Generalized Linear Model

Name – Krupal Shah
Student ID – 200313719
Unity ID – khshah2

For the problem above, 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 –

```
Coefficients: (4 not defined because of singularities)
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -4.563e-01  1.379e+00  -0.331  0.74071
sellerRating                -4.447e-05  1.532e-05  -2.903  0.00370 **
ClosePrice                   1.053e-01  1.139e-02   9.241  < 2e-16 ***
OpenPrice                   -1.176e-01  1.230e-02  -9.559  < 2e-16 ***
Category_Computer           -2.388e-01  1.362e+00  -0.175  0.86086
Category_Automotive         -9.341e-01  1.368e+00  -0.683  0.49473
Category_Electronics         1.139e-02  1.387e+00   0.008  0.99345
`Category_Antique/Art/Craft` -3.016e-01  1.356e+00  -0.223  0.82391
`Category_Coins/Stamps`     -1.717e+00  1.416e+00  -1.213  0.22524
`Category_Home/Garden`      -1.499e-01  1.392e+00  -0.108  0.91428
`Category_Health/Beauty`    -1.904e+00  1.437e+00  -1.325  0.18512
Category_Photography               NA         NA      NA       NA
currency_US                  6.872e-01  2.482e-01   2.769  0.00562 **
currency_GBP                 2.053e+00  5.778e-01   3.553  0.00038 ***
currency_EUR                       NA         NA      NA       NA
Duration_5                   4.075e-02  3.176e-01   0.128  0.89790
Duration_3                  -4.915e-01  2.725e-01  -1.804  0.07127 .
Duration_1                  -6.647e-01  8.122e-01  -0.818  0.41314
Duration_10                        NA         NA      NA       NA
endDay_Mon                   7.467e-01  2.417e-01   3.089  0.00201 **
endDay_Fri                   3.094e-01  2.092e-01   1.479  0.13916
endDay_Thu                  -2.455e-01  5.574e-01  -0.441  0.65956
endDay_Sat                         NA         NA      NA       NA
---
```

From the summary of fit.all we can see that the predictor with the highest estimate is X_h = currency_GBP
After fitting a regression model using only currency_GBP as the predictor we get the following equation.
Thus, x = currency_GBP

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.11347    0.06056   1.873  0.06100 .
currency_GBP  0.63011    0.23355   2.698  0.00698 **
---
```

a. Probabilities - $Prob(Y = Yes \mid X_h = x) = p = \frac{1}{1+e^{-(0.11347+0.63011x)}}$

b. Odds - $Prob(Y = Yes) = \frac{p}{1-p} = e^{0.11347+0.63011x}$

c. Logit - $\ln \frac{p}{1-p} = 0.11347 + 0.63011x$

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) –

   The predictors with the highest absolute values are: currency_GBP , Category_Automotive, Category_Coins/Stamps and Category_Health/beauty.

   a. Logit - $\ln \frac{p}{1-p} = -0.4563 + 2.053 * currency\_GBP - 0.9341 * Category\_Automotive - 1.717 * Category\_Coins|Stamps - 1.904 * Category\_Health|Beauty$

   b. Odds - $\frac{p}{1-p} = e^z, where\ z = -0.4563 + 2.053 * currency\_GBP - 0.9341 * Category\_Automotive - 1.717 * Category\_Coins|Stamps - 1.904 * Category\_Health|Beauty$

   c. Probability - $p = \frac{1}{1+e^{-z}}$, $where\ z = -0.4563 + 2.053 * currency\_GBP - 0.9341 * Category\_Automotive - 1.717 * Category\_Coins|Stamps - 1.904 * Category\_Health|Beauty$

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. 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.

   Here $X_h = Currency\_GBP$ and the rest of the predictors are constant.

$$\frac{odds(X_1+1, X_2, \dots X_q)}{odds(X_1, X_2, \dots X_q)} = \frac{e^{\beta_0+\beta_1(X_1+1)+\beta_2X_2+\dots.+\beta_qX_q}}{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_qX_q}} = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_4X_4+\beta_1}}{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_4X_4}} = e^{\beta_1}$$

   Since, the estimate of $Currency\_GBP = 2.053, therefore\ e^{\beta_1} = 7.7912$

   This means that for a unit increase in response variable, $Currency\_GBP$, the response variable will change by the factor 7.7912 for logistic regression. Since the coefficient is positive, the odds increase exponentially.

   However, for linear regression the change is proportional to 2.053 times. That is, for a unit increase in $Currency\_GBP$, the response variable will increase by 2.053 times 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.

```
Model 1: Competitive ~ sellerRating + ClosePrice + OpenPrice + currency_US +
    currency_GBP + endDay_Mon
Model 2: Competitive ~ sellerRating + ClosePrice + OpenPrice + Category_Computer +
    Category_Automotive + Category_Electronics + `Category_Antique/Art/Craft` +
    `Category_Coins/Stamps` + `Category_Home/Garden` + `Category_Health/Beauty` +
    Category_Photography + currency_US + currency_GBP + currency_EUR +
    Duration_5 + Duration_3 + Duration_1 + Duration_10 + endDay_Mon +
    endDay_Fri + endDay_Thu + endDay_Sat
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1177     1231.8
2      1165     1178.9 12   52.862 4.359e-07 ***
---
```

The reduced model is built from taking statistically significant predictors. We performed chi-square test after fitting the reduced model and we found from the results that the p-value is 4.359e-07 which signifies that the difference is significant and both models are not equivalent.

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

```
Overdispersion test Obs.Var/Theor.Var Statistic p-value
      binomial data          0.4609032  545.2485       1
```

The dispersion of the model can be evaluated by the following formula:

$$\emptyset = \frac{Residual\ Deviance}{Degree\ of\ Freedom} = \frac{1232}{1177} = 1.046 \approx 1$$

Also, the overdispersion test was run on the data and

$$\frac{Obs.Variance}{Theoritical\ Variance} = 0.4609$$

And the p-value is also 1, and the value 0.4609 is not significantly different than 1, hence the model is not over dispersed.