Homework 1

Ankur Garg, agarg12@ncsu.edu

Q1.

The predictor with the highest estimate (in terms of absolute value) for its regression coefficient is:
CategoryPhotography
So, X_h = CategoryPhotography

Equations for fit.single using this Category predictor are as follows:

(a) Probability:

$$\text{Prob}(\text{Y} = \text{Yes}|X_h \ = \ \text{x}) \ = \frac{1}{1 + e^{-(0.1823+14.3837*CategoryPhotography)}}$$

(b) Odds = P(Y=Yes)/ P(Y=No)

$$\frac{Prob(Y \ = \ Yes)}{1 - P(Y = Yes)} \ = \ e^{(0.1823+14.3837*CategoryPhotography)}$$

(c) Logit = log(odds)

$$\log\left(\frac{Prob(Y \ = \ Yes)}{1 - P(Y = Yes)}\right) = \ (0.1823 + 14.3837 * CategoryPhotography)$$

Q2.

The four predictors for which the absolute value estimates of regression coefficients are highest are:

CategoryPhotography = CP, currencyEUR = EU, endDayMon = EDM, CategoryElectronics = CE

a. The equation for logit is:

$$logit = \ 0.3529363 + 14.58834 * CP + 1.913506 * EUR + 1.894083 * EDM + 1.769387 * CE$$

b. The equation for odds is:

$$odds = \ e^{logit}$$

$$odds = e^{0.3529363+14.58834*CP+1.913506*EUR+1.894083*EDM+1.769387*CE}$$

c. Probability as a function of predictors is:

$$Probability = \frac{odds}{1 + odds}$$

$$Probability = \frac{1}{1 + e^{-(0.3529363+14.58834*CP+1.913506*EUR+1.894083*EDM+1.769387*CE)}}$$

Q3.

The predictor with the highest estimate (absolute value) for its regression coefficient in fit.all is: CategoryPhotography (as mentioned in Q1)

So, X_h = CategoryPhotography, let coefficient of X_h = t

Let sum of terms related to all other predictors = S

$$\frac{odds(X_h + 1, X_2, X_3 \dots, X_q)}{odds(X_h, X_2, X_3 \dots, X_q)} = \frac{e^{S+t*(X_h+1)}}{e^{S+t*(X_h)}}$$

$$\frac{odds(X_h + 1, X_2, X_3 \dots, X_q)}{odds(X_h, X_2, X_3 \dots, X_q)} = e^{t}$$

here, the coefficient of X_h = CategoryPhotography = 14.58834

So, the ratio,

$$\frac{odds(X_h + 1, X_2, X_3 \dots, X_q)}{odds(X_h, X_2, X_3 \dots, X_q)} = e^{14.58834} = 2.16589 \times 10^{6}$$

For Logistic Regression, Unit change in a variable will lead to change equal to coefficient of that predictor in the log(odds).

If it was Linear Regression, it would be lead to change equal to coefficient of that predictor in the overall prediction of the model. So, in that case, it would lead to change in Y equal to 14.58834

Q4.

Significant predictors are the ones which have the corresponding p-value < 0.05 [or has a star in the summary of model fit.all]. These are the predictors:

CategoryAntique/Art/Craft, CategoryBusiness/Industrial, CategoryElectronics, currencyEUR, sellerRating, endDayFri, endDayMon, ClosePrice, OpenPrice

Comparing the two models: The two models can be compared using anova test. P-value returned for the Chisq test is 0.02176. Since it is quite low (less than α = 0.05), it is statistically significant. So, the Null hypothesis is ignored. That is, the two models are not equivalent.

Q5.

The dispersion of the model fit.reduced = Residual Deviance / Residual degrees of freedom = 1215/1173 = 1.0358

The over-dispersion test was run on the data (qcc.overdispersion.test)
Obs.Variance/Theoretical Variance = 0.4528323
And the p-value is 1.
So, the value 0.4528323 is not statistically significantly different from 1.
Therefore, the model is not over-dispersed.