Generalized Linear Models

Wen-Han Hu (whu24)

Problem 1 (25 points: 5 points each question): Building and analyzing the logistic regression model

For the problem below, build the logistic regression model (fit.all) using all the predictors and answer the following questions by including the corresponding R code and showing all the required mathematical derivations used to answer these questions:

1. Let X_h be the predictor with the highest estimate (in terms of its absolute value) for its regression coefficient. Build a single predictor logistic regression model (*fit.single*) using X_h as the predictor. Write the equations relating the dependent variable (Response) to the explanatory variable in terms of:
   $X_h = CategoryEverythingElse$
   a. Probabilities: $Prob(Y = Yes \mid X_h = x)$
   $$Prob(Y = Yes \mid X_h = x) = \frac{1}{1 + e^{-(0.1246 - 2.3219 \times X_h)}}$$
   b. Odds: $Prob(Y = Yes)$
   $$\frac{Prob(Y = Yes)}{1 - Prob(Y = Yes)} = e^{(0.1246 - 2.3219 \times X_h)}$$
   c. Logita
   $$logit = log(odds) = log(\frac{Prob(Y = Yes)}{1 - Prob(Y = Yes)}) = (0.1246 - 2.3219 \times X_h)$$

2. Write the estimated equation for the *fit.all* model in all three formats (if the number of predictors is more than four, then include only those four predictors whose absolute value estimates are the highest):
   $$X_1 = CategoryEverythingElse, X_2 = CategoryBusiness/Industrial,$$
   $$X_3 = CategoryElectronics, X_4 = currencyGBP$$
   a. The logit as a function of the predictors.
   $$logit = (-1.21 - 1.58 \times X_1 + 1.29 \times X_2 + 1.27 \times X_3 + 1.16 \times X_4)$$
   b. The odds as a function of the predictors.
   $$odds = e^{(-1.21 - 1.58 \times X_1 + 1.29 \times X_2 + 1.27 \times X_3 + 1.16 \times X_4)}$$
   c. The probability as a function of the predictors
   $$p = \frac{1}{1 - e^{-(-1.21 - 1.58 \times X_1 + 1.29 \times X_2 + 1.27 \times X_3 + 1.16 \times X_4)}}$$

3. Let X_h be the predictor with the highest estimate (in terms of its absolute value) for its regression coefficient in the *fit.all*. Compute the odds ratio that estimated a single unit increase in X_h, holding the other predictors constant. For example, if $X_h = 1\ then$:
   $$\frac{odds(X_1 + 1, X_2, \dots, X_q)}{odds(X_1, X_2, \dots, X_q)}$$
   Provide the interpretation for this regression coefficient. If it were a linear regression model, how would the interpretation change for a single unit increase in X_h.
   $assume\ c = coefficient\ of\ X_h$
   $$\frac{odds(X_1 + 1, X_2, \dots, X_q)}{odds(X_1, X_2, \dots, X_q)} = e^c = e^{-1.58} \approx 0.206$$

Thus, odds ratio increases 0.206 when a single unit increases in X_h.

If it were a linear regression model, the increase of single unit will depend on its coefficient and directly reflect to Y. So, Y will increase -1.58.

4. Build a reduced logistic regression model (*fit.reduced*) using only the predictors that are statistically significant. Assess if the reduced model is equivalent to the full model. Justify your answer.

The following picture is the screenshot of *fit.all:*

```
Call:
glm(formula = `Competitive?` ~ ., family = binomial(link = "logit"),
    data = train_dummy)

Deviance Residuals:
   Min       1Q  Median      3Q     Max
-4.564  -0.868   0.000   0.833   2.197

Coefficients: (4 not defined because of singularities)
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -1.21e+00   3.26e-01   -3.71  0.00020 ***
`CategoryAntique/Art/Craft`   5.63e-03   3.33e-01    0.02  0.98651
`CategoryBusiness/Industrial` 1.29e+00   7.84e-01    1.65  0.09846 .
CategoryCollectibles          6.93e-01   2.90e-01    2.39  0.01696 *
CategoryElectronics           1.27e+00   6.25e-01    2.03  0.04237 *
CategoryEverythingElse       -1.58e+00   1.10e+00   -1.44  0.15017
`CategoryHealth/Beauty`      -1.05e+00   5.35e-01   -1.96  0.05054 .
`CategoryHome/Garden`         4.55e-01   3.51e-01    1.30  0.19490
CategoryJewelry              -4.54e-01   4.04e-01   -1.12  0.26154
CategoryPhotography           7.41e-01   1.35e+00    0.55  0.58334
`CategoryPottery/Glass`             NA         NA      NA       NA
currencyEUR                  -3.58e-01   2.54e-01   -1.41  0.15941
currencyGBP                   1.16e+00   5.80e-01    2.00  0.04542 *
currencyUS                          NA         NA      NA       NA
sellerRating                 -3.76e-05   1.51e-05   -2.49  0.01267 *
Duration10                   -8.38e-02   2.76e-01   -0.30  0.76138
Duration5                     2.51e-01   2.11e-01    1.19  0.23318
Duration7                           NA         NA      NA       NA
endDayMon                     8.57e-01   2.36e-01    3.64  0.00027 ***
endDaySun                     5.44e-01   1.96e-01    2.77  0.00562 **
endDayThu                    -1.84e-01   5.31e-01   -0.35  0.72946
endDayWed                           NA         NA      NA       NA
ClosePrice                    1.35e-01   1.34e-02   10.03  < 2e-16 ***
OpenPrice                    -1.47e-01   1.44e-02  -10.20  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Those variables that the p-value <0.05 are significant predictors.
Then, do the *fit.reduced* with significant predictors. And the screenshot of comparing *fit.reduced to fit.all* as following:

```
Analysis of Deviance Table

Model 1: `Competitive?` ~ CategoryCollectibles + CategoryElectronics +
    currencyGBP + sellerRating + endDayMon + endDaySun + ClosePrice +
    OpenPrice
Model 2: `Competitive?` ~ `CategoryAntique/Art/Craft` + `CategoryBusiness/Industrial` +
    CategoryCollectibles + CategoryElectronics + CategoryEverythingElse +
    `CategoryHealth/Beauty` + `CategoryHome/Garden` + CategoryJewelry +
    CategoryPhotography + `CategoryPottery/Glass` + currencyEUR +
    currencyGBP + currencyUS + sellerRating + Duration10 + Duration5 +
    Duration7 + endDayMon + endDaySun + endDayThu + endDayWed +
    ClosePrice + OpenPrice
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1174       1199
2      1163       1171 11     27.7    0.0036 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the P-value ≤ α(0.05), we conclude that there is a statistically significant association between the variables. Thus, the reduced model is equivalent to the full model.

5. Compute the dispersion of your model and run the dispersion diagnostic test. If the constructed model is overdispersed, then discuss the ways to deal with the issue.

$$\phi = \frac{Residual\ Deviance}{Residual\ df} >> 1, \phi(fit.reduced) = \frac{1199}{1174} = 1.02, \phi\ (fit.all) = \frac{1171}{1163} = 1.01$$

Thus, no overdispersion on data.