Generalized Linear Model

Name: Rutvik Kolhe
Unity ID: rkolhe

Soultions:

Question1)
After bulding the logistic regression model for all predictors, we can infer that currency_GBP is the predictor with the highest estimate for its coefficient. We then build fit.single with using currency_GBP as the predictor. The equation of relation of Competitive? to currency_GBP is:

$$Competitive? \ = \ 0.2193 - 1.7233 * \ categoryEverythingElse$$

a)

The odds that Y=yes for the given set of X values is given by $\dfrac{P(Y = yes)}{1 - P(Y = yes)}$

The relationship of log-odds with the predicted variable is:

$$ln\left(\frac{P(Y = yes)}{1 - P(Y = yes)}\right) = \beta_o + \beta_1 * X_1 + \beta_2 * X_2$$

$$ln\ (odds) = 0.2193 - 1.7233 * \ categoryEverythingElse$$

$$odds \ = \ e^{0.2193-1.7233* \ categoryEverythingElse}$$

$$Let\ e^{0.2193-1.7233* \ categoryEverythingElse} \ = \ e^{z}$$

$$Hence,\ P(Y) \ = \ e^{z}\,(1 - P(Y))$$

$$P(Y)\left(1 - \ e^{z}\right) \ = \ e^{z}$$

$$P(Y) = \frac{e^{z}}{1 - e^{z}}$$

$$P(Y = yes) = \frac{1}{1 + e^{0.2193-1.7233* \ categoryEverythingElse}}$$

b)

$$ln\ (odds) = 0.2193 - 1.7233 * \ categoryEverythingElse$$

$$odds \ = \ e^{0.2193 - 1.7233 * \ categoryEverythingElse}$$

c)

$$Logit = ln\left(\frac{P(Y = yes)}{1 - P(Y = yes)}\right) = 0.2193 - 1.7233 * \ categoryEverythingElse$$

Question 2)

The top four predictors for fit_all model are: category_EverythingElse, currency_GBP, currency_US, category_Business/Industrial. Based on these predictors, we can express the different equations as follows:

a)

$$\begin{aligned} Logit \ = \ & -2.5518 * categoryEverythingElse + 2.202 * currencyGBP \\ & + 0.8917 * categoryBusinessIndustrial + 0.8735 * currencyUS \end{aligned}$$

b)

$$ln\ (odds) \ = \ \begin{matrix} -2.5518 * categoryEverythingElse + 2.202 * currencyGBP \\ +0.8917 * categoryBusinessIndustrial + 0.8735 * currencyUS \end{matrix}$$

$$odds \ = \ e^{\substack{-2.5518 * categoryEverythingElse + 2.202 * currencyGBP + \\ 0.8917 * categoryBusinessIndustrial + 0.8735 * currencyUS}}$$

c)

As shown in solution 1a)

$$ln\left(\frac{P(Y = yes)}{1 - P(Y = yes)}\right) = \beta_o + \beta_1 * X_1 + \beta_2 * X_2 ....$$

$$P(Y) = \frac{e^z}{1 - e^z}$$

$$P(Y) = \frac{1}{1 + e^{\substack{-\ (-2.5518 * categoryEverythingElse + 2.202 * currencyGBP + \\ 0.8917 * categoryBusinessIndustrial + 0.8735 * currencyUS\ )}}}$$

Question 3)

The highest predictor for fit_all is category_EverythingElse. The generalized equation for odds can be given as:

$$odds \ = \ e^{\beta_o+\beta_1*X_1+\beta_2*X_2}$$

Hence, the value of odds for fit_all is:

$$odds \ = e^{-2.5518*categoryEverythingElse+2.202*currencyGBP+ 0.8917*categoryBusinessIndustrial+0.8735*currencyUS}$$

Now, if we increse the value of currency_GBP by 1 and keep the value of coefficents constant, the corressponding equation for odds is:

$$odds^{'} \ = e^{-2.5518*\ \mathbf{(categoryEverythingElse\ +\ 1)}+2.202*currencyGBP+ 0.8917*categoryBusinessIndustrial+0.8735*currencyUS}$$

Hence the odds ratio is:

$$\frac{odds^{'}}{odds} \ = \ \frac{e^{-2.5518*\ \mathbf{(categoryEverythingElse\ +\ 1)}+2.202*currencyGBP+ 0.8917*categoryBusinessIndustrial+0.8735*currencyUS}}{e^{-2.5518*categoryEverythingElse+2.202*currencyGBP+ 0.8917*categoryBusinessIndustrial+0.8735*currencyUS}}$$

$$= \frac{e^{-2.5518*categoryEverythingElse+2.202*currencyGBP+ 0.8917*categoryBusinessIndustrial+0.8735*currencyUS} * e^{-2.5518}}{e^{-2.5518*categoryEverythingElse+2.202*currencyGBP+ 0.8917*categoryBusinessIndustrial+0.8735*currencyUS}}$$

$$= e^{-2.5518}$$

Therefore, if the value of currency_GBP increases by 1, the value of response changes by a factor of $e^{-2.5518}$.

If it was linear regression, then the value of response would change by the factor of 2.5518 (coefficient) times. Since, the value of logistic regression gives us the output of logit function, we have to find the value of $e^{\ value}$. Whereas in linear regression the coefficient output directly affects the response

Question 4)

We can use anova test to check if the two models- fit_reduced and fit_all are equivalent to each other or nont. After running anova test on the two models. The p_value obtained is 0.7086 . Since p-value is greater than 0.05, we can conclude that the two models do not significantly differ from each other.

Question 5)

Overdispersion occurs when the value of $\phi = \dfrac{Residual\ deviance}{Residual\ df} \gg 1$. In this case,

$\phi\ =\ $ 0.992 . Since the value of ϕ is close to 1, there is no overdispersion present in the model.

If there is any overdispersion, then we can use quasi-binomial distribution instead of binomial family distribution.