Generalized Linear Models

Name: Rachit Shah (rshah25)

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:**

   ANS:

   From the summary of fit.all, we can see that the predictor with highest estimate is $X_h = currency\_GBP$ with an estimate of 2.014

```
C:/Users/rachi/OneDrive/NCSU/sem 2/adbi/Projects/GLM Logistic Regression HW Resources-20190206/
(Intercept)                       5.724e-01  7.049e-01   0.812 0.416731
sellerRating                     -2.505e-05  1.604e-05  -1.561 0.118431
ClosePrice                        1.162e-01  1.167e-02   9.959  < 2e-16 ***
OpenPrice                        -1.394e-01  1.376e-02 -10.127  < 2e-16 ***
`Category_Health/Beauty`         -2.410e+00  8.108e-01  -2.973 0.002951 **
Category_Books                   -1.197e+00  7.815e-01  -1.531 0.125657
`Category_Home/Garden`           -9.348e-01  7.565e-01  -1.236 0.216615
Category_Collectibles            -9.264e-01  7.227e-01  -1.282 0.199924
`Category_Toys/Hobbies`          -8.255e-01  7.145e-01  -1.155 0.247897
`Category_Antique/Art/Craft`     -1.011e+00  7.279e-01  -1.389 0.164942
Category_Automotive              -1.479e+00  7.358e-01  -2.010 0.044473 *
`Category_Music/Movie/Game`      -8.925e-01  7.135e-01  -1.251 0.210992
Category_Electronics             -4.559e-01  9.159e-01  -0.498 0.618651
`Category_Coins/Stamps`          -1.964e+00  8.713e-01  -2.254 0.024224 *
Category_Jewelry                 -1.244e+00  7.701e-01  -1.615 0.106257
`Category_Business/Industrial`   -1.495e+00  1.056e+00  -1.416 0.156708
Category_Computer                -1.427e+00  9.681e-01  -1.474 0.140464
`Category_Clothing/Accessories`  -2.950e+00  8.417e-01  -3.505 0.000456 ***
Category_SportingGoods           -1.170e+00  7.894e-01  -1.482 0.138244
Category_EverythingElse          -1.187e+01  7.998e+01  -0.148 0.882035
Category_Photography              1.765e-01  1.414e+00   0.125 0.900646
currency_GBP                      2.014e+00  5.612e-01   3.589 0.000332 ***
Duration_5                        4.458e-01  2.208e-01   2.019 0.043487 *
Duration_10                      -1.295e-01  2.461e-01  -0.526 0.598909
Duration_3                        2.033e-02  2.775e-01   0.073 0.941590
Duration_1                       -1.330e+00  9.475e-01  -1.403 0.160489
endDay_Sun                        1.790e-01  2.209e-01   0.811 0.417627
endDay_Wed                       -3.827e-01  4.357e-01  -0.878 0.379784
endDay_Thu                       -1.103e+00  5.151e-01  -2.142 0.032194 *
endDay_Mon                        5.658e-01  2.207e-01   2.564 0.010340 *
endDay_Tue                        1.292e-01  2.963e-01   0.436 0.662788
```

   a. Probabilities:

$$Prob(Y = Yes \mid X_h = x) = \frac{1}{1 + e^{-(0.07930 + 0.69140 * currencyGBP)}}$$

   b. Odds:

$$Odds = \frac{p}{1-p} = \frac{\frac{1}{1+e^{-(0.07930+0.69140*currencyGBP)}}}{1-\frac{1}{1+e^{-(0.07930+0.69140*currencyGBP)}}} = e^{(0.07930+0.69140*currencyGBP)}$$

c. Logit

$$Logit = \log(odds) = \log\left(e^{(0.07930+0.69140*currencyGBP)}\right)$$
$$= 0.07930 + 0.69140 * currencyGBP$$

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):**

ANS:

The 4 predictors with the highest estimates are: currencyGBP, endDayMon, Duration5 and endDaySun.

a. The logit as a function of the predictors.

$$Logit = 0.5724 + 2.0141 * currencyGBP + 0.5658 * endDayMon + 0.4458 * Duration5 + 0.179 * endDaySun$$

b. The odds as a function of the predictors.

$$Odds = e^{logit}$$
$$= e^{0.5724+2.0141*currencyGBP+0.5658*endDayMon+0.4458*Duration5+\ .179*endDaySun}$$

c. The probability as a function of the predictors

$$Prob = \frac{odds}{1+odds} = \frac{e^{logit}}{1+e^{logit}} = \frac{1}{1+e^{-logi}}$$
$$= \frac{1}{1+e^{-(0.5724+2.0141*currencyGBP+0.5658*endDayMon+\ .4458*Duration5+\ .179*endDaySun)}}$$

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.

ANS:

Here, $X_h = currencyGBP$ and the rest of the predictors $X_2, X_3, \dots X_q$ are constant. Hence,

$$\frac{odds(X_1+1, X_2, \dots, X_q)}{odds(X_1, X_2, \dots, X_q)} = \frac{e^{Int+coef*(X_h+1)}}{e^{Int+coef*X_h}} = e^{Int-Int+coef*X_h-co\quad *X_h+coef} = e^{coef}$$

Since the estimate for currencyGBP is 2.014, $e^{coef} = e^{2.014} = 7.493$

This means that for a unit increase of currencyGBP, the response variable will change 7.493 times for logistic regression. For 10 times increase in currencyGBP will cause 7.493^10 increase in response variable.

However, for linear regression, the change would be proportional to 2.014 and not its exponential.

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.**
   ANS:
   The statistically significant predictors which we can ascertain from fit.all are: ClosePrice, OpenPrice, `Category_Health/Beauty`, Category_Automotive, `Category_Coins/Stamps`, `Category_Clothing/Accessories`, currency_GBP, Duration_5, endDay_Thu and endDay_Mon.
   After fitting this reduced model and performing chi-square anova test we can find whether they are equivalent or not. From the result of the test, the p-value is 0.3162 which states that the difference is not significant and hence they are equivalent. Hence, we should choose the simpler model.

```
Model 1: Competitive ~ ClosePrice + OpenPrice + `Category_Health/Beauty` +
    Category_Automotive + `Category_Coins/Stamps` + `Category_Clothing/Accessories` +
    currency_GBP + Duration_5 + endDay_Thu + endDay_Mon
Model 2: Competitive ~ sellerRating + ClosePrice + OpenPrice + `Category_Health/Beauty` +
    Category_Books + `Category_Home/Garden` + Category_Collectibles +
    `Category_Toys/Hobbies` + `Category_Antique/Art/Craft` +
    Category_Automotive + `Category_Music/Movie/Game` + Category_Electronics +
    `Category_Coins/Stamps` + Category_Jewelry + `Category_Business/Industrial` +
    Category_Computer + `Category_Clothing/Accessories` + Category_SportingGoods +
    Category_EverythingElse + Category_Photography + currency_GBP +
    Duration_5 + Duration_10 + Duration_3 + Duration_1 + endDay_Sun +
    endDay_Wed + endDay_Thu + endDay_Mon + endDay_Tue
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1172     1177.4
2      1152     1155.0 20   22.459   0.3162
>
```

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.**
   ANS:
   The dispersion of the model can be calculated by the formula:

$$Dispersion\ \emptyset = \frac{Residual\ Deviance}{Degrees\ of\ Freedom} = \frac{1155}{1152} = 1.00464 \approx 1$$

   Hence, the dispersion is not too great then 1. Also, the dispersion diagnostic test in qcc package returns a p-value of 1 signifying that the model is not overdispersed.
   If the test had resulted positive and there was overdispersion in our model, then we would have to refit our model with quasi-binomial distribution instead of binomial.

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1634.7  on 1182  degrees of freedom
Residual deviance: 1155.0  on 1152  degrees of freedom
AIC: 1217

Number of Fisher Scoring iterations: 10
```

```
Dispersion of model is 1.00464075767012
> sample_size=rep(100, length(train_df$Competitive))
> qcc.overdispersion.test(train_df$Competitive, size=sample_size, type="binomial")

Overdispersion test Obs.Var/Theor.Var Statistic p-value
      binomial data         0.4695094  554.9601       1
> |
```

Competitive Auctions on eBay.com. The file eBayAuctions.xls contains information on 1972 auctions transacted on eBay.com during May–June 2004. The goal is to use these data to build a model that will distinguish competitive auctions from noncompetitive ones. A competitive auction is defined as an auction with at least two bids placed on the item being auctioned. The data include variables that describe the item (auction category), the seller (his or her eBay rating), and the auction terms that the seller selected (auction duration, opening price, currency, day of week of auction close). In addition, we have the price at which the auction closed. The goal is to predict whether or not the auction will be competitive.

Data Preprocessing. Create dummy variables for the categorical predictors. These include Category (18 categories), Currency (USD, GBP, euro), EndDay (Monday–Sunday), and Duration (1, 3, 5, 7, or 10 days). Split the data into training and validation datasets using a 60% : 40% ratio.

a. Create pivot tables for the average of the binary dependent variable (Competitive?) as a function of the various categorical variables (use the original variables, not the dummies). Use the information in the tables to reduce the number of dummies that will be used in the model. For example, categories that appear most similar with respect to the distribution of competitive auctions could be combined.

See R code