Generalized Linear Models Homework 1 Logistic Regression

Question #1

The predictor with the highest regression coefficient – **Category_Photography**

```
> max_reg_coeff
[1] "CategoryPhotography"
> max(edf)
[1] 13.19277
> edf
                                  est
CategoryAutomotive          -1.034235e+00
CategoryBooks               -1.141407e+00
CategoryBusiness/Industrial -1.607676e-01
CategoryCoins/Stamps        -1.032930e+00
CategoryElectronics          5.954112e-01
CategoryEverythingElse      -2.137351e+00
CategoryHealth/Beauty       -1.547228e+00
CategoryPhotography          1.319277e+01
currencyGBP                  1.146657e+00
sellerRating                -3.133604e-05
Duration                     7.844235e-02
endDayMon                    9.238976e-01
endDayThu                   -8.489126e-02
ClosePrice                   8.520579e-02
OpenPrice                   -9.806807e-02
```

Equations after using category_photography as a single predictor

a) Probabilities – $Prob(Y = Yes | X_h = x) = 1/(1+ e^{-(-0.3604+13.19277*Category\_Photography)})$
b) Odds – $P(Y = Yes)/ (1 - P(Y = Yes)) = e^{(-0.3604+13.19277*Category\_Photography)}$
c) Logit = $\log(P(Y = Yes)/ (1 - P(Y = Yes))) = (-0.3604+13.19277*Category\_Photography)$

Question #2

The four predictors that have the highest absolute value estimates (as seen in the estimates dataframe edf) are –

CP = Category_Photography, CE = Category_EverythingElse, CH = Category_Health/Beauty, CUR = Currency_GBP

a) Equation for logit = -0.3604 +13.19277*CP – 2.1373*CE – 1.5472*CH + 1.1466*CUR
b) Odds = $e^{(-0.3604 +13.19277*CP – 2.1373*CE – 1.5472*CH + 1.1466*CUR)}$
c) Probability = $1/(1+ e^{-(-0.3604 +13.19277*CP – 2.1373*CE – 1.5472*CH + 1.1466*CUR)})$

Question #3

The Xh in the first question is Category_Photography. The coefficient of Xh is 13.19277.

The ratio, $odds(X_{h+1}, X_2 \dots X_q)/ odds(X_h, X_2 \dots X_q) = e^{13.19277} = 5.364 * 10^5$

In the case of logistic regression, one unit increase in the variable leads to the increase equal to the coefficient of that predictor in terms of log(odds) after holding all other predictors constant.

In the case of linear regression, one unit increase in the variable increases the response by 13.19277 after holding all other predictors constant.

Question #4

The statistically significant predictors with p-values less than the significance level (0.05) in the full model are –

CategoryAutomotive, CategoryBooks, CategoryCoins/Stamps, CategoryHealth/Beauty, currencyGBP, sellerRating, endDayMon, ClosePrice, OpenPrice

A reduced model is fit using these predictors. The reduced model can be compared with the full model using anova and Chi-square test. The p-value corresponding to the chi-square test is 5.691e-6 which provides strong evidence that the two models are not equivalent.

```
> anova(fit.reduced, fit.all, test='Chisq')
Analysis of Deviance Table

Model 1: `Competitive?` ~ Category_Automotive + Category_EverythingElse +
    `Category_Health/Beauty` + currency_GBP + currency_EUR +
    sellerRating + endDay_Mon + ClosePrice + OpenPrice
Model 2: `Competitive?` ~ Category + currency + sellerRating + Duration +
    endDay + ClosePrice + OpenPrice + `Category_Coins/Stamps` +
    `Category_Business/Industrial` + Category_Photography + Category_Electronics +
    `Category_Antique/Art/Craft` + Category_Automotive + Category_Books +
    `Category_Health/Beauty` + Category_EverythingElse + currency_EUR +
    currency_GBP + endDay_Mon + endDay_Fri + endDay_Thu
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1174     1243.7
2      1167     1207.1  7   36.556 5.691e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Question #5

The over-dispersion test results for the training data

```
> qcc.overdispersion.test(train_data1$`Competitive?`, size=s, type="binomial")

Overdispersion test Obs.Var/Theor.Var Statistic p-value
      binomial data         0.4621393  546.2486       1
```

P-value = 1

Observed variance/ theoretical variance = 0.4621

This infers that the data is not over-dispersed as the statistic is not significant.