edX (https://www.edx.org)

MITx: 15.071x The Analytics Edge

zhushun0008 (/dashboard) ▼

Courseware (/courses/MITx/15.071x/1T2014/courseware)   Course Info (/courses/MITx/15.071x/1T2014/info)

Discussion (/courses/MITx/15.071x/1T2014/discussion/forum)   Progress (/courses/MITx/15.071x/1T2014/progress)

Syllabus (/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/)

Schedule (/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/)

Help

UNDERSTANDING WHY PEOPLE VOTE

In August 2006 three researchers (Alan Gerber and Donald Green of Yale University, and Christopher Larimer of the University of Northern Iowa) carried out a large scale field experiment in Michigan, USA to test the hypothesis that one of the reasons people vote is social, or extrinsic, pressure. To quote the first paragraph of their 2008 research paper (http://www.apsanet.org/imgtest/apsrfeb08gerberetal.pdf):

> Among the most striking features of a democratic political system is the participation of millions of voters in elections. Why do large numbers of people vote, despite the fact that ... "the casting of a single vote is of no significance where there is a multitude of electors"? One hypothesis is adherence to social norms. Voting is widely regarded as a citizen duty, and citizens worry that others will think less of them if they fail to participate in elections. Voters' sense of civic duty has long been a leading explanation of vote turnout...

In this homework problem we will use both logistic regression and classification trees to analyze the data they collected.

THE DATA

The researchers grouped about 344,000 voters into different groups randomly - about 191,000 voters were a "control" group, and the rest were categorized into one of four "treatment" groups. These five groups correspond to five binary variables in the dataset.

1. "Civic Duty" (variable **civicduty**) group members were sent a letter that simply said "DO YOUR CIVIC DUTY - VOTE!"

2. "Hawthorne Effect" (variable **hawthorne**) group members were sent a letter that had the "Civic Duty" message plus the additional message "YOU ARE BEING STUDIED" and they were informed that their voting behavior would be examined by means of public records.

3. "Self" (variable **self**) group members received the "Civic Duty" message as well as the recent voting record of everyone in that household and a message stating that another message would be sent after the election with updated records.

4. "Neighbors" (variable **neighbors**) group members were given the same message as that for the "Self" group, except the message not only had the household voting records but also that of neighbors - maximizing social pressure.

5. "Control" (variable **control**) group members were not sent anything, and represented the typical voting situation.

Additional variables include **sex** (0 for male, 1 for female), **yob** (year of birth), and the dependent variable **voting** (1 if they voted, 0 otherwise).

PROBLEM 1.1 - EXPLORATION AND LOGISTIC REGRESSION (1 point possible)

We will first get familiar with the data. Load the CSV file gerber.csv (/c4x/MITx/15.071x/asset/gerber.csv) into R. What proportion of people in this dataset voted in this election?

PROBLEM 1.2 - EXPLORATION AND LOGISTIC REGRESSION (1 point possible)

Which of the four "treatment groups" had the largest fraction of voters?

- ○ Civic Duty
- ○ Hawthorne Effect
- ○ Self
- ○ Neighbors

PROBLEM 1.3 - EXPLORATION AND LOGISTIC REGRESSION (1 point possible)

Build a **logistic regression** model for *voting* using the four treatment group variables as the independent variables (civicduty, hawthorne, self, and neighbors). Use all the data to build the model (DO NOT split the data into a training set and testing set). Which of the following coefficients are significant in the logistic regression model?

- ☐ Civic Duty
- ☐ Hawthorne Effect
- ☐ Self
- ☐ Neighbors

PROBLEM 1.4 - EXPLORATION AND LOGISTIC REGRESSION (1 point possible)

Using a threshold of **0.3**, what is the accuracy of the logistic regression model? (When making predictions, you don't need to use the newdata argument since we didn't split our data.)

PROBLEM 1.5 - EXPLORATION AND LOGISTIC REGRESSION (1 point possible)

Using a threshold of **0.5**, what is the accuracy of the logistic regression model?

PROBLEM 1.6 - EXPLORATION AND LOGISTIC REGRESSION (1 point possible)

Compare your previous two answers to the percentage of people who did not vote (the baseline accuracy) and compute the AUC of the model. What is happening here?

○ Even though all of the variables are significant, this is a weak predictive model.

○ The model's accuracy doesn't improve over the baseline, but the AUC is high, so this is a strong predictive model.

Show Answer   *You have used 0 of 1 submissions*

PROBLEM 2.1 - TREES (1 point possible)

We will now try out trees. Build a CART tree for *voting* using all data and the same four treatment variables we used before. Don't set the option *method="class"* - we are actually going to create a regression tree here. We are interested in building a tree to explore the fraction of people who vote, or the probability of voting. We'd like CART to split our groups if they have different probabilities of voting. If we used method='class', CART would only split if one of the groups had a probability of voting above 50% and the other had a probability of voting less than 50% (since the predicted outcomes would be different). However, with regression trees, CART will split even if both groups have probability less than 50%.

Leave all the parameters at their default values. You can use the following command in R to build the tree:

```
CARTmodel = rpart(voting ~ civicduty + hawthorne + self + neighbors, data=gerber)
```

Plot the tree. What happens, and if relevant, why?

○ Only the "Neighbors" variable is used in the tree - it is the only one with a big enough effect.

○ All variables are used - they all make a difference.

○ No variables are used (the tree is only a root node) - none of the variables make a big enough effect to be split on.

Show Answer   *You have used 0 of 1 submissions*

PROBLEM 2.2 - TREES (1 point possible)

Now build the tree using the command:

```
CARTmodel2 = rpart(voting ~ civicduty + hawthorne + self + neighbors, data=gerber, cp=0.0)
```

to force the complete tree to be built. Then plot the tree. What do you observe about the order of the splits?

○ Civic duty is the first split, neighbor is the last.

○ Neighbor is the first split, civic duty is the last.

Show Answer   *You have used 0 of 1 submissions*

PROBLEM 2.3 - TREES (1 point possible)

Using only the CART tree plot, determine what fraction (a number between 0 and 1) of "Civic Duty" people voted:

PROBLEM 2.4 - TREES (2 points possible)

Make a new tree that includes the "sex" variable, again with cp = 0.0. Notice that sex appears as a split that is of secondary importance to the treatment group.

In the control group, which gender is more likely to vote?

- ○ Men (0)
- ○ Women (1)

In the "Civic Duty" group, which gender is more likely to vote?

- ○ Men (0)
- ○ Women (1)

PROBLEM 3.1 - INTERACTION TERMS (1 point possible)

We know trees can handle "nonlinear" relationships, e.g. "in the 'Civic Duty' group **and** female", but as we will see in the next few questions, it is possible to do the same for logistic regression. First, let's explore what trees can tell us some more.

Let's just focus on the "Control" treatment group. Create a regression tree using **just the "control" variable**, then create another tree with the "control" and "sex" variables, both with cp=0.0.

In the "control" only tree, what is the absolute value of the difference in the predicted probability of voting between being in the control group versus being in a different group? You can use the absolute value function to get answer, i.e. abs(Control Prediction - Non-Control Prediction). Add the argument "digits = 6" to the prp command to get a more accurate estimate.

PROBLEM 3.2 - INTERACTION TERMS (1 point possible)

Now, using the second tree (with control and sex), determine who is affected **more** by NOT being in the control group (being in any of the four treatment groups):

- ○ Men, by a margin of more than 0.001
- ○ Women, by a margin of more than 0.001
- ○ They are affected about the same (change in probability within 0.001 of each other).

PROBLEM 3.3 - INTERACTION TERMS (1 point possible)

Going back to logistic regression now, create a model using "sex" and "control". Interpret the coefficient for "sex":

- ○ Coefficient is negative, reflecting that women are less likely to vote
- ○ Coefficient is negative, reflecting that women are more likely to vote
- ○ Coefficient is positive, reflecting that women are less likely to vote
- ○ Coefficient is positive, reflecting that women are more likely to vote

---

PROBLEM 3.4 - INTERACTION TERMS (1 point possible)

The regression tree calculated the percentage voting exactly for every one of the four possibilities (Man, Not Control), (Man, Control), (Woman, Not Control), (Woman, Control). Logistic regression has attempted to do the same, although it wasn't able to do as well because it can't consider exactly the joint possibility of being a women and in the control group.

We can quantify this precisely. Create the following dataframe (this contains all of the possible values of sex and control), and evaluate your logistic regression using the predict function (where "LogModelSex" is the name of your logistic regression model that uses both control and sex):

```
Possibilities = data.frame(sex=c(0,0,1,1),control=c(0,1,0,1))
predict(LogModelSex, newdata=Possibilities, type="response")
```

The four values in the results correspond to the four possibilities in the order they are stated above ((Man, Not Control), (Man, Control), (Woman, Not Control), (Woman, Control)). What is the **absolute difference** between the tree and the logistic regression for the **(Woman, Control)** case? Give an answer with five numbers after the decimal point.

PROBLEM 3.5 - INTERACTION TERMS (1 point possible)

So the difference is not too big for this dataset, but it is there. We're going to add a new term to our logistic regression now, that is the **combination of the "sex" and "control" variables** - so if this new variable is 1, that means the person is a woman AND in the control group. We can do that with the following command:

```
LogModel2 = glm(voting ~ sex + control + sex:control, data=gerber, family="binomial")
```

How do you interpret the coefficient for the new variable **in isolation**? That is, how does it relate to the dependent variable?

- ○ If a person is a woman or in the control group, the chance that she voted goes up.
- ○ If a person is a woman and in the control group, the chance that she voted goes up.
- ○ If a person is a woman or in the control group, the chance that she voted goes down.
- ○ If a person is a woman and in the control group, the chance that she voted goes down.

PROBLEM 3.6 - INTERACTION TERMS (1 point possible)

Run the same code as before to calculate the average for each group:

predict(LogModel2, newdata=Possibilities, type="response")

Now what is the difference between the logistic regression model and the CART model for the (Woman, Control) case? Again, give your answer with five numbers after the decimal point.

Show Answer   *You have used 0 of 3 submissions*

PROBLEM 3.7 - INTERACTION TERMS (1 point possible)

This example has shown that trees can capture nonlinear relationships that logistic regression can not, but that we can get around this sometimes by using variables that are the combination of two variables. Should we always include all possible interaction terms of the independent variables when building a logistic regression model?

○ Yes
○ No

Show Answer   *You have used 0 of 1 submissions*

Please remember not to ask for or post complete answers to homework questions in this discussion forum.

Show Discussion   ✎ **New Post**