Hearst Challenge : A Framework to Learn Magazine Sales Volume

Reza Omrani

1 Introduction

This project is based on a 25000$ competition by the Hearst Magazines posted on www.kdnuggets.com . The goal of this project is to predict the sales of 10 different magazine titles published by the Hearst Magazines at each newsstand location across United States, to optimize the overall contribution of the newsstand locations. To predict the magazines' sales volume, Hearst magazines has provided a database of the sales of 10 different magazines for almost 40390 newsstands across the US for the period of $2006-2009$. The data set includes 10000000 training examples. In addition they have provided different statistical demographical information from the different sources. We need to find a learning algorithm which can use these information to learn the pattern of each title's sales based on newsstand statistical demographic and time frame.

There are 5 different category of Information provided by the Hearst Magazines as follows:

1. Store/Chain table : This table has the information of the 40390 training newsstands. Each store's information includes two category of data: some descriptive information about the stores plus the demographic of the neighborhood stand is located in. The store description data includes:

   (a) Chain Key: unique identifier denoting a particular chain. Each chain has numerous stores under it.

   (b) Store key: unique identifier denoting a store

   (c) City

   (d) Zip Code: $(5+4)$ zip code

   (e) State

   (f) Store Type: BOOKSTO, C-STORE, CLUB ST, COLLEGE, DEPARTM, DISCOUN, DRUG, ENTERTA, HOME &, LIQUOR, MASS ME, MILITAR, NEWSSTA, OFFICE/, OTHER, SPORTS, SUPERCE, SUPERMA, TERMINA

   Demographic information includes average statistical information of the 5 digit zip code area in which the store is located:

   (a) Number of households

   (b) Number of Individuals older than 18

   (c) Number of households with different type of occupations: professional/technical; sales/service; farm related; blue collar; other; retired; unknown

   (d) Number of households with each educational level: high school diploma; some college; Bachelor degree; Graduate degree; less than high school diploma; unable to determine education

   (e) Individuals age distribution: 19; 20; 21; 22-24; 25-29; 30-34; 35-39; 40-44; 45-49; 50-54; 55-59; 60-64; 65-69; 70-75; > 75

   (f) Gender distribution: male; femal; unknown

   (g) Marital status distribution: married; single; other (including divorced and widowed)

   (h) Household distribution: head of household; spouse; young adult; elderly individual

   (i) Income distribution: $< 15K$; $15K$-$25K$; $25K$-$35K$; $35K$-$50K$; $50K$-$75K$; $75K$-$100K$; $100K$-$125K$; $125K$-$150K$; $150K$-$175K$; $175K$-$200K$; $200K$-$250K$; $> 250K$; unknown

   (j) Home ownership distribution: owner; renter; unknown

   (k) Residence type distribution: single family; multi-family; multi-family or marginal; post office box indication

   (l) Length of residence distribution (in years): < 1; $= 1$; $2-5$; $6-10$; > 10

   (m) Number of units at address distribution (in units): 1; 2; 3; 4; 5-9; 10-19; 20-49; 50-100; > 100

   (n) Median house value distribution: $< 100K$; $100K$-$200K$; $200K$-$300K$; $300K$-$500K$; $500K$-$1M$; $> 1M$; unknown

   (o) Some extra summarized area level statistics: A-H

   (p) Vehicle distribution in (5 + 4) zip code: number of vehicles, number of households with a vehicle, number of households without a vehicle, number of new cars, number

of new light trucks, number of used cars, number of used light trucks, average MSRP for vehicle, average current vehicle retail value

2. Zip+4 table: includes the same statistical data as above for every 9 digit $(5+4)$ zip code in US.

3. Sales table: includes the sales data of various newsstands(declared with store key) for a given issue of a magazine title.

4. Issue table: Gives price, on sale date, and off sale date for every issue in the sales table.

5. Wholesaler table: Gives wholesaler geographical information for each newssatand wholesaler.

Furthermore a blind test data of size 10000 is provided, once your algorithm works good enough you can run the test data and predict the sales of some different newsstand and submit it to the competition to get their feedback on your predictions.

Now let's look at the problem more carefully. As it is observed, there are too many parameters in the model, and we don't have enough training data (already a huge training set) to use all of them. In addition, many of these parameters are highly correlated and using them may not be useful. That is why we need to apply some feature extraction method to data to focus on the useful data. Furthermore we need to come up with a good model which works good both on trainig data and test data. Finally we should consider the time variations, as we have training points for different points of the time. One idea may be to use separate model for each time of interest (48 months).

Once we learned the sales volume for the training examples, for any given new store, with given descriptive data, the demographic data can be extracted from Zip+4 data. Once we have all the parameter about the store we can feed it to our model and get an estmate on sales volume.

In following, we focus on learning the sales of one title, say title 1 for a given period. Once we have estimated the sales of one title, the same process can be repeated for all the other titles. Furthermore, as mentioned above, we can repeat the same procedure for any given period. Once we are done with modeling one title's sales for a given period, we will repeat the same procedure with several other title/period combination to make sure our algoithm is working satisfactorily. In follwoing we investigate these issues in more detail.

2 Data Processing

All the simulations for this project are done, using R programming language. As described above the data sets are too big, and they are not clean (there are many missing and repeated data), that is why we are using some preprocessing to reformat data sets to more useful forms. First of all, the sales data is broken to 10 different files for each title. The next step is to combine issue table and sales table, as we wish to have monthly prediction of sales while not all of the titles are monthly periodicals(one of them is weekly, one is bi-weekly, and one is bi-monthly). Furthermore we don't have the sales information of each title for all data points, and that makes the problem even more complicated, as for some time intervals we have much larger training sets compared to other time intervals, and our algorithm should take that into the consideration too.

We choose a random issue of title 1, and extract all the sales data corresponding to the issue. Then we link every sales point for the given issue with the store's information. Looking at sales histogram, the sales volume beyond 50 issue is very sporadic, and including these points just biases our estimation, that is why, we get rid of the sales$>$ 50 data in following.

3 Linear Model

After all we get to the actual modelling. As a first step, let's try to fit a linear model based on the statistical demographics of the newsstands for thr specific title for a random time interval.

To do this the training data is randomly assigned to 80% training and 20% cross validation portions, and we used R to fit a linear model to the training portion. To fit the model, we need to get rid of some correlated numbers. Since the summation of numbers in each category adds up to either number of households, or number of individuals, we need to get rid of one of the subcategories in each group (otherwise they won't be independent and a linear model can't be fit). After doing these modifications, we fit a full model to the training data and use a backward elimination using p-value of each variable to reduce the model. The backward elimination continues untill the p-value of all the variables in the model are smaller than 0.05(all the parameters are significant). Repeating this for various time intervals gave almost uniform results in terms of number of variables in the model and the major variables in the model.

The initial full model starts with 85 parameters with residual standard error equal to 10.4, and the final reduced model has 40 parameters with standard error equal to 10.4. The derived model parameters show some nice positive and negative correlation be-

haviors of sales with different parameters. Although it should be noticed that this is newsstand sales and doesn't include subscription to magazines through mail and internet.

But how good is this model? The mean of sales for this data is 10.4 and the standard deviation of sales is 10.7 . So our model is barely doing better than the average estimator. Let's see why is that? We know that, the main assumption of linear regression is that the residuals should be samples of a normal distribution, but if we look at the plot of the residuals for all training examples in Figure 1, we see that the distribution of residuals doesn't look like to be samples of a normal distribution, and they are asymmetric around zero. If the residuals were samples of a normal distribution, then the Q-Q plot of studentized residual vs. standard normal which depicted in Figure 1 should be a straigh line, but it can be seen that it is far from a straigh line, and it has curvature in middle, which is not common.

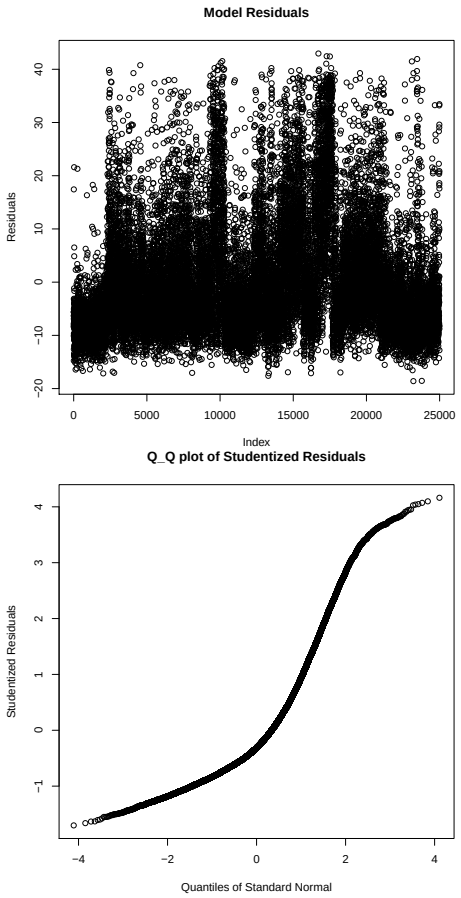

Figure 1: Residual plots for linear model based on just demographics

If we study the data and residuals more closely, we see that the data is dominated with a lot of small sales values, such that more than 60% of the sales values are below 10, and more than 80% of sales values are less than 20, but the tail is not decreasing fast enough and still more than 2% of sales values are above 40(for a normal distribution we expect less than 0.001 probability at $\mu+3\sigma$). So the linear model behavior is governed by the small sales values, but we have enough large sales values which increase the standard deviation, and make the residuals asymmetric.

4 Categorical Linear Model

The above observation, leads to a model, which uses different models for different portions of sales data. Let's assuume that we have a genie that tells us the nighborhood of sales values, say if $sales > 20$ or not. Let's see how informative that information is. Let's treat this genie information as a categorical parameter and fit a model to sales using the parameters from the last part plus this genie information.

The initial full model starts with 86 parameters with residual standard error equal to 5.9, and the final reduced model has 34 parameters with standard error equal to 5.9, and obviously the smallest p-value corresponds to the genie information. So the genie information reduces standard error to almost half. Essentially the same model with two different constant intercept for training examples with $sales \leq 20$ and $sales > 20$ improved model a lot.

Obviously, the next step is to go to more genie information say $sales \leq 5$, $5 < sales \leq 10$, and $10 < sales \leq 20$. The initial full model starts with 88 parameters with residual standard error equal to 3.6, and the final reduced model has 12 parameters with standard error equal to 3.6 .

Plots int the Figure 2 show the residual plots for the above mentioned genie models. As it can be seen the residuals are getting more symmetric and the variation of residuals is decreasing too, and that is the reason why the residual standard error is improving.

Unfortunately we have no genie, to tell us the neighborhoo of the sales (we can not use modeled parameter in its own estimation). Let's see if we can use the discriptive data in our data base to categorize sell. Lets look at the store type distribution in the original data set (training + test) for $sales > 20$ and $sales \leq 20$ depicted in Figure 3. The store type distributions are apparantly diffeent in two cases, and we can use this difference in distributions to estimate the genie information.

Let's use the original demographic parameters and additional categorial variables on type of the stores to fit a linear model to the sales data. The initial full model starts with 103 parameters with residual standard error equal to 8.7, and the final reduced

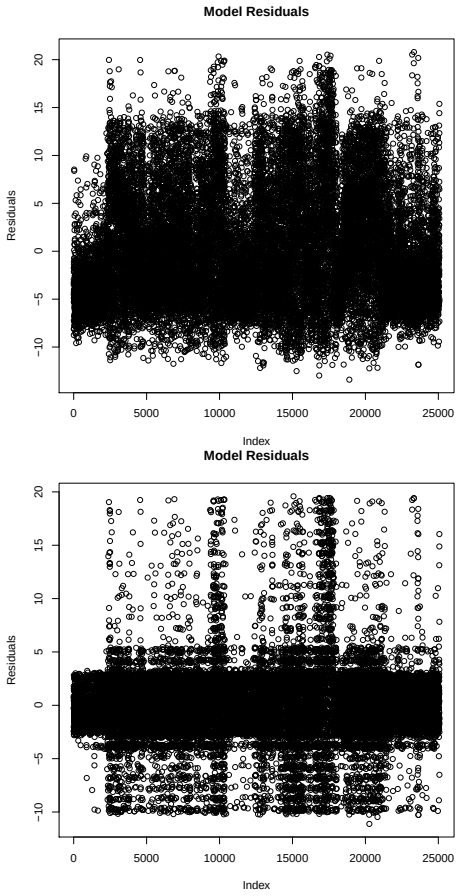

Figure 2: Residual plots for model with extra genie parameter $sales > 20$; and the model for extra genie parameters $sales \leq 5$, $5 < sales \leq 10$, and $10 < sales \leq 20$

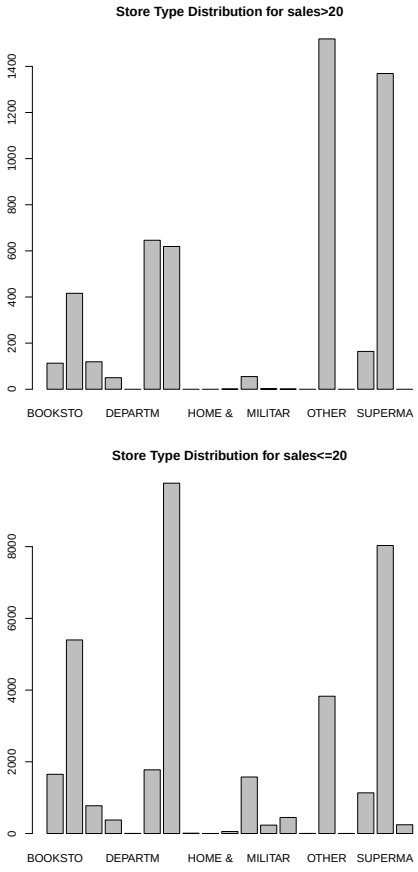

Figure 3: Store type distribution for $sales > 20$ and $sales \leq 20$ in the whole data set

model has 66 parameters with standard error equal to 8.7, and majority of categorical parameters are present in the final reduced model. Similarly, we can add the categorical data of which state the store is located in, to get even more information, and if we do so, the initial full model starts with 136 parameters with residual standard error equal to 8.6, and the final reduced model has 86 parameters with standard error equal to 8.6. Including state information we get a little bit of gain over store type model, but if we check state information, we will find out that, all the data are from only 34 states. If we use the model to estimate sales of a store from some other state, it should be OK, as it uses all non-state regressors to estimate the sales(assuming the mean of the left out states doesn't differ significantly with the mean of the present 34 states). As, we are not gaining much by adding state information, and it complicates the model, we focus on the earlier model with only store type data added to demographical data. This model is presented in Figure 5.

Obviously, by going to this model, we have improved the residual standard error a lot, although it is not as good as genie information. Let's look at the residual plots of the model presented in Figure 4. It can be seen that the residual distribution is more symmetric around zero, and the Q-Q plot of studentized residuals vs. standard normal is much closer to straight line compared to Figure 1. All of these indicate that our residuals distribution is close to normal. If the normal and iid assumptions holds for residuals, then we cannot improve our model, and as the residual plots show enough resemblance to normal distribution, we can stop here.

Now that our model is working good enough on our training data, we need to cross validate it with the test data. If the above model is used, to predict the sales value for the test portion of data(which is not used yet), we get some estimate for the sales of each store in the test data. If we compute the residual error between the estimated sales value vs. actual sales value for all training data, the quantity has mean equal to 0.11 and standard deviation equal to 8.8, which is close enough to our model residual mean on training data, ≈ 0 and standard error $= 8.7$.

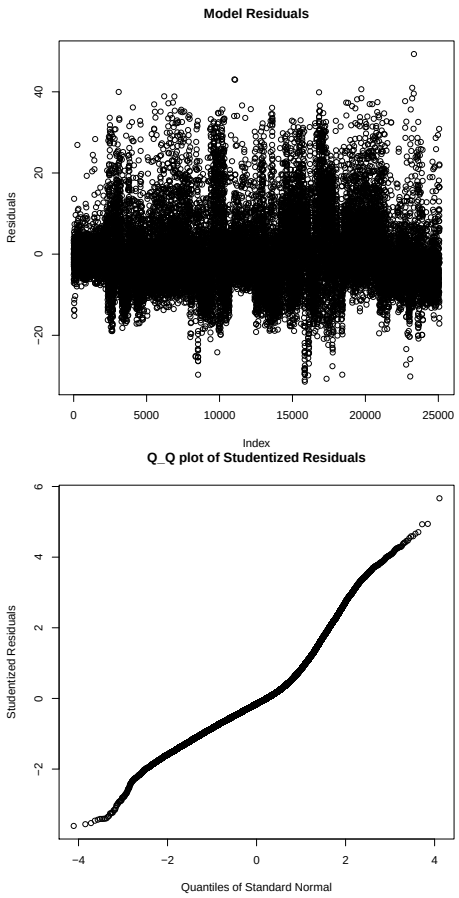

Figure 4: Residual plots for linear model based on demographic information and store type

```
Residuals:
    Min      1Q  Median      3Q     Max
-31.423  -5.342  -1.365   3.438  49.314

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       2.785e+01  1.488e+00  18.711  < 2e-16 ***
HOUSEHOLDS                       -2.527e-03  7.074e-04  -3.572 0.000355 ***
OCCUPATION_PROF_TECHNICAL        -3.854e-04  1.364e-04  -2.825 0.004738 **
OCCUPATION_RETIRED               -3.416e-03  4.185e-04  -8.163 3.43e-16 ***
EDUCATION_BACHELOR_DEGREE        -5.330e-04  1.628e-04  -3.275 0.001060 **
EDUCATION_LESS_THAN_HIGH_SCHOOL  -8.958e-04  1.016e-04  -8.815  < 2e-16 ***
AGE_19                           -4.300e-03  1.492e-03  -2.882 0.003958 **
AGE_21                           -3.848e-03  8.359e-04  -4.604 4.17e-06 ***
AGE_25_TO_29                     -2.015e-03  5.912e-04  -3.408 0.000656 ***
AGE_30_TO_34                     -1.876e-03  6.479e-04  -2.896 0.003787 **
AGE_35_TO_39                     -1.315e-03  4.718e-04  -2.787 0.005323 **
AGE_40_TO_44                     -4.101e-03  4.538e-04  -9.037  < 2e-16 ***
AGE_50_TO_54                     -2.774e-03  5.234e-04  -5.300 1.17e-07 ***
AGE_60_TO_64                     -3.462e-03  6.973e-04  -4.965 6.92e-07 ***
AGE_65_TO_69                     -3.215e-03  7.904e-04  -4.068 4.75e-05 ***
GENDER_MALE                       1.839e-03  2.071e-04   8.876  < 2e-16 ***
GENDER_FEMALE                     1.306e-03  1.686e-04   7.749 9.60e-15 ***
MARITAL_STATUS_YES                3.000e-04  8.008e-05   3.746 0.000180 ***
INCOME__UNDER__15000              3.235e-03  7.101e-04   4.556 5.25e-06 ***
INCOME__15000_TO__24999           5.686e-03  8.215e-04   6.922 4.56e-12 ***
INCOME__25000_TO__34999           3.182e-03  7.659e-04   4.155 3.26e-05 ***
INCOME__35000_TO__49999           1.953e-03  7.267e-04   2.687 0.007213 **
INCOME__50000_TO__74999           3.041e-03  6.850e-04   4.439 9.08e-06 ***
INCOME__75000_TO__99999           3.322e-03  7.038e-04   4.720 2.37e-06 ***
INCOME__100000_TO__124999         3.803e-03  8.663e-04   4.389 1.14e-05 ***
INCOME__125000_TO__149999         4.218e-03  1.135e-03   3.717 0.000202 ***
INCOME__150000_TO__174999         3.217e-03  7.286e-04   4.416 1.01e-05 ***
INCOME__200000_TO__249999         4.061e-03  8.698e-04   4.668 3.05e-06 ***
INCOME__250000_                   4.054e-03  8.338e-04   4.862 1.17e-06 ***
HOMEOWNER                        -3.229e-04  9.694e-05  -3.331 0.000866 ***
DWELLING_TYPE_SINGLE_FAMILY_SFD   4.653e-04  1.644e-04   2.831 0.004649 **
DWELLING_TYPE_MULTI_FAM_CONDO     1.192e-03  1.973e-04   6.041 1.55e-09 ***
LENGTH_OF_RESIDENCE_1_YR          9.971e-04  3.270e-04   3.049 0.002299 **
DWELLING_SIZE___2_UNITS          -7.127e-04  1.333e-04  -5.347 9.03e-08 ***
DWELLING_SIZE___3_UNITS          -7.148e-04  1.362e-04  -5.248 1.55e-07 ***
DWELLING_SIZE___4_UNITS           1.063e-03  3.050e-04   3.487 0.000490 ***
DWELLING_SIZE___10_TO_19_UNITS   -4.793e-04  1.468e-04  -3.266 0.001094 **
DWELLING_SIZE___20_TO_49_UNITS   -5.408e-04  1.704e-04  -3.173 0.001510 **
MVH___UNDER__101000              -8.121e-05  2.731e-05  -2.974 0.002946 **
SUMMARIZED_AREA_LVL_STATISTICS_A  5.910e-01  1.389e-01   4.255 2.10e-05 ***
SUMMARIZED_AREA_LVL_STATISTICS_B -4.022e-02  1.053e-02  -3.820 0.000134 ***
SUMMARIZED_AREA_LVL_STATISTICS_D  7.923e-05  2.528e-05   3.134 0.001724 **
SUMMARIZED_AREA_LVL_STATISTICS_E -1.159e-01  1.401e-02  -8.275  < 2e-16 ***
SUMMARIZED_AREA_LVL_STATISTICS_G  5.829e-02  1.704e-02   3.421 0.000624 ***
SUMMARIZED_AREA_LVL_STATISTICS_H -7.075e-01  1.382e-01  -5.121 3.06e-07 ***
NUMBER_OF_VEHICLES_IN_ZIP_ZIP_4   2.605e-01  5.261e-02   4.952 7.41e-07 ***
NUMBER_OF_USED_CARS              -1.912e-01  3.755e-02  -5.092 3.57e-07 ***
NUMBER_OF_USED_LIGHT_TRUCKS      -4.725e-01  9.829e-02  -4.807 1.54e-06 ***
AVG_VEHICLE_MSRP                 -3.525e-04  6.304e-05  -5.592 2.26e-08 ***
AVG_CURRENT_VEHICLE_RETAIL_VALU   5.724e-04  1.099e-04   5.209 1.92e-07 ***
X1                               -1.269e+01  1.000e+00 -12.691  < 2e-16 ***
X2                               -1.881e+01  9.803e-01 -19.186  < 2e-16 ***
X3                               -8.672e+00  1.075e+00  -8.064 7.69e-16 ***
X4                               -1.411e+01  1.098e+00 -12.848  < 2e-16 ***
X6                               -1.740e+01  1.146e+00 -15.186  < 2e-16 ***
X7                               -1.216e+01  9.749e-01 -12.475  < 2e-16 ***
X8                               -1.732e+01  4.468e+00  -3.877 0.000106 ***
X9                               -1.996e+01  8.778e+00  -2.274 0.022991 *
X10                              -2.043e+01  1.677e+00 -12.189  < 2e-16 ***
X11                               6.982e+00  1.031e+00   6.772 1.30e-11 ***
X12                              -1.256e+01  1.192e+00 -10.538  < 2e-16 ***
X13                              -1.849e+01  1.110e+00 -16.655  < 2e-16 ***
X14                              -2.302e+01  8.779e+00  -2.622 0.008743 **
X15                              -1.755e+01  9.898e-01 -17.734  < 2e-16 ***
X16                              -1.707e+01  6.249e+00  -2.731 0.006316 **
X17                              -7.608e+00  1.045e+00  -7.279 3.46e-13 ***
X18                              -5.067e+00  9.769e-01  -5.187 2.16e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.722 on 25009 degrees of freedom
Multiple R-squared: 0.3402, Adjusted R-squared: 0.3385
F-statistic: 195.4 on 66 and 25009 DF,  p-value: < 2.2e-16
```

Figure 5: Linear model based on demographical information and store types($X1$-$X18$ are categorical parameters for store type 1-18)

5 Other Approaches

The great success of the genie approach encouraged us to look for more efficient methods to partition sales. One method we came up with, was to learn a supervised classification problem of estimating say $sales > 20$, and then feed this value into linear method to get closer to genie results. We used logistic regression to estimate $sales > 20$. This method gave results similar to the final model in the previus Section.

Furthermore, We tried to come up with some non-linear model including powers of parameters and log of some parameters like Number of Households and Number of Individuals, and none of them did as good as the model we derived in the previous Section.

References

[1] S. Chatterjee and A. S. Hadi, *Regression Analysis By Examples*. Wiley, New Jersey, 2006.

[2] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li *Applied Linear Statistical Models*. McGraw-Hill, New York, 2005.

[3] T. L. Lai and H. Xing *Statistical Models and Methods for Financial Markets*. Springer, New York, 2008.