Problem

Wayne Kenney

2023-06-22

Explanation of Problem

The situation is not an easy one to sort out.

The data comes from a large database of 226,838,098 million weibos (Chinese tweets). Each weibo has been labeled as being censored (1) or not censored (0). For each weibo, particulars of the language used has been computed into 99 qualities. These qualities are things like posemo and negemo, indicating the number of positive and negative emotions. All of these 99 qualities are integer values from 0 to the max amount found in the data, typically up to 30 or as high as 240. In addition, there is tokencount, the number of Chinese tokens found in the tweet, and totallen, the total number of characters in the tweet. The dataset then has 226 million rows and 100 columns, 99 qualities and one target column for censored or not censored. There is a high level of multicollinearity between many of these predictors.

The overall number of censored tweets is 86,083, and the overall number of not censored is 226,752,015, for an overall probability of about 0.00322.

Now I have three hypotheses as the reason for censorship, and have picked out three sets of qualities that would support these hypothesis or refute them. I would like to fit a model to each of these hypothesis, look at significance, and compare all these models to each other to determine which is most likely.

In addition, I would like to do a broad search of all the 99 qualities in order to find a set that may explain the censorship column better than the sets for the three hypotheses.

Preprocessing the Data

Because of rarity of censored in relation to not censored, and the high number of rows, I decided that I did not want to solve this problem with sampling, and fitting a logistic regression on the sampled data. I wanted to use all 86k censored tweets to use all the available information. For the fact that all the predictors are integers and not continuous, it became possible to preprocess the data so that it may take up less space.

All the three hypotheses together formed a set of 7 qualities. I aggregated on all the unique integer combinations of the 7 predictors and summed up the censored and not censored for those combinations. This derivative dataset has 8,419,694 rows and 9 columns, 7 predictors and 2 target columns (# of censored, # of not censored). head(df) yields the following:

percept	cogproc	negemo	ppron	posemo	tokencount	relativ	censored	not_censored
0	0	0	0	0	1	0	3543	8 752 360
0	0	0	0	0	1	1	68	186 799
0	0	0	0	0	2	0	15700	53 699 600
0	0	0	0	0	2	1	197	458 375
0	0	0	0	0	2	2	21	29 083

This allows us to have a manageable dataset of only 8 million rows, with no loss of information. Further, logistic regression on the original dataset with 0 and 1's will yield the same coefficients and standard errors

as a binomial regression on the aggregated dataset. This is simple to check with a toy dateset.

The problem

The main problem is that of model adequacy checking. The residuals vs fitted for these models look off. They exhibit heteroskedasticity.

For example a binomial model censored,not_censored ~ posemo*negemo + tokencount. (This model was chosen to see the effect of postive and negative tokens, controlling for complexity of tweet measured by tokencount).

The R code that generated these residuals is simply

```
form <- formula(cbind(censored,not_censored) ~ posemo*negemo + tokencount)
m <- glm(form, df, family=binomial())

fit = fitted(m)
res = residuals(m)
```

1. Residual vs Fitted
2. Residual vs tokencount
3. Residual vs posemo
4. Residual vs negemo

While this is only one model, the other hypotheses have similar patterns. How should I proceed?