr>
<td>CHEESE_mean</td>
<td>Cheese Sales in Dollars</td>
</tr>
<tr>
<td>CONVFOOD_mean</td>
<td>Conventional Foods Sales in Dollars</td>
</tr>
<tr>
<td>CUSTCOUN_mean</td>
<td>Customer Count</td>
</tr>
<tr>
<td>DAIRY_mean</td>
<td>Dairy Sales in Dollars</td>
</tr>
<tr>
<td>DELI_mean</td>
<td>Deli Sales in Dollars</td>
</tr>
<tr>
<td>FISH_mean</td>
<td>Fish Sales in Dollars</td>
</tr>
<tr>
<td>FLORAL_mean</td>
<td>Floral Sales in Dollars</td>
</tr>
<tr>
<td>FROZEN_mean</td>
<td>Frozen Items Sales</td>
</tr>
<tr>
<td>GM_mean</td>
<td>General Merchandise Sales in Dollars</td>
</tr>
<tr>
<td>HABA_mean</td>
<td>Health and Beauty Aids Sales in Dollars</td>
</tr>
<tr>
<td>MEAT_mean</td>
<td>Meat Sales in Dollars</td>
</tr>
<tr>
<td>MEATFROZ_mean</td>
<td>Meat-Frozen Sales in Dollars</td>
</tr>
<tr>
<td>PRODUCE_mean</td>
<td>Produce Sales in Dollars</td>
</tr>
<tr>
<td>SALADBAR_mean</td>
<td>Salad Bar Sales in Dollars</td>
</tr>
<tr>
<td>SPIRITS_mean</td>
<td>Spirits Sales in Dollars</td>
</tr>
<tr>
<td>WINE_mean</td>
<td>Wine Sales in Dollars</td>
</tr>
</tbody>
</table>
Business Questions

• How is Grocery_mean related to the other variables?

• Does increasing/decreasing the sales on one variable reduce or enhance the Grocery_mean sales?

• What role does customer count play on the Grocery_mean sales?

• Should we model the correlation among the same store sales?

Three Stores for Selected Weeks

• Use a 60/40 split
  – 60 is training set
  – 40 is validation set
### Regression

#### Effect Estimate Pr > |t|
- DAIRY\_mean 1.80 <.0001
- FROZEN\_mean 1.35 <.0001
- MEAT\_mean 0.75 <.0001
- MEATFROZ\_mean 1.09 <.0001
- FISH\_mean -1.60 <.0001
- PRODUCE\_mean 0.27 <.0001
- BULK\_mean 0.50 <.0001
- SALAD\_BAR\_mean 0.50 <.0001
- FLORAL\_mean 0.04 0.5952
- DEJU\_mean 0.98 <.0001
- CHEESE\_mean -0.66 <.0001
- BAKERY\_mean -0.42 <.0001
- GM\_mean 0.39 <.0001
- HABA\_mean 0.19 <.0001
- BEER\_mean 0.15 0.0577
- WINE\_mean -0.78 <.0001
- SPIRITS\_mean 0.29 0.073
- CUST\_COUN\_mean -0.68 <.0001

What seems to be of interest here?

---

### Regression

#### Fit Statistics
- R-Square 0.9671
- Adj R-Sq 0.9659
- AIC 119040.2652
- BIC 119062.4269
- SBC 121052.5235
- C(p) 286.0000

Mean Square Error 1382894

What seems to be of interest here?
Regression

<table>
<thead>
<tr>
<th>TARGET</th>
<th>Fit statistic</th>
<th>Statistics</th>
<th>Train</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROCERY_m__AC_</td>
<td>Akaike's Information</td>
<td>119040.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__ASE_</td>
<td>Average Squared Error</td>
<td>1335004</td>
<td>1340034</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__AVERR_</td>
<td>Average Error</td>
<td>1335004</td>
<td>1340034</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__DFE_</td>
<td>Degrees of Freedom</td>
<td>11113</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__DFM_</td>
<td>Model Degree</td>
<td>286</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__DFT_</td>
<td>Total Degree of Freedom</td>
<td>8399</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__DFW_</td>
<td>Degrees of Freedom for ASE</td>
<td>8399</td>
<td>5599</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__ERR_</td>
<td>Error Function</td>
<td>1122010</td>
<td>750059</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__FPE_</td>
<td>Final Prediction Error</td>
<td>1420084</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__MAX_</td>
<td>Maximum ASE</td>
<td>3136501</td>
<td>1104833</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__MSE_</td>
<td>Mean Square Error</td>
<td>1382384</td>
<td>1340034</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__NOBS_</td>
<td>Sum of Frequencies</td>
<td>8399</td>
<td>5599</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__NPS_</td>
<td>Number of Pairs</td>
<td>206</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__RASE_</td>
<td>Root Average Squared Error</td>
<td>115527</td>
<td>1157946</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__RFM_</td>
<td>Root Final Prediction Error</td>
<td>1196019</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__RME_</td>
<td>Root Mean Error</td>
<td>1175965</td>
<td>1157945</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__SDE_</td>
<td>Schwarz's B</td>
<td>121052.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__SSF_</td>
<td>Sum of Squares</td>
<td>1122010</td>
<td>750059</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GROCERY_m__SUMV_</td>
<td>Sum of Cases</td>
<td>8399</td>
<td>5599</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

What seems to be of interest here?

Run Regression As Proc Mixed -

- **REPEATED WEEK / SUBJECT=STORE TYPE=VC** ;

- Same Results as with Regression
Run Mixed with TYPE = CS

- **REPEATED WEEK / SUBJECT=STORE**
  
  
  ```
  TYPE=CS ;
  
  \[
  \begin{bmatrix}
  \sigma^2 + \sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\
  \sigma_1 & \sigma^2 + \sigma_1 & \sigma_1 & \sigma_1 \\
  \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1 & \sigma_1 \\
  \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1 
  \end{bmatrix}
  \]
  
  Compound Symmetry CS
  ```

- Here we have a 268 by 268 covariance matrix

- Different Results

| Effect          | Estimate VC | Pr > |t| Estimate VC | Pr > |t| Estimate CS | Pr > |t| Estimate CS |
|-----------------|-------------|------|-------------|------|-------------|------|-------------|------|-------------|
| DAIRY_mean      | 1.80        | <.0001 | 1.8030 | <.0001 |
| FROZEN_mean     | 1.35        | <.0001 | 1.3086 | <.0001 |
| MEAT_mean       | 0.75        | <.0001 | 0.4463 | <.0001 |
| MEATFROZ_mean   | 1.09        | <.0001 | 0.6136 | <.0001 |
| FISH_mean       | -1.60       | <.0001 | -0.4204 | <.0001 |
| PRODUCE_mean    | 0.27        | <.0001 | 0.1542 | <.0001 |
| BULK_mean       | 0.50        | <.0001 | 0.5855 | <.0001 |
| SALADBAR_mean   | 0.50        | <.0001 | 1.1951 | <.0001 |
| FLORAL_mean     | 0.04        | 0.5952 | -0.3734 | <.0001 |
| DELI_mean       | 0.98        | <.0001 | 0.1402 | 0.0025 |
| CHEESE_mean     | -0.66       | <.0001 | -0.3511 | 0.0317 |
| BAKERY_mean     | -0.42       | <.0001 | -0.5755 | <.0001 |
| GM_mean         | 0.39        | <.0001 | 0.3047 | <.0001 |
| HABA_mean       | 0.19        | <.0001 | 0.3665 | <.0001 |
| BEER_mean       | 0.15        | 0.0577 | 0.7665 | <.0001 |
| WINE_mean       | -0.78       | <.0001 | 0.6275 | <.0001 |
| SPIRITS_mean    | 0.29        | 0.073  | -0.9958 | <.0001 |
| CUSTCOUN_mean   | -0.68       | <.0001 | 0.5686 | <.0001 |
Fit Statistics

Null Model Likelihood Ratio Test

<table>
<thead>
<tr>
<th></th>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>3889.46</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>STORE</td>
<td>1383624</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>826177</td>
</tr>
<tr>
<td>Rho</td>
<td></td>
<td>.65</td>
</tr>
</tbody>
</table>

What seems to be of interest here?

Prediction Error

```
libname me 'C:';

data me.pred_cs;
set me.tmp1_cs;
predms_cs = ( GROCERY_mean-pred )**2;
run;
quit;

PROC MEANS DATA=me.pred_cs
   FW=12 MEAN ;
   VAR predms_CS;
RUN;
```
Proc Mixed - CS

<table>
<thead>
<tr>
<th>dataset</th>
<th>mse_vc</th>
<th>mse_cs</th>
<th>cov_cs</th>
<th>Prediction_VC</th>
<th>Prediction_CS</th>
<th>Percent Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>train1 60/40</td>
<td>1382894</td>
<td>826177</td>
<td>1383624</td>
<td>1340603.95</td>
<td>776236.49</td>
<td>-42.10</td>
</tr>
</tbody>
</table>

What seems to be of interest here?

Enterprise Miner - SAS®

Data Partition 60/40

Mixed CS

Prediction_error_CS

Prediction_error_VC

Mixed_VC

Regression
Conclusion for Dominick’s Finer Foods

- Although $R^2$ is large (.97), there is correlation among the same stores sales.

- The regression model gives misleading information.

- Using Compound Symmetry to model the correlation produces a model with significantly less prediction mean square error (-42%).

Part 2: Logistic regression in data mining

- When should you include quadratic predictor terms?
- When should you include interactions between predictors?
- Marginal model plots for checking model validity
- Case study
Logistic regression: why assume the logistic S-shaped curve?

Suppose that Y is a binary random variable (i.e., takes values 0 and 1) and that X is a single predictor variable. Then we seek a model for

\[ P(Y = 1 \mid X = x) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)} \]

A popular choice for the S-shaped function for \( \theta(x) \) is the logistic function, that is,

\[ \exp(-\beta_0 - \beta_1 x) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)} \]

Solving this last equation for \( \beta_0 + \beta_1 x \) gives

\[ \log \left( \frac{\beta_0 + \beta_1 x}{1 - \theta(x)} \right) = \beta_0 + \beta_1 x \]

In what circumstances would you consider adding quadratic or interaction terms to a logistic regression model?

Logistic regression: Justification for linear and quadratic terms

Suppose that \( X \mid Y=\gamma, \gamma=0,1 \) is normally distributed with mean \( \mu_\gamma \) and variance \( \sigma_\gamma^2 \), \( \gamma=0,1 \). Then it can be shown that,

\[ \log \left( \frac{\theta(x)}{1 - \theta(x)} \right) = \beta_0 + \beta_1 x + \beta_2 x^2 \]

where

\[ \beta_2 = \frac{1}{2} \left( \frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2} \right) \]

Thus, when the predictor variable X is normally distributed with a different variance for the two values of Y, the log odds are a quadratic function of x.

When \( \sigma_0^2 - \sigma_1^2 = \sigma^2 \), the log odds simplifies to

\[ \log \left( \frac{\theta(x)}{1 - \theta(x)} \right) = \beta_0 + \beta_1 x \]

where

\[ \beta_1 = \left( \frac{\mu_1 - \mu_0}{\sigma^2} \right) \]
Logistic regression: Justification for interactions and transformations

The last result can be extended to the case where we have $p$ predictor variables which have multivariate normal conditional distributions. If the variance-covariance matrix of the predictors differs across the two groups then the log odds are a function of $x_i, x_i^2$ and $x_i \times (i, j = 1, ..., p; i + j)$.

Simulated normal data:
means=0,2; sds = 1,1

$$\theta(x) = P(Y = 1 \mid X = x)$$

where $$\log \left( \frac{\theta(x)}{1 - \theta(x)} \right) = \beta_0 + \beta_1 x$$
Simulated normal data: means=0,0; sds = 1,3

\[ \theta(x) = P(Y = 1 \mid X = x) \]

where \( \log\left(\frac{\theta(x)}{1 - \theta(x)}\right) = \beta_0 + \beta_1 x + \beta_2 x^2 \)

Building a marginal model plot - Simulated normal data: means=0,2; sds = 1,1

\[ \theta(x) = P(Y = 1 \mid X = x) = E(Y \mid X = x) \]

where \( \log\left(\frac{\theta(x)}{1 - \theta(x)}\right) = \beta_0 + \beta_1 x \)

\[ E(\hat{Y} \mid X = x) \]

\[ E(Y \mid X = x) \]
Building a marginal model plot - Simulated normal data: means=0,0; sds = 1,3: Incorrect model

\[ \theta(x) = P(Y = 1 \mid X = x) = E(Y \mid X = x) \]

where \( \log \left( \frac{\theta(x)}{1 - \theta(x)} \right) = \beta_0 + \beta_1 x \)

\[ E(\hat{Y} \mid X = x) \]

\[ E(Y \mid X = x) \]

---

Building a marginal model plot - Simulated normal data: means=0,0; sds = 1,3: Correct model

\[ \theta(x) = P(Y = 1 \mid X = x) = E(Y \mid X = x) \]

where \( \log \left( \frac{\theta(x)}{1 - \theta(x)} \right) = \beta_0 + \beta_1 x + \beta_2 x^2 \)

\[ E(\hat{Y} \mid X = x) \]

\[ E(Y \mid X = x) \]
CASE STUDY:

- Illustrates the importance of checking whether the current model is a valid model
- Highlights the dangers of a black box approach to modeling
- Demonstrates the value of marginal model plots in data mining

Direct Mail Fundraising data set:
3120 data points with 50% training & 50% validation

5.1% Response Rate
Average Donation = $13.00
Cost To Solicit = $0.68
Marginal Model Plots for the full model without transformations and extra terms (MAIN EFFECTS)

Description of the steps taken to get to the final model

- Perform exploratory analysis on dataset
- Ran logistic regression
- Diagnostics indicated problems
- Used an ad-hoc Box-Cox procedure to get transformations
- Created a new factor of zip*wealth
- Conjectured that the assumptions of equal slopes by the transformed variables might not be valid
- Reran the logistic with transformed variables and interactions – did forward selection
- Diagnostics were much better
Final Variables and Their Transformations

- (Average Home Value)^-.2
- (Median Family Income)^.4
- (Average Family Income)
- Sqrt(Number of Promotions)
- Log($ Amount of Lifetime Gifts)
- ($ Amount of Largest Gift)^.2
- Sqrt(Average Gift)

Final Model

Target_B = zip_wealth
genderdummy
homeownerdummy
HV_T*zip_wealth
lcmed_T*zip_wealth
lcavg_T*zip_wealth
numprom_T*zip_wealth
ramntall_T*zip_wealth
maxramnt_T*zip_wealth
avggft_T*genderdummy
Marginal Model Plots for the Final model after transformations and extra terms
Comparison of pay-offs for original and final models – Average Profit

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Transformation</td>
<td>0.066</td>
<td>0.048</td>
</tr>
<tr>
<td>Transformations</td>
<td>0.075</td>
<td>0.059</td>
</tr>
</tbody>
</table>

Transformations Model is 14% better for the Training Dataset and 23% better than Validation Dataset

Summary

- Illustrated the importance of checking whether the current model is a valid model
- Highlighted the dangers of a black box approach to modeling
- Demonstrated the value of marginal model plots in data mining
- Demonstrated that all this can be achieved in SAS® and SAS® Enterprise Miner
Conclusion

• Yes, it takes time to do checks on the validity of the model
• But not doing the checks can lead to incorrect conclusion and less than optimal models
• Statisticians working together with the data miners should be able to develop methods for doing checks on the validity of the model which take less time.

References

• Dominicks’ Dataset
  – [http://research.chicagogsbg.edu/marketing/databases/dominicks/index.aspx](http://research.chicagogsbg.edu/marketing/databases/dominicks/index.aspx)

• “Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner” by Galit Shmueli, Nitin R. Patel, Peter C. Bruce. Wiley-Interscience (December 11, 2006)
About The Authors

F. Michael Speed is a Professor of Statistics at Texas A&M University
If you have any questions/comments you can contact Mike at:
Texas A&M University
3143 TAMU
College Station, TX 77843-3143
Voice(979 - 854 - 4141)
Email mspeed@stat.tamu.edu

Simon Sheather is a Professor and Head of the Department of Statistics
Contact information is the same as above
Email sheather@stat.tamu.edu