IEMS 462: COURSE PROJECT

PREDICTING NPO'S PAST DONOR CONTRIBUTION

Yiwen Hu Taorui Peng

Northwestern University

June 4, 2017

Executive Summary

This report aims at predicting the non-profit organization's Fall 2010 donation amount using past donors' contribution records between October 2001 and June 2010. Two specific questions are investigated: who will donate and how much will they contribute.

Through various predictive models, we have found that a past donor's likelihood to contribute in Fall 2010 tends to be negatively correlated to their most recent donation amount, and the time length since their latest contributions. In addition, self-identified males and married couples are more likely to donate than female-identified donors.

As for predicting specific donation amount, we found that donors' latest contribution amount always serves as a good positive indication of their Fall 2010 donation level. In addition, donors who have not donated for a long time since their last contribution tend to donate more. And interestingly, evidence showed that number of solicitations received by a donor does not necessarily translate into higher probabilities of future donation, nor the contribution size.

As a result of our models, we select top 1000 expected donors with their actual donation to be $10,160.05. Based on their past donation records, we recommend the management to focus on maintaining good relationships with long-time loyal donors, whose contribution would make a majority of the organization's donation income, and schedule solicitations appropriately to those fresher and less-established donors, keeping them tuned but not too frequently.

Contents

1. Introduction

Given the donation record database of a non-profit organization, we plan to answer two basic questions through statistical learning: 1) based on past donation behavior, which donors are more likely to donate during Q4 2010; 2) among those who did donate in Fall 2010, which part of their donation history could effectively explain the variations in their dollar amount contribution. After exploring and analyzing different models, we assess their predictive power on the test set, which were not ever used during model training, hence constituting a fair ground for model validation.

The approach we adopt here in this project is purely empirical, hoping that employing logistic regression and multiple regression model will help us make good predictions. Intuitively, when assessing the potential whether a donor will make a donation, we assume that it's correlated with the amount of the latest few donations, time elapse since latest donations, gender, regions and maybe other predictors. Particularly, when measuring the amount of target donation, we assume that the target amount of donation will have a higher correlation with the last three donation amount and the average donation amount in a donor's lifetime.

The rest of the report is organized into four parts: Section 2 – Model Fitting, Section 3 – Model Validation, Section 4 – Conclusions, and the Appendix.

Section 2, the core of our report, details how we build up the prediction models. It is presented with three subsections: Exploratory Analysis and Data Preprocessing, Logistic Regression, Multiple Regression. In Exploratory Analysis and Data Preprocessing section, we examine the nature of data by generating plots for univariate and bivariate variables. Particularly, we also explain how we decide not to use a single model to predict the test set but to come up with the idea to separate data into three subsets and fitting models for each of them. Therefore, the readers are expected to see total six models for predicting three sub datasets. In Logistic Regression, we would explain the overall procedure and the details of how we conduct predictors selection, interactions, model fitting, correct classification rate, model diagnosis. In Multiple Linear Regression Model, in addition to similar procedures of logistic regression, we also do data transformation, such as log and square root. Section 3 tests our models built in Section 2 on the validation set, including many key performance measures. Section 4 concludes the whole investigation by noting important discoveries, comments, and directions for further improvement.

2. MODEL FITTING

2.1 Exploratory Analysis and Data Preprocessing

Before delving into the core modeling and analysis part of this project, we first examine the dataset, identify practical issues which may arise and become obstacles later on in the model building stage, and try to address them via reasonable and intuitive pre-processing transformations.

At beginning, we do a univariate exploration of the data. Since the response that we care about is the target donation dollar, we first plotted the target donation dollar - `TARGOL`. It can be noticed from Appendix Figure1 that there's one donation for $1500, which is much higher than other donations. We would omit this data point as one of the outliers when doing multiple linear regression.

In addition to univariate exploration, we also conduct a bivariate exploration of the data. Considering the target donation amount may have positive correlation with the average donation amount. We plot the two variables in Appendix Figure2, which indeed shows us a sign of positive correlation. We also hypothesize that the largest donated amount have some correlation with the largest amount donated `CONLARG` and we observe Appendix Figure 3 for intuition. We continue to do similar exploration process for other variables to help us better understand the nature of the data.

After exploratory analysis and a better understanding of the data, we shift our attention to address the problem of its vast missing data. Most missing values stem from variables related to a donor's 2nd and 3rd latest contribution record, as over two thirds of them (67,402 out of 99,200 total observations; 44,941 out of 66134 training samples) have only one or two donating records over the 10-year course. This varied degree of data missing leads to some non-negligible discrepancies between those different segments —- for example, if we look at data samples that correspond to donors with only one contribution record in their entire lives, then their latest contribution amount (`CNDOL1`), lifetime contribution (`CNTRLIF`), largest contribution (`CONLARG`), and first contribution amount (`CONTRFST`) are all identical, hence essentially conveying just one piece of information, i.e. *the first-time contribution dollars*. Thus it is clear that the mentioned difference above is fundamental in such a way that the intrinsic data dimensions present and available for use in predictive models have severely shrink.

On the other hand, based on business intuition, one would expect that this group of relatively newer and fresher donors having only a couple of donation records on file preserve a higher level of uncertainty toward their future willingness to donate and the monetary contribution potentials, as much less is known about them compared to their long-time loyal counterparts. So in response to this concern, we consider one natural solution, namely, to divide all data samples into three groups:

1. People that have donated only once in the past (`Once`);

2. Donors that have precisely two contribution records (`Twice`);

3. Loyal donors who have contributed at least three times (`Loyal`).

In the training stage, we fit models to each group separately, which to a large extent eliminates much need on manual cleaning for data-missing samples. Correspondingly, when we validate models on the test set, each of them gets applied to the respective group in the validation set as well. As a result we would have three models to predict the results of the separated test sample data sets.

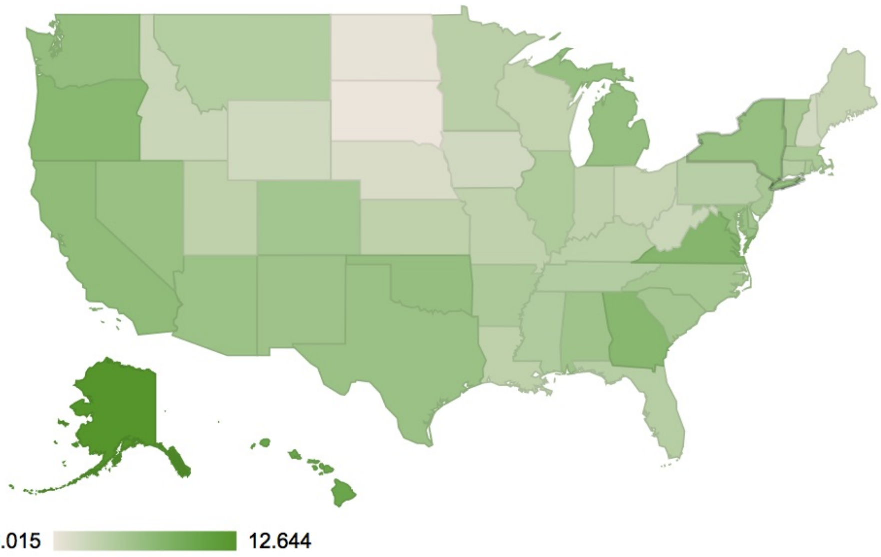
Table 1 briefly summarizes the number of observations categorized into each of these groups on both training and validation sets. As can be seen from the statistics, one-time donors and loyal donors take up the majority, so in actual model building, we may focus our learning efforts on the `Once` and `Loyal` groups, while modeling for the `Twice` group can be built upon experiences gained from learning the previous two. Intuitively, we would expect that a successful predictive model for two-time donors should inherit some attributes from the good models fitted for the `Once` group, as they both target people that have much less revealed information available to this non-profit organization.

	Training		Validation	
Group	All	`TARGDOL>0`	All	`TARGDOL>0`
`Once`	39,655	8,990	19,770	4,480
`Twice`	5,286	1,470	2,691	754
`Loyal`	21,193	7,679	10,605	3,835
Total	66,134	18,139	33,066	9,069

Table 1: Summary of Group Division by Times of Past Contribution

In addition, given information of the Contribution Code and Solicitation Code, we create new columns named `cnc1, cnc2, cnc3, sol1, sol2, sol3` to reflect the respective category (`A, B, C, D, or M`) that each code belongs to, using `dmef1code.csv`.

Another main concern is that the `STATE` variable has too many levels to be a potentially useful categorical predictor. So if we want to incorporate this geographic information into our modeling, one inevitable question is how to synthesize them into just several categories in a way that has practical meaning such that useful patterns/structures be largely preserved. For this project, we consider two classifications: 1) based on Census 2) by median household income level. Both data are downloaded from U.S. government website.



Data: statMean[c(2:12, 14:54),] • Chart ID: GeoChartID262366f21c1 • googleVis-0.6.2
R version 3.3.2 (2016-10-31) • Google Terms of Use • Documentation and Data Policy

Figure 1: Color Map of Lifetime Mean Contribution Size per Donor Averaged on State-Level

We also re-group `SEX` into four main categories: female, male, married couple, and others (`U` and `C` combined); Since we don't know what C represents, we would replace C with U. Also, we use female as the reference level in most models.

Deemed as a potential predictor variable, we also create a derived categorical variable – `solTime` – by counting how many solicitations had been sent to each donor during the 10-year course by looking at solicitation codes `SLCOD1`, `SLCOD2`, and `SLCOD3`. Donors having three complete codes are labeled as one group, and those that had been solicited just once or twice each form a group.

2.2 Predict Donating Probabilities: Logistic Regressions

To assess the probability of donation in the test samples, we will employ binary logistic regression model. We choose this model because we are interested in measuring a response which has two values, which are whether a donor will donate or not donate. Binary logistic regression model satisfies our requirement, because it has binary response as 0 and 1. Before building up the model, we need to create a new label, which is a new column in the dataset named `TARGDOL_BI`. We denote `TARGDOL_BL` = 1 to indicate that the donor will donate(`TARGDOL_BL` > 0), and `TARGDOLBL` = 0 to indicate that the donor will not donate(`TARGDOL_BL` = 0). After creating this new label, we will have all the predictors needed to build our binary logistic regression model using `TARGDOL_BL` as our binary response.

We would conduct some analysis before fitting the model using all the predictors available. Since we have already separated the datasets into three, we would employ different strategies for each of them. For example, on one hand, when we are fitting model 1, which only relates to the donors who have donated only once, we would not consider predictors that related to donors who donated twice or more. So we would not use `CNDOL2, CNCOL3, CNMON2, CNMON3, CNDAT2, CNDAT3, cnc2, cnc3`. On the other hand, when we are fitting model 3, we would consider these predictors. In addition, we notice that there's a predictor for donors' ID. We regard this as a random factor so we will not consider this predictor into our models. As instructed, we also know that ĈNMON and CNDAT will give us the same information. Considering the number of of date is less relevant than the time length between two donations. So we would use ĈNMON only to incorporate the time factor.

Let's begin by building the model for training set 1. We know from our classification that training set 1 contains the donors who donated only once. Therefore, their Latest Contribution, Dollars Contribution Lifetime, Largest Contribution, First Contribution will all be the same.

Therefore, we will only use `CNDOL1` instead of repetitively all of same other information. Moreover, any predictors related to their 2nd or 3rd contribution will also be 0 as well. Knowing the facts above will help us simplify our model building process.

We would employ backward stepwise method to select candidate models for training set 1. So we first build a full mode, which will include non-repetitive predictors, such as `CNDOL1, SEX, REGION, CNDAT1, cnc1, sol1, sol2, sol3`. Please refer to appendix for Model Summary [x]. We have noticed that there are many data points eliminated due to missing values in our full model. This is due to the nature of the data. We decide not to manually replace missing data with other entries to increase bias. Observing the P values, we first find that `region, incZone` are highly insignificant in our model because majority of their P values are close or greater than 0.5. So we first remove `REGION, incZone` from out model and run with the rest of predictors again. In this new model, we first notice that AIC has decreased which means the model is better than our full model. Here we would conduct likelihood ration test. We assess the goodness of fit by comparing the Deviance with chisqure (n-p-1), overall significance test $9669.4 - 9539.3 = 130.1 > 37.65$. So it passes the overall significance test. We did this trials and errors several times and based on the lowest AIC value, we choose the model

$$\texttt{TARGDOL\_BI} \sim \texttt{CNDOL1} + \texttt{CNMON1} + \texttt{SEX} + \texttt{cnc1} + \texttt{sol1}$$

We begin to calculate accuracy our built model 1. By plotting ROC curve, we could find roughly find the AUC, and pick our cutoff rate p* to be 0.4. Applying $predict$ and our model back to training dataset 1, We get the following confusion matrix

Thus our calculated correct classification rate is

$$\frac{30511 + 80}{30551 + 114 + 8910 + 80} = 77.24\% .$$

Confusion	FALSE	TRUE
0	30551	114
1	8910	80

After selecting training model 1, we begin to build model for training set 2. Training dataset 2 has more predictors than training dataset 1. Similar to the process of building the first one, we would employ backward stepwise procedure to select our best candidate model. First we add all the predictors and try a full model. The full model involves predictors as follows: CNDOL1, CNDOL2, CNMON1, CNMON2, CONLARG, CONTRFST, SEX, cnc1, cnc2, cnc3, sol1, sol2, sol3, incZone. Observing the P values from statistics summary, we find out that regions and income level zones are highly non-significant. So at the first step, we would eliminate these two predictors. Eliminating the region and incZone, we observe the new P values of predictors. Using similar thought process as before, we eliminate sol1 and sol3. In addition, we hypothesize that CNDOL1 and CNDOL1 will an important role in fitting our model but both of them show insignificance. So here we would introduce an interaction by adding a new column CNDOL1 * CNDOL2. Employing this new predictor, after trying adding and eliminating predictors, we find out that our best model for training set 2 is

$$\texttt{TARGDOL\_BI} \sim \texttt{CNDOL1} + \texttt{I(CNDOL1} * \texttt{CNDOL2)} + \texttt{CONLARG} + \texttt{CONTRFST} + \texttt{SEX} + \texttt{CNMON1} + \texttt{sol2}$$

Next we will verify our assumption by measuring correct classification rate using training dataset 2. We first plot its ROC curve and find its cutoff probability rate to be 0.4. Choosing the cutoff rate, we implement our model to dataset 2, and get the confusion matrix.

The correct classification rate is

$$\frac{3190+570}{3190+604+890+570} = 71.51\% \,.$$

Confusion	FALSE	TRUE
0	3190	604
1	890	570

Now we switch our focus to fitting model for our training dataset 3. Different from the training dataset 1 and 2, this dataset contains the donors who have donated three or more times. So we would include `CNTMLIF`. Again, we would implement the backward stepwise procedure. By observing the P values and AIC, we could first eliminate regions. Similarly we do this process for a few more steps. When building model 3, we are satisfied to find that `CNMON1, CNMON2, CNMON3, cnc1, cnc3` are more significant than before. To achieve a lower AIC and getting a more significant P values, we add interactions for `CNDOL1 * CNDOL3, CNDOL2 * CNDOL3` . We also eliminate sol1, sol2 by comparing AIC. For training dataset 3, we come up with the model as follows

$$\begin{aligned}\texttt{TARGDOL\_BL} \sim \texttt{CNDOL1} + \texttt{CNDOL2} + \texttt{CNDOL3} + \texttt{I(CNDOL1} * \texttt{CNDOL2)} + \texttt{I(CNDOL2} * \texttt{CNDOL3)}\\ \texttt{CNTMLIF} + \texttt{CONLARG} + \texttt{CNMON1} + \texttt{CNMON2} + \texttt{CNMON3} + \texttt{cnc2} + \texttt{cnc3} + \texttt{sol3}\end{aligned}$$

We verify our model 3 by measuring its correct classification rate. By simply applying $predict$ function to training dataset 3 and tabulate the results, we first plot the ROc curve and choose appropriate cut off rate 0.45. Then we get the following confusion matrix.

Confusion	FALSE	TRUE
0	10457	3057
1	3209	4470

We achieve an correct classification rate

$$\frac{10457 + 4470}{10457 + 3057 + 3209 + 4470} = 70.43\%\,.$$

2.3 Predict Contribution Amount: Multiple Regressions

To train models on each donor group, we always consider *multiple linear regressions* as the starting point, and then go on to diagnostics, propose relevant transformations on responses and predictors, and finally to iteratively refine the model specifications for a few more times and determine a couple of good candidate models ready for testing. The detailed procedure followed in this model training process is outlined below:

1. multiple linear regression including all potentially useful predictors from the dataset;
2. stepwise regression (backward elimination) to narrow down predictors to a smaller subset;
3. re-fit using variables identified in Step 2 and detect outliers;
4. screen outliers manually and determine on those that deserve exclusion;
5. run diagnostics for the new model fitted using the subsample free of severe outliers (such as Residuals vs Fitted plot, AV/CR plots, multicollinearity test, etc.);
6. if necessary, propose variable transformations on the response, predictors or both, e.g. log transformation and other general power transformations;
7. re-fit several models using transformed variables, through linear regressions, GLM of non-Gaussian families, or penalized regressions like Ridge and Lasso when needed;
8. re-run diagnostics on new fitted models including influential observation tests;
9. make discretionary decisions on which influential observations to leave out based on Cook's distances, and repeat some of the Steps in 2-7.

Group Once

To illustrate this, we briefly show the model building work-flow and results for the training on Once group. As already mentioned, we first consider a fully linear multiple regression model incorporating almost all useful variables from the training group, i.e.

Model 1.1 (full linear): $\texttt{TARGDOL} \sim \texttt{CNDOL1} : \texttt{CNCOD1} + \texttt{CNMON1} + \texttt{SEX} + \texttt{region} + \texttt{incZone} + \texttt{solTime}\,.$

After backward elimination procedure performed on this full model, variables CNDOL1 : CNCOD1, CNMON1, SEX, and incZone are retained. Then, after examining residuals, we try excluding a handful of outliers (Row ID 28, 32, 34, 35, 37 and 146) to test other linear regression assumptions. Figure 2 shows an Added-Variable plot for the main interaction term CNDOL1 : CNCOD1 in the new model, from which we actually see that two of the four contribution categories consist of not more than half a dozen observations, and code levels A and D don't differ from each other a lot given their estimated coefficients being 0.83 and 0.82 respectively, both with numerically zero p-values. Hence, we consider eliminate the interaction term in modeling one-time donors, and add region back to the model as it would now produce some significant coefficients.

Figure 3 (a) shows the Residuals vs Fitted plot for the revised model: TARGDOL ~ CNDOL1 +CNMON1 +SEX +incZone +region, which implies some sort of heteroscedasticity, with slight signs of concavity in the middle part of the plot. Thus we modify the original main term to a log-transformed one $\ln$ CNDOL1 to introduce concavity. After that, we screen a series of **Box-Cox power transformations** (i.e. $y \mapsto \left(y^{\lambda} - 1\right) / \lambda$) under different values of λ on the response variable, to see if any more need on transforming TARGDOL arises. The results are portrayed in Figure 3 (b).

From this chart, we see that the best power is about $1/9 \approx 0.11$, which is fairly close to 0; hence we can try both $9\left(\texttt{TARGDOL}^{1/9} - 1\right)$ and $\ln$ TARGDOL as our new responses, and get the following

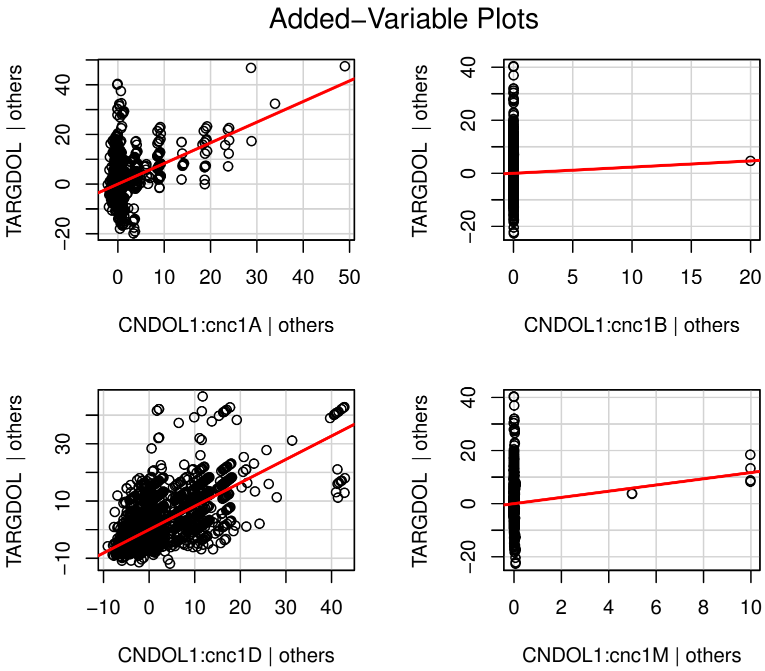


Figure 2: AV Plot for `CNDOL1:CNCOD1` in Model `TARGDOL ~ CNDOL1:CNCOD1 +CNMON1 +SEX +incZone` on the `Once` Group. (somewhat indicating the non-necessity of using interaction here)

two models:

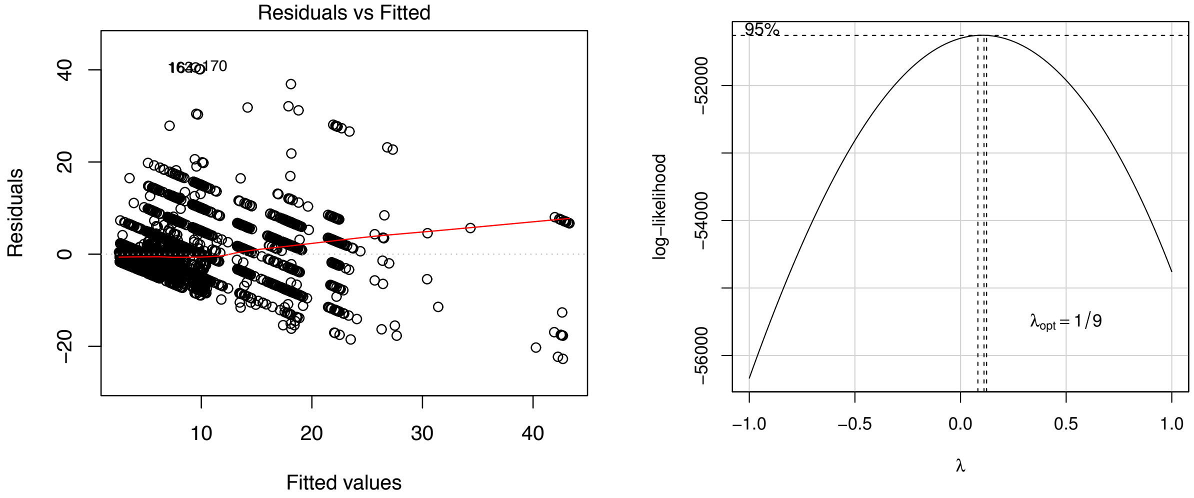
$$\text{Model 1.2 (Box-Cox):}\quad \frac{\mathtt{TARGDOL}^{1/9}-1}{1/9} \sim \mathtt{CNDOL1} + \ln(\mathtt{CNDOL1}) + \mathtt{CNMON1} + \mathtt{SEX} + \mathtt{region}\,;$$

$$\text{Model 1.3 (log-log):}\quad \ln(\mathtt{TARGDOL}) \sim \ln(\mathtt{CNDOL1}) + \mathtt{CNMON1} + \mathtt{SEX} + \mathtt{region} + \mathtt{incZone}\,.$$

We fit Models 1.2 and 1.3 using standard linear model and GLM with Poisson family, respectively, and get pretty decent results as shown in Figures 7(a)-(b) in the Appendix.

Group `Twice`

The group of people that had donated exactly twice is a relatively small segment, as compared to those who donated only once and more loyal and long-time donors. And modeling for this



(a) Residuals vs Fitted **(b)** Log-Likelihood vs $\lambda \in (-1, 1)$ under Box-Cox

Figure 3: Residuals vs Fitted Plot (left) & Box-Cox Transformation Screening Chart (right) for the `Once` Group. (note: the LHS model is `TARGDOL ~ CNDOL1 +CNMON1 +SEX +incZone +region`, while the RHS screening is done under `TARGDOL ~ log(CNDOL1) +CNMON1 +SEX +incZone +region`.)

group of people seems to be more straightforward as well, since most of the time, linear models are already good enough in representing the data.

Following a similar procedure as outlined at the beginning of this section and illustrated for the `Once` group, we construct predictive model 2.1 as shown below:

$$\text{Model 2.1 (linear):}\quad \texttt{TARGDOL} \sim \texttt{CNDOL1}:\texttt{CNCOD1} + \texttt{CNDOL2}:\texttt{CNCOD2} + \texttt{CONLARG}^2 + \texttt{region} + \texttt{solTime}\,.$$

We tried using some of the predictors proven to be significant for one-time donors, but these turned out to be not quite significant in this group of donors — such variables include time since latest and 2nd latest contributions, sex, and income zone. Past solicitation codes also seem unimportant in modeling contribution size among these people.

On the other hand, we include the squared term of largest contribution into the model, and the estimated coefficient yielded to be highly significant. This was done to help signal those who may have large potentials to become important donors in the future, and we did see that the adjusted R^2 improved as well after incorporating this variable. Please see the detailed regression output attached in the Appendix (Figure 8(a)), as well as the Residuals vs Fitted plot (Figure 8(b)).

Some key take-aways from modeling results on this group are:

- A linear combination of the donors' two contribution dollar amount records could to a large extent predict their donation size in Fall 2010;
- Compared to donors who were previously solicited for only once, those that got two solicitations tend to donate about \$2 less for Fall 2010, significant at 10% with a p-value of 0.086.

Group `Loyal`

In modeling loyal donors, we have more useful variables to utilize, hence, in some sense, we need more care and caution in order to handle these additional information wisely.

Two important predictors we derived from the original dataset to adopt in our predictive models are:

1. `AVG`: the lifetime average contribution amount, defined as `CNTRLIF/CNTMLIF`;
2. `vip`: the "high potential" donors, selected based on whether one's largest donation amount exceeds \$100 or not. Those exceeding \$100 are labeled as `vip` donors.

Hence, the model we determined as the best for our loyal donors is as follows:

$$\text{Model 3.1 (linear):}\quad \texttt{TARGDOL} \sim \texttt{AVG} + \texttt{vip} + \texttt{CNDOL1} : \texttt{CNCOD1} + \texttt{CNDOL2} : \texttt{CNCOD2}$$

$$+\texttt{CNDOL3}:\texttt{CNCOD3}+\sqrt{\texttt{CONLARG}}+\texttt{CNMON1}+\texttt{SLCOD2}\,.$$

Please see detailed regression output and the corresponding residual plot in Figures 9(a)-(b) of the Appendix.

Although contribution dollar amount records tend to correlate with each other, as well as with AVG, estimated coefficients are still highly significant for most of the predictors included in Model 3.1. Indeed, when we use Ridge regression under the same specification as presented in Model 3.1, 100 runs of 10-fold cross-validations on the training set gives us a best penalizing constant λ (chosen to be the one yielding the lowest average MSE overall) valued at around 0.03. Thus, it implies that the need for adding a penalizing term to the original linear model is actually minimal, indicating to some degree that multicollinearity is not a big problem for this model.

Some key lessons learned from this model are:

- Again, a linear combination with decreasing weights of the latest three contribution amount can explain variations in donation amount for Fall 2010 to a great deal (adjusted R^2 already exceeding 0.79 when these interaction terms are sole predictors);

- Past average donation size positively influences a donor's Fall 2010 contribution amount; to be specific, for every one dollar's increase on AVG, TARGDOL is expected to increase by 23 cents;

- VIP donors have a significant big margin in their Fall 2010 dollar contribution potential, i.e. expected to donate about $21 more on average.

3. Model Validation

3.1 Yes/No Validation

We conduct validation for our logistic regression model, we first apply logistic model 1 to our testing dataset 1. Applying cutting off probability 0.4, we could get our confusion matrix as follows

Confusion	FALSE	TRUE
0	15225	65
1	4441	39

The correct classification rate rate is

$$\frac{15225+39}{15225+65+4441+39} = 77.21\%$$

Applying logistic model 2 to our testing dataset 2, using cutoff probability 0.4, we get our confusion matrix as follows.

Confusion	FALSE	TRUE
0	1598	330
1	459	287

Similarly, We find out that correct classification rate for test dataset 2 is

$$\frac{1598+287}{1598+330+459+287} = 70.50\%$$

Applying logistic regression model 3 to testing data set 3 and using cutting off rate 0.45, we get the following confusion matrix.

Confusion	FALSE	TRUE
0	5228	1532
1	1626	2197

Calculating the correct classification rate, we get

$$\frac{5228+2197}{5228+1532+1626+2197} = 70.15\%$$

3.2 Donating Amount Validation

To validate the goodness of our multiple regression results, we isolate test set observations with `TARGDOL`>0, apply Models 1.2, 1.3, 2.1 and 3.1 to these donors in their respect category, and calculate validation Root Mean Square Errors (RMSE). Table 2 presents these results:

Model	Group	Test RMSE (`TARGDOL`>0 Only)
1.2 (Box-Cox)	`Once`	\$4.7347
1.3 (log-log)	`Once`	\$4.6183
2.1 (linear)	`Twice`	\$4.0061
3.1 (linear)	`Loyal`	\$4.6503

Table 2: Validation Results for Multiple Regressions on Each Group for People Who Donated in Fall 2010.

As can be seen from Table 2 that RMSE across all groups for `TARGDOL`>0 observations are all in the range of \$4-5, which looks very solid. In particular, the result for two-time donors is the best, with lowest RMSE at about \$4. For people who had only one past donation record, Model 1.2 & 1.3 perform similarly, with the GLM estimated log-log model yielding an even lower RMSE — so for the combined validation part just below, we use Model 1.3 only to predict Fall 2010 donation amount for the `Once` group on the entire validation set.

3.3 Combined Validation on Actual vs Expected Contribution

Now, we combine both parts of our models, binary logistic classification and multiple regression results, to calculate *expected* donation dollars amount for each sample of data in the validation set as instructed by the assignment statement, i.e. via conditional expectation $\mathtt{E(TARGDOL)} = \mathtt{P(TARGDOL} > 0) \cdot \mathtt{E(TARGDOL|TARGDOL} > 0)$. Table 3 contains all RMSE results on each group separately and for the entire dataset as well. Also presented at the bottom rows are the actual sum of TARGDOL values for 1,000 donors with highest E(TARGDOL); as a reference, we also include the theoretical maximum amount of this measure by summing the highest 1,000 TARGDOL values across the whole validation samples.

Model	Group	RMSE of E(TARGDOL)
1.3 (log-log)	Once	$4.4927
2.1 (linear)	Twice	$4.8382
3.1 (linear)	Loyal	$6.3861
Combined	All Observations	$5.2006
Actual Donation by Most Promising 1,000 People:		$10,160.05
Maximum Possible (sum of highest 1k TARGDOL's):		$23,760.98

Table 3: Combined Validation Results Using $\mathtt{E(TARGDOL)} = \mathtt{P(TARGDOL} > 0) \cdot \mathtt{E(TARGDOL|TARGDOL} > 0)$.

We think that the test results outlined in the table above are decent for two reasons: 1) the RMSE values do not differ too much from corresponding values in Table 2, indicating good extensibility; 2) the payoff calculated by summing actual TARGDOL values for one thousand donors with highest expected donation is almost half the theoretical maximum. Since among those 1,000 most promising donors, 498 didn't actually donate in Fall 2010, we think this payoff result may be further improved by adopting more sophisticated modern techniques in statistical & machine learning to achieve higher classification accuracies, as well as collecting more useful donor attributes and expanding the intrinsic dimensions of the NPO's donor database.

4. Conclusions

For each separated datasets, we have one Logistic Regression and one Multiple Regression Model. They are used to predict the probability for each donor to make a prediction and to predict the amount of target donation respectively. Thus, we have six distinct models which consists of different significant factors.

For dataset 1, which contains the donors who have donated only one time, Logistic Regression model shows that `CNDOL1, COMON1, SEX` are the most significant predictors. Similarly, Multiple Linear Regression shows in addition these three predictors, `region` is also one significant predictor. The value of coefficients of Logistic Regression Model reveals that if a donor donated more at last donation, he is less likely to make another donation this time. Also, if the time length from this coming donation and the last donation is longer, a donor is less likely to make a donation. Moreover, a married couple and a male is more likely to make a donation than a female. The value of coefficients of Multiple Linear Regression model gives us another perspective. The donation amount a donor will make can be positively explained by `CNDOL1`, `COMON1`, etc.

For dataset 2, which consists of the donors who have donated twice, Logistic Regression Model reveals that `CNDOL1, CNDOL1*CNDOL2, CONLARG, CONMON1, sol2` are the most significant predictors. It shows that the probability of a donor who will make a donation this time is negatively correlated with the donation amount of last two times separately. However, it is positively correlated with `SEX, sol2`. In Multiple Regression Model, we observe that `CONDOL1,CONDOL2,CONLARG`^{2}`,region,solTime` are the most significant factors. Except for a positive correlation with a `CONDOL1,CONDOL2,CONLARG`^{2}, it's worth noting that target donation amount has a negative correlation with `solTime`, which reveals that the more times that a donor is solicited, the less amount he will donate.

For dataset 3, which consists of the donors who have donated three times or more, Logistic Regression Model shows that `CNDOL1, CNDOL2, CNDOL3, CNDOL2 * CNDOL3, CNTMLIF, cnc1, cnc2, sol3, CNMON1, CNMON2, CNMON3` are the most significant factors. It reveals that the larger amount of the latest 2nd donation and of the latest 3rd donation, the less likely that a donor will make a donation this time. However, for this group of donors, the larger of the the latest donation, the more likely that they will make a donation this time. Also, the longer of the months since the previous donations, the less likely that a donor will be willing to make a donation this time. The contribution code and solicitation code also have a significant relationship with the probability, but since we don't know the meaning of each code, so we will leave that for future analysis. For Multiple Regression model, we find that `AVG`(the average donation amount), `sqrt(CONLARG), vip, sol2, CNMON1, CNDOL1, CNDOL2, CNDOL3` are significant factors. `AVG, CNMON1, CNDOL1, CNDOL2, CODOL3` have positive correlation with the amount of donation. While `sol2,sqrt(CONLARG)` have negative correlation.

After running models and data analysis, we find that our relevant data points are still limited. For example, we lack specific meaning behind each Contribution and Solicitation Code. In addition, we also lack the exact dates of solicitation and donation time because we may want to consider whether weekends or holidays will affect a donor's willingness to make a donation and its amount.

At last, based on calculations done in the validation part, we obtain the root mean square prediction error for `TARGDOL` to be $5.2. Selecting the 1000 donors from the test set who have the highest `TARGDOL` based on our models, we have computed their total actual donations to be $10,160.05.

Appendix

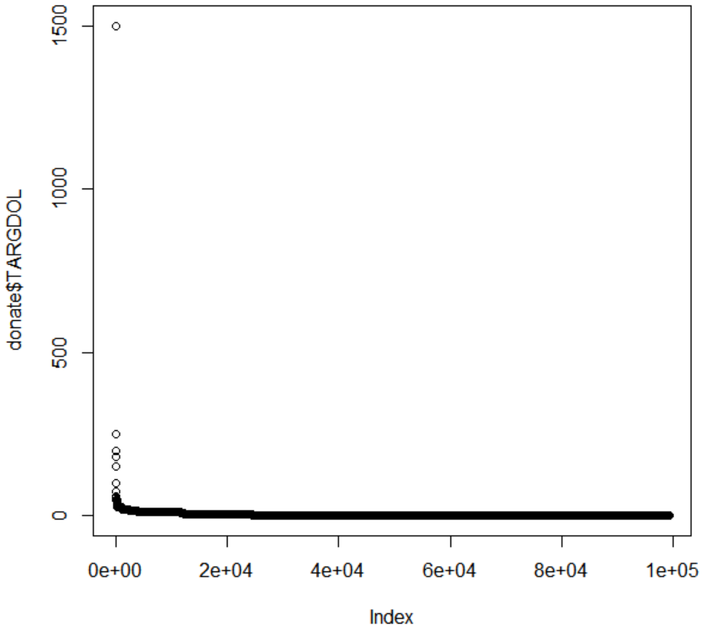


Figure 1: TARGDOL

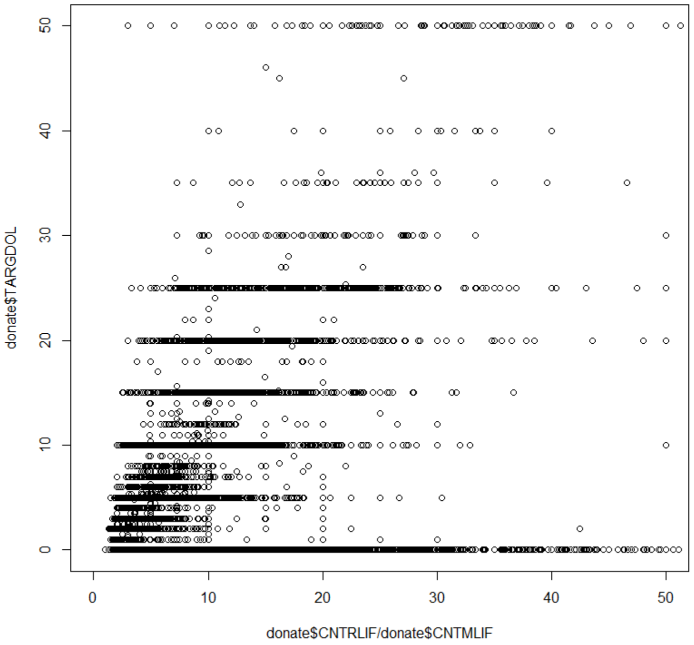


Figure 2: TARGDOL VS CNTMLF

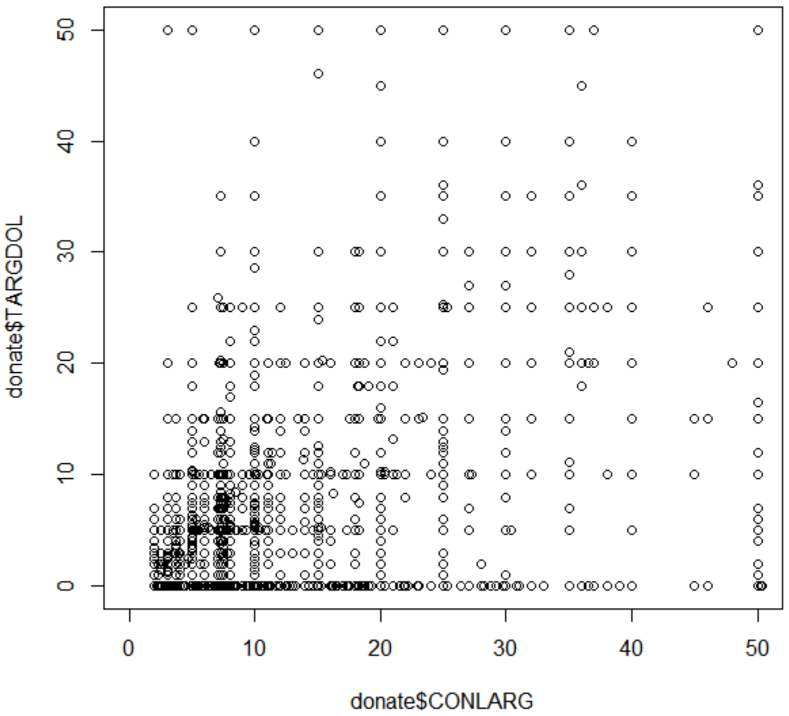
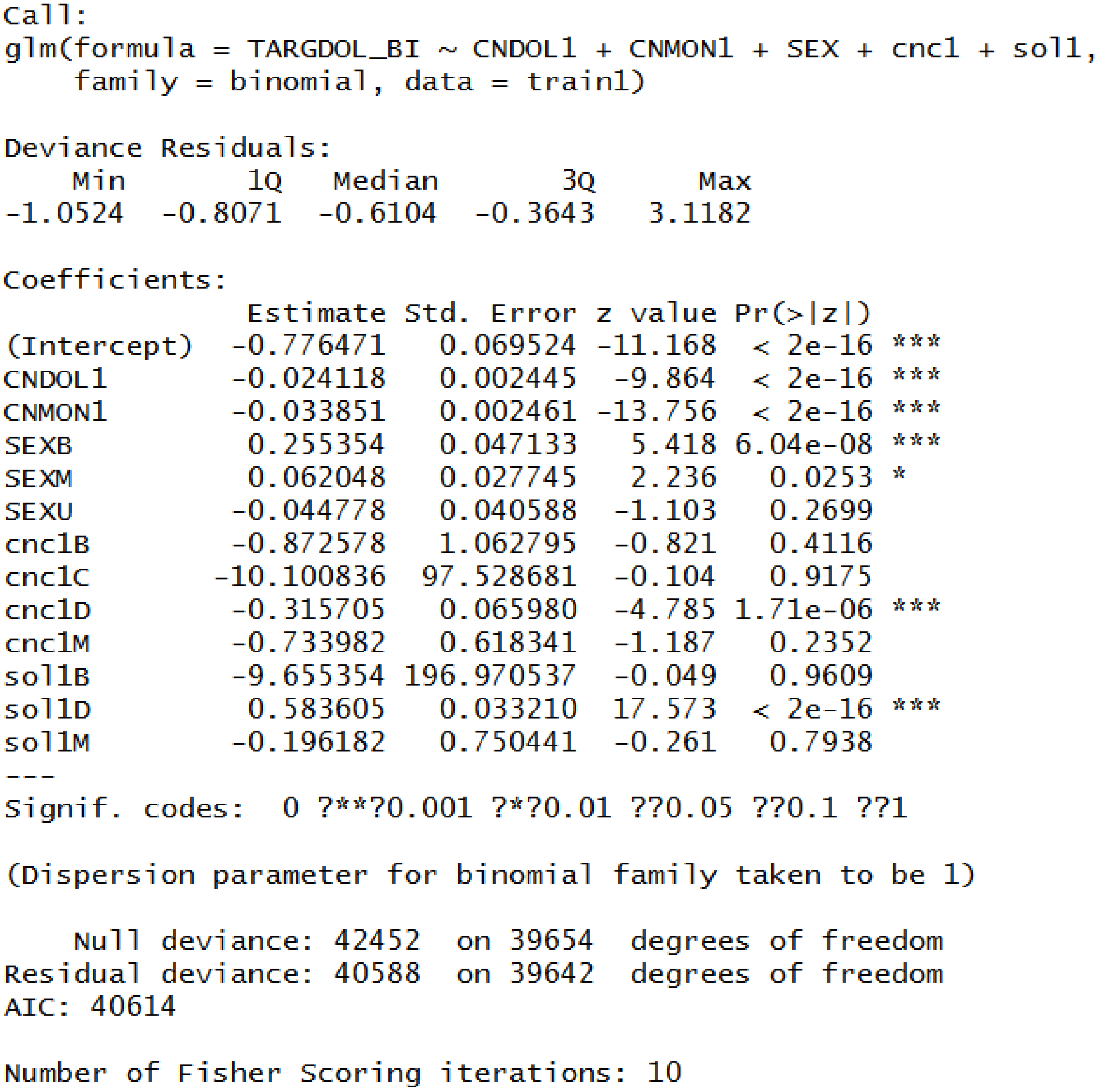
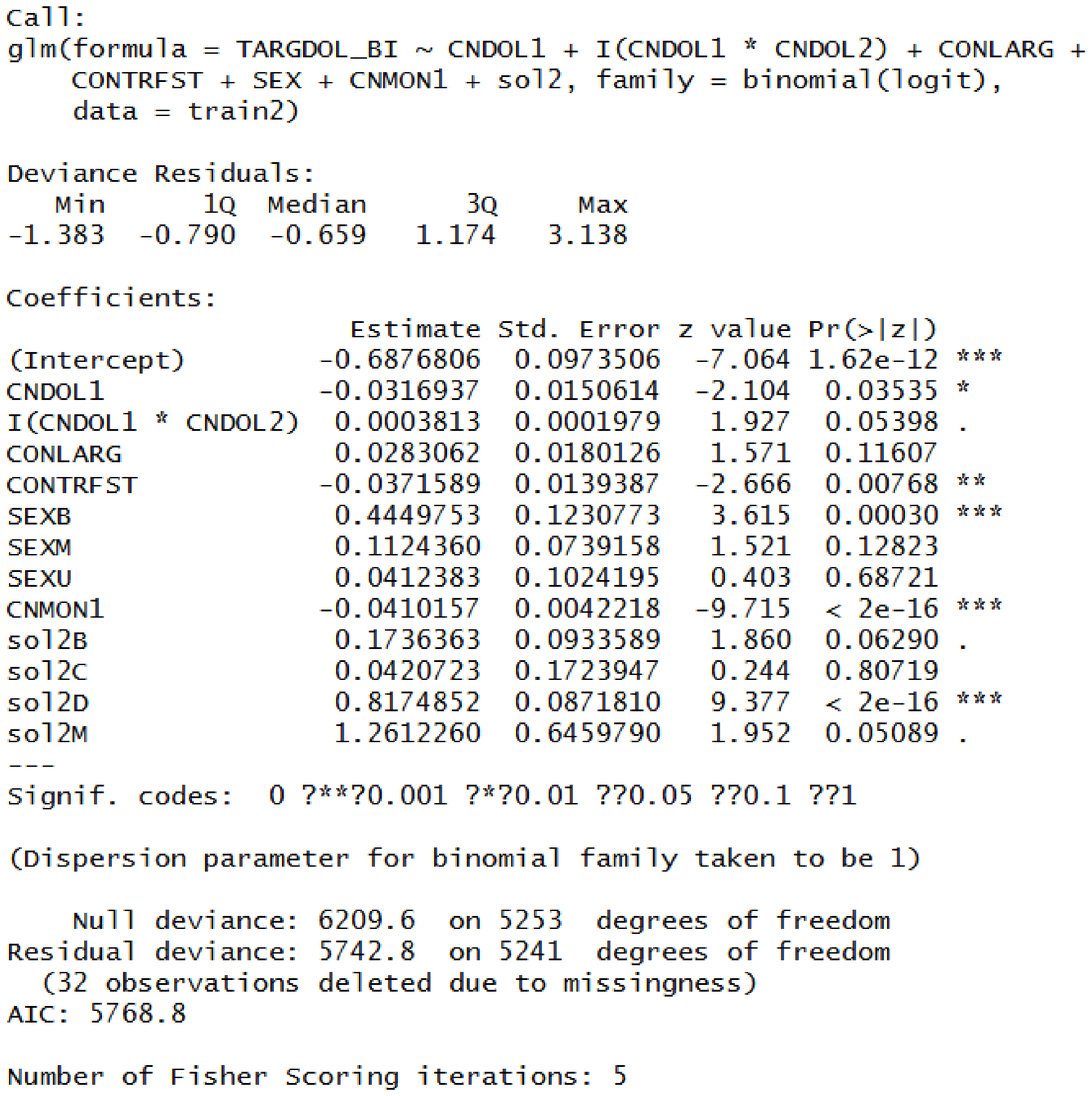


Figure 3: TARGDOL vs CONLARG



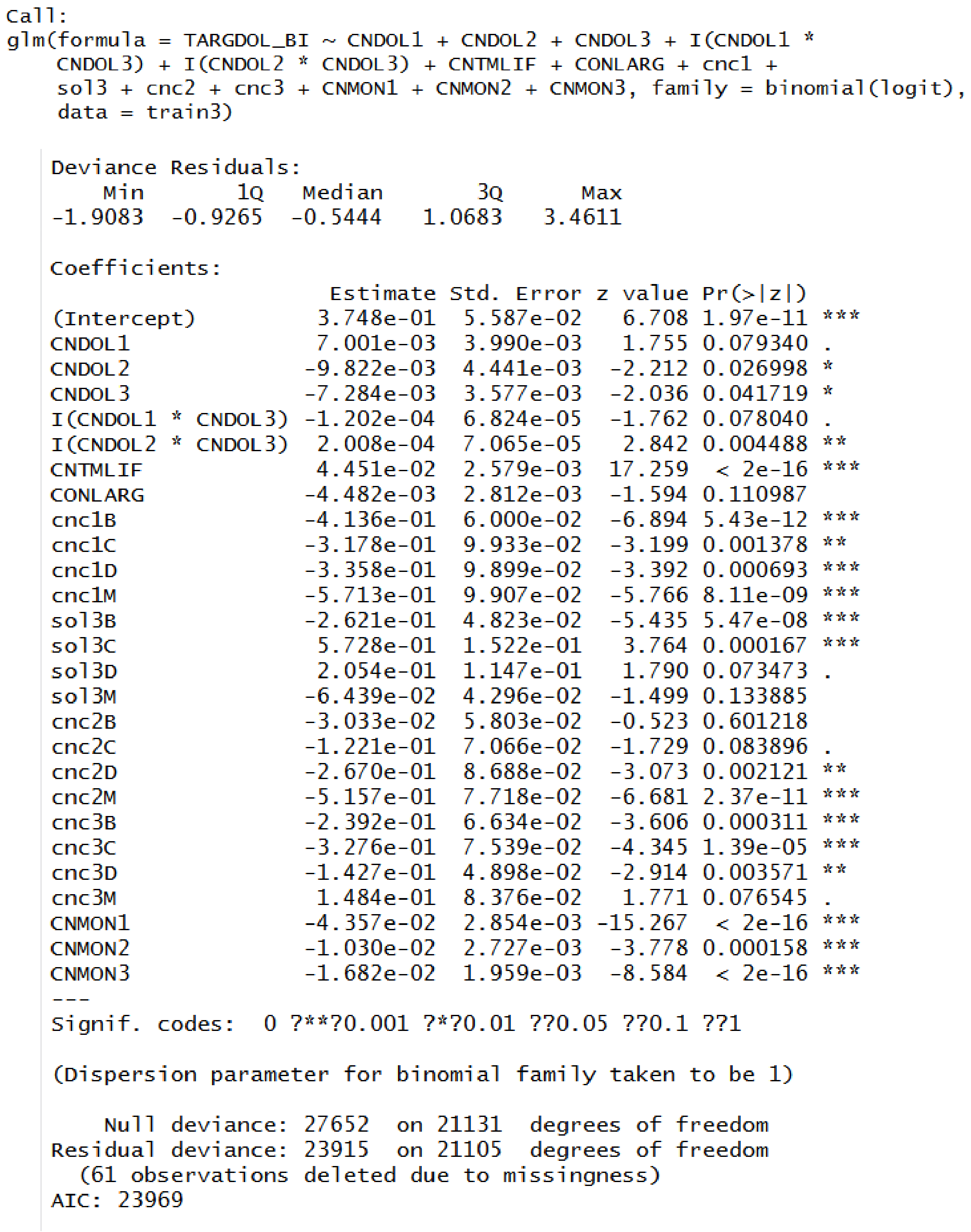
```
Call:
glm(formula = TARGDOL_BI ~ CNDOL1 + CNMON1 + SEX + cnc1 + sol1,
    family = binomial, data = train1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0524  -0.8071  -0.6104  -0.3643   3.1182

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.776471   0.069524 -11.168  < 2e-16 ***
CNDOL1       -0.024118   0.002445  -9.864  < 2e-16 ***
CNMON1       -0.033851   0.002461 -13.756  < 2e-16 ***
SEXB          0.255354   0.047133   5.418 6.04e-08 ***
SEXM          0.062048   0.027745   2.236   0.0253 *
SEXU         -0.044778   0.040588  -1.103   0.2699
cnc1B        -0.872578   1.062795  -0.821   0.4116
cnc1C       -10.100836  97.528681  -0.104   0.9175
cnc1D        -0.315705   0.065980  -4.785 1.71e-06 ***
cnc1M        -0.733982   0.618341  -1.187   0.2352
sol1B        -9.655354 196.970537  -0.049   0.9609
sol1D         0.583605   0.033210  17.573  < 2e-16 ***
sol1M        -0.196182   0.750441  -0.261   0.7938
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 42452  on 39654  degrees of freedom
Residual deviance: 40588  on 39642  degrees of freedom
AIC: 40614

Number of Fisher Scoring iterations: 10
```

Figure 4: Logistic Regression Model 1



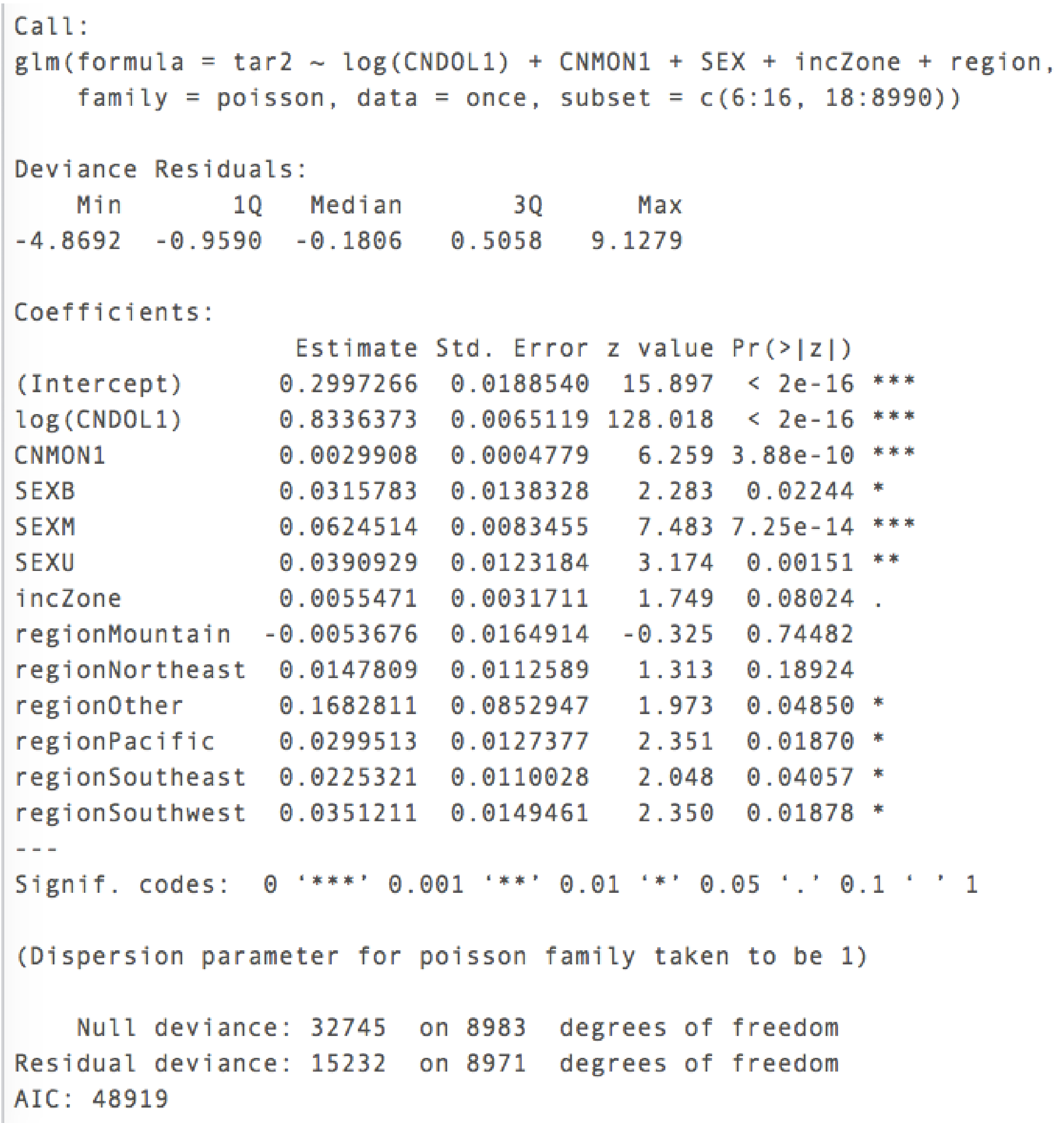
```
Call:
glm(formula = TARGDOL_BI ~ CNDOL1 + I(CNDOL1 * CNDOL2) + CONLARG +
    CONTRFST + SEX + CNMON1 + sol2, family = binomial(logit),
    data = train2)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.383  -0.790  -0.659   1.174   3.138

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.6876806  0.0973506  -7.064 1.62e-12 ***
CNDOL1             -0.0316937  0.0150614  -2.104  0.03535 *
I(CNDOL1 * CNDOL2)  0.0003813  0.0001979   1.927  0.05398 .
CONLARG             0.0283062  0.0180126   1.571  0.11607
CONTRFST           -0.0371589  0.0139387  -2.666  0.00768 **
SEXB                0.4449753  0.1230773   3.615  0.00030 ***
SEXM                0.1124360  0.0739158   1.521  0.12823
SEXU                0.0412383  0.1024195   0.403  0.68721
CNMON1             -0.0410157  0.0042218  -9.715  < 2e-16 ***
sol2B               0.1736363  0.0933589   1.860  0.06290 .
sol2C               0.0420723  0.1723947   0.244  0.80719
sol2D               0.8174852  0.0871810   9.377  < 2e-16 ***
sol2M               1.2612260  0.6459790   1.952  0.05089 .
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6209.6  on 5253  degrees of freedom
Residual deviance: 5742.8  on 5241  degrees of freedom
  (32 observations deleted due to missingness)
AIC: 5768.8

Number of Fisher Scoring iterations: 5
```

Figure 5: Logistic Regression Model 2



```
Call:
glm(formula = TARGDOL_BI ~ CNDOL1 + CNDOL2 + CNDOL3 + I(CNDOL1 *
    CNDOL3) + I(CNDOL2 * CNDOL3) + CNTMLIF + CONLARG + cncl +
    sol3 + cnc2 + cnc3 + CNMON1 + CNMON2 + CNMON3, family = binomial(logit),
    data = train3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9083  -0.9265  -0.5444   1.0683   3.4611

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.748e-01  5.587e-02   6.708 1.97e-11 ***
CNDOL1              7.001e-03  3.990e-03   1.755 0.079340 .
CNDOL2             -9.822e-03  4.441e-03  -2.212 0.026998 *
CNDOL3             -7.284e-03  3.577e-03  -2.036 0.041719 *
I(CNDOL1 * CNDOL3) -1.202e-04  6.824e-05  -1.762 0.078040 .
I(CNDOL2 * CNDOL3)  2.008e-04  7.065e-05   2.842 0.004488 **
CNTMLIF             4.451e-02  2.579e-03  17.259  < 2e-16 ***
CONLARG            -4.482e-03  2.812e-03  -1.594 0.110987
cnc1B              -4.136e-01  6.000e-02  -6.894 5.43e-12 ***
cnc1C              -3.178e-01  9.933e-02  -3.199 0.001378 **
cnc1D              -3.358e-01  9.899e-02  -3.392 0.000693 ***
cnc1M              -5.713e-01  9.907e-02  -5.766 8.11e-09 ***
sol3B              -2.621e-01  4.823e-02  -5.435 5.47e-08 ***
sol3C               5.728e-01  1.522e-01   3.764 0.000167 ***
sol3D               2.054e-01  1.147e-01   1.790 0.073473 .
sol3M              -6.439e-02  4.296e-02  -1.499 0.133885
cnc2B              -3.033e-02  5.803e-02  -0.523 0.601218
cnc2C              -1.221e-01  7.066e-02  -1.729 0.083896 .
cnc2D              -2.670e-01  8.688e-02  -3.073 0.002121 **
cnc2M              -5.157e-01  7.718e-02  -6.681 2.37e-11 ***
cnc3B              -2.392e-01  6.634e-02  -3.606 0.000311 ***
cnc3C              -3.276e-01  7.539e-02  -4.345 1.39e-05 ***
cnc3D              -1.427e-01  4.898e-02  -2.914 0.003571 **
cnc3M               1.484e-01  8.376e-02   1.771 0.076545 .
CNMON1             -4.357e-02  2.854e-03 -15.267  < 2e-16 ***
CNMON2             -1.030e-02  2.727e-03  -3.778 0.000158 ***
CNMON3             -1.682e-02  1.959e-03  -8.584  < 2e-16 ***
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27652  on 21131  degrees of freedom
Residual deviance: 23915  on 21105  degrees of freedom
  (61 observations deleted due to missingness)
AIC: 23969
```

Figure 6: Logistic Regression Model 3



```
Call:
glm(formula = tar2 ~ log(CNDOL1) + CNMON1 + SEX + incZone + region, 
    family = poisson, data = once, subset = c(6:16, 18:8990))

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-4.8692  -0.9590  -0.1806   0.5058   9.1279  

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)      0.2997266  0.0188540  15.897  < 2e-16 ***
log(CNDOL1)      0.8336373  0.0065119 128.018  < 2e-16 ***
CNMON1           0.0029908  0.0004779   6.259 3.88e-10 ***
SEXB             0.0315783  0.0138328   2.283  0.02244 *  
SEXM             0.0624514  0.0083455   7.483 7.25e-14 ***
SEXU             0.0390929  0.0123184   3.174  0.00151 ** 
incZone          0.0055471  0.0031711   1.749  0.08024 .  
regionMountain  -0.0053676  0.0164914  -0.325  0.74482    
regionNortheast  0.0147809  0.0112589   1.313  0.18924    
regionOther      0.1682811  0.0852947   1.973  0.04850 *  
regionPacific    0.0299513  0.0127377   2.351  0.01870 *  
regionSoutheast  0.0225321  0.0110028   2.048  0.04057 *  
regionSouthwest  0.0351211  0.0149461   2.350  0.01878 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 32745  on 8983  degrees of freedom
Residual deviance: 15232  on 8971  degrees of freedom
AIC: 48919
```

Figure 7(a): Multiple Regression Model 1.3

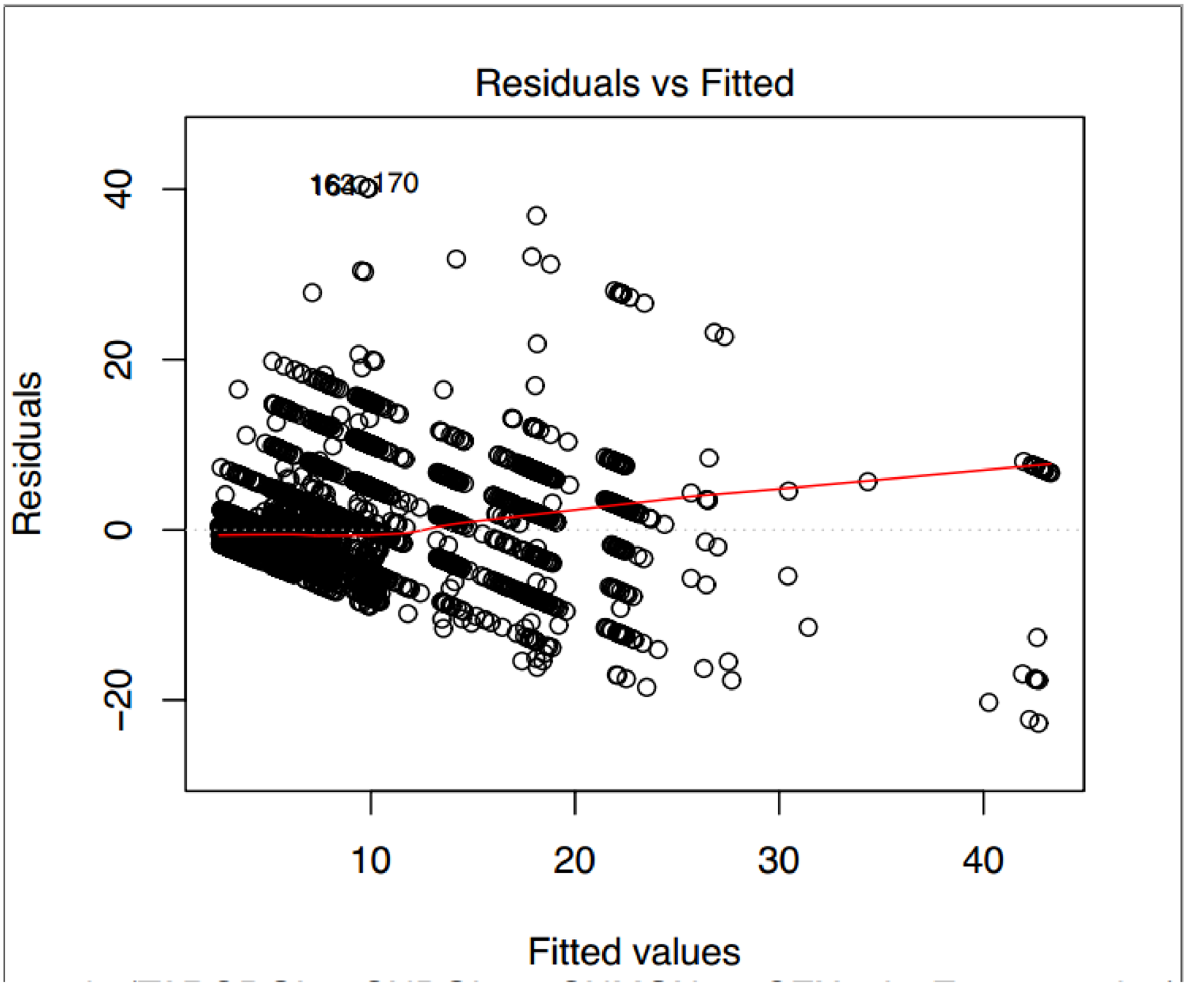
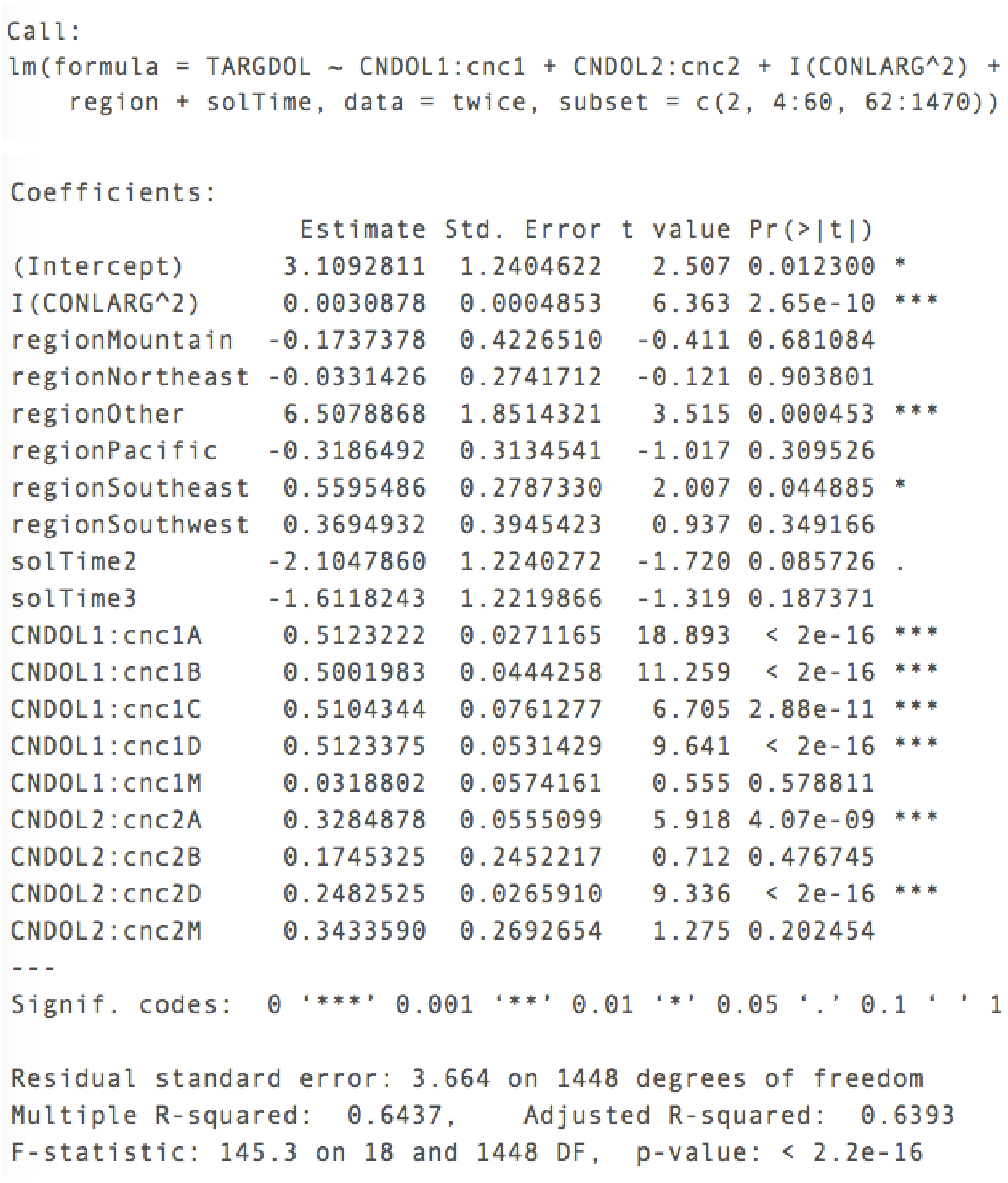


Figure 7(b): Residual Plot Multiple Regression Model 1.3



```
Call:
lm(formula = TARGDOL ~ CNDOL1:cnc1 + CNDOL2:cnc2 + I(CONLARG^2) +
    region + solTime, data = twice, subset = c(2, 4:60, 62:1470))

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      3.1092811  1.2404622   2.507 0.012300 *  
I(CONLARG^2)     0.0030878  0.0004853   6.363 2.65e-10 ***
regionMountain  -0.1737378  0.4226510  -0.411 0.681084    
regionNortheast -0.0331426  0.2741712  -0.121 0.903801    
regionOther      6.5078868  1.8514321   3.515 0.000453 ***
regionPacific   -0.3186492  0.3134541  -1.017 0.309526    
regionSoutheast  0.5595486  0.2787330   2.007 0.044885 *  
regionSouthwest  0.3694932  0.3945423   0.937 0.349166    
solTime2        -2.1047860  1.2240272  -1.720 0.085726 .  
solTime3        -1.6118243  1.2219866  -1.319 0.187371    
CNDOL1:cnc1A     0.5123222  0.0271165  18.893  < 2e-16 ***
CNDOL1:cnc1B     0.5001983  0.0444258  11.259  < 2e-16 ***
CNDOL1:cnc1C     0.5104344  0.0761277   6.705 2.88e-11 ***
CNDOL1:cnc1D     0.5123375  0.0531429   9.641  < 2e-16 ***
CNDOL1:cnc1M     0.0318802  0.0574161   0.555 0.578811    
CNDOL2:cnc2A     0.3284878  0.0555099   5.918 4.07e-09 ***
CNDOL2:cnc2B     0.1745325  0.2452217   0.712 0.476745    
CNDOL2:cnc2D     0.2482525  0.0265910   9.336  < 2e-16 ***
CNDOL2:cnc2M     0.3433590  0.2692654   1.275 0.202454    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.664 on 1448 degrees of freedom
Multiple R-squared:  0.6437,    Adjusted R-squared:  0.6393 
F-statistic: 145.3 on 18 and 1448 DF,  p-value: < 2.2e-16
```

Figure 8(a): Multiple Regression Model 2.1

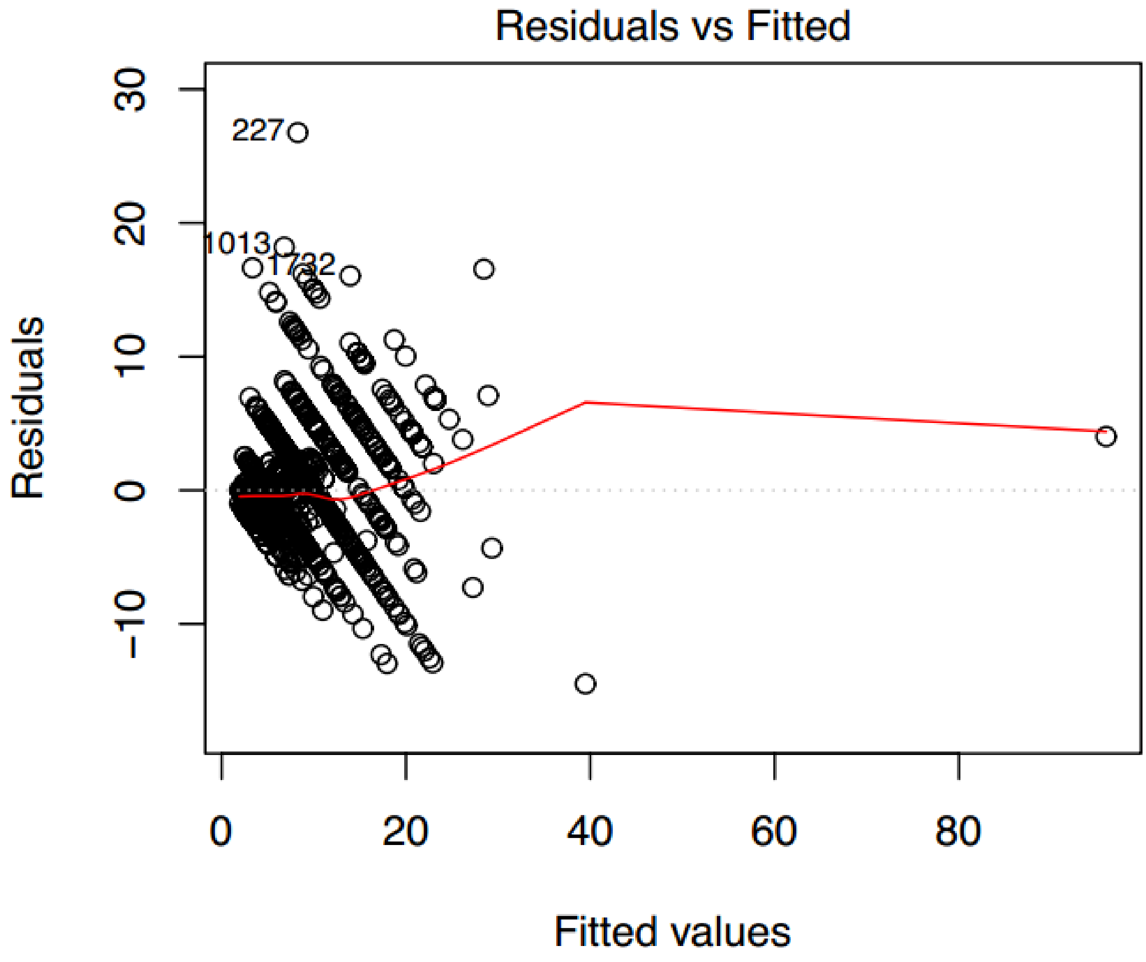
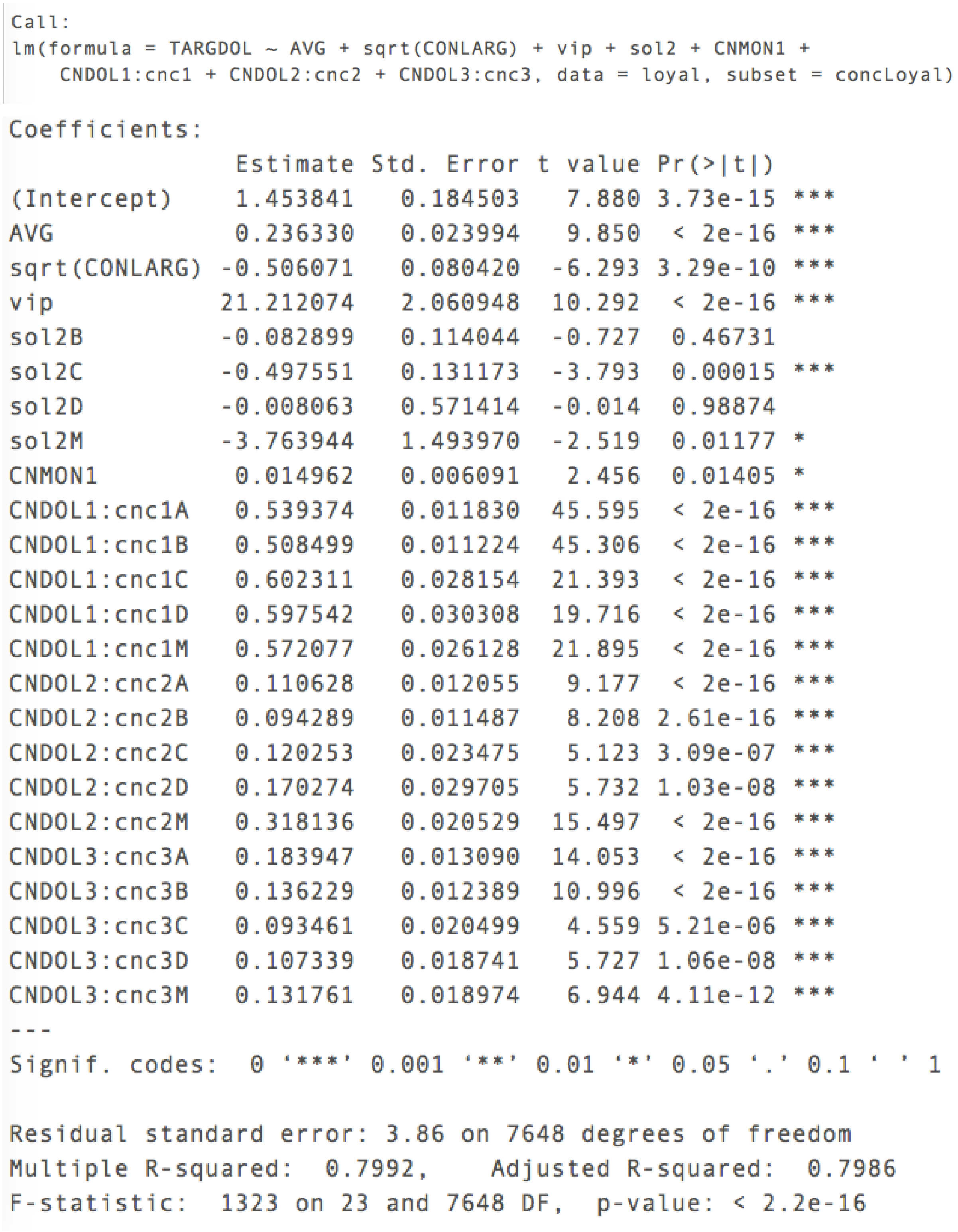


Figure 8(b): Multiple Regression Model 2.1



```
Call:
lm(formula = TARGDOL ~ AVG + sqrt(CONLARG) + vip + sol2 + CNMON1 +
    CNDOL1:cnc1 + CNDOL2:cnc2 + CNDOL3:cnc3, data = loyal, subset = concLoyal)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.453841   0.184503   7.880 3.73e-15 ***
AVG            0.236330   0.023994   9.850  < 2e-16 ***
sqrt(CONLARG) -0.506071   0.080420  -6.293 3.29e-10 ***
vip           21.212074   2.060948  10.292  < 2e-16 ***
sol2B         -0.082899   0.114044  -0.727  0.46731
sol2C         -0.497551   0.131173  -3.793  0.00015 ***
sol2D         -0.008063   0.571414  -0.014  0.98874
sol2M         -3.763944   1.493970  -2.519  0.01177 *
CNMON1         0.014962   0.006091   2.456  0.01405 *
CNDOL1:cnc1A   0.539374   0.011830  45.595  < 2e-16 ***
CNDOL1:cnc1B   0.508499   0.011224  45.306  < 2e-16 ***
CNDOL1:cnc1C   0.602311   0.028154  21.393  < 2e-16 ***
CNDOL1:cnc1D   0.597542   0.030308  19.716  < 2e-16 ***
CNDOL1:cnc1M   0.572077   0.026128  21.895  < 2e-16 ***
CNDOL2:cnc2A   0.110628   0.012055   9.177  < 2e-16 ***
CNDOL2:cnc2B   0.094289   0.011487   8.208 2.61e-16 ***
CNDOL2:cnc2C   0.120253   0.023475   5.123 3.09e-07 ***
CNDOL2:cnc2D   0.170274   0.029705   5.732 1.03e-08 ***
CNDOL2:cnc2M   0.318136   0.020529  15.497  < 2e-16 ***
CNDOL3:cnc3A   0.183947   0.013090  14.053  < 2e-16 ***
CNDOL3:cnc3B   0.136229   0.012389  10.996  < 2e-16 ***
CNDOL3:cnc3C   0.093461   0.020499   4.559 5.21e-06 ***
CNDOL3:cnc3D   0.107339   0.018741   5.727 1.06e-08 ***
CNDOL3:cnc3M   0.131761   0.018974   6.944 4.11e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.86 on 7648 degrees of freedom
Multiple R-squared:  0.7992,	Adjusted R-squared:  0.7986
F-statistic:  1323 on 23 and 7648 DF,  p-value: < 2.2e-16
```

Figure 9(a): Multiple Regression Model 3.1

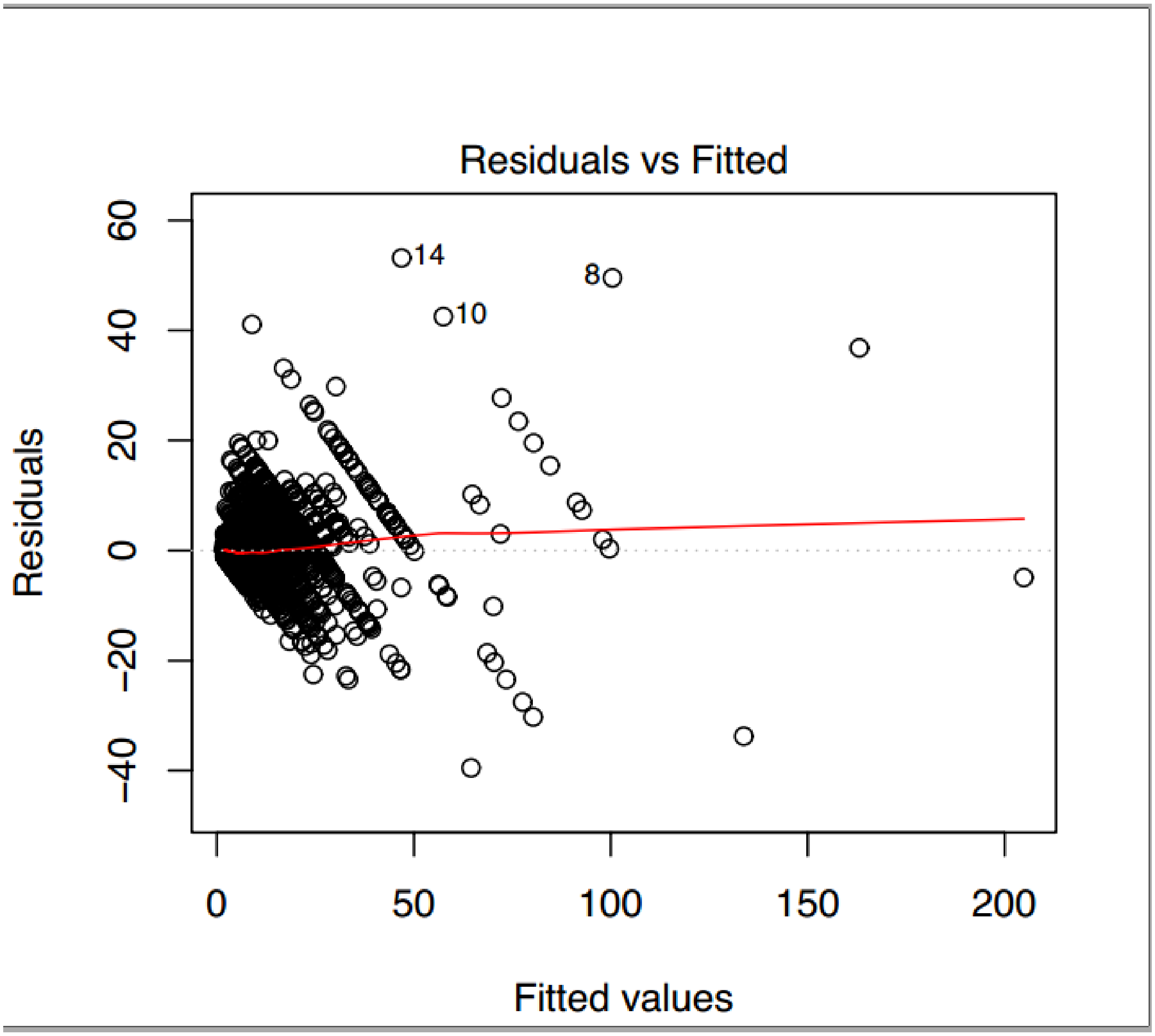


Figure 9(b): Multiple Regression Model 3.1

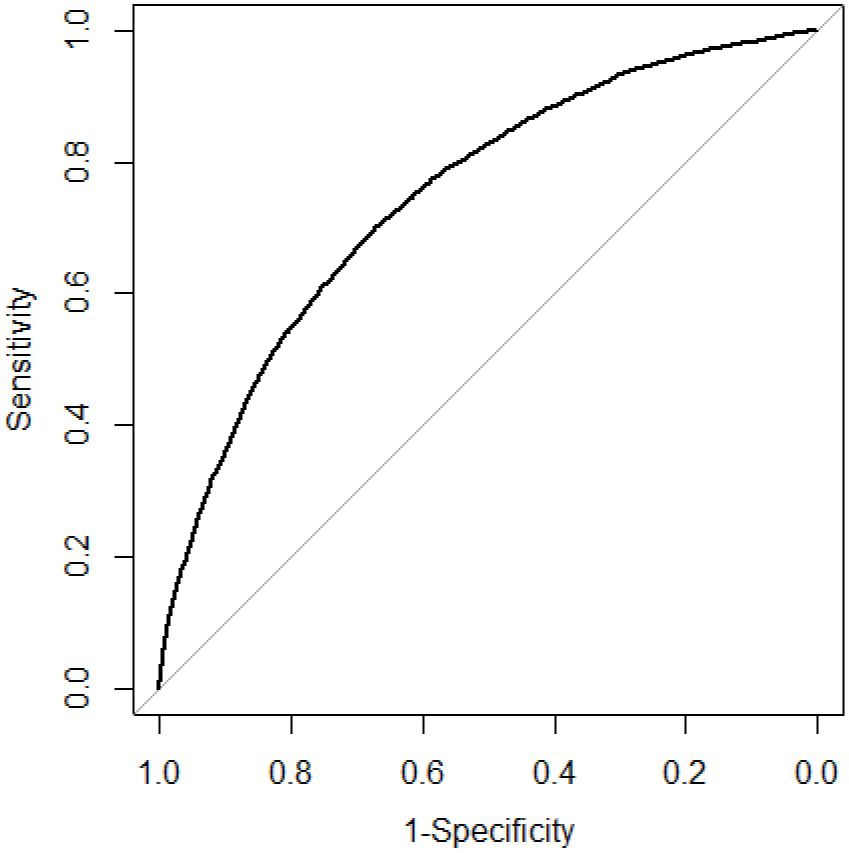


Figure 10: ROC for Training Data Set 3