Generalized Linear Models Homework

Unity_id - ssamudr

Question1)

After training the logistic regression model for all predictors using the training data, we can infer that “Category_Coins/Stamps” is the predictor with the highest estimate for its coefficient. We then train model.single using only “Category_Coins/Stamps” as the predictor. The equation of relation of response variable “Competitive” to “Category_Coins/Stamps” is:

$$Competitive = -0.507 - 2.07233 * Category_Coins/Stamps$$

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

$$\log\left(P(Y = yes)/(1 - P(Y = yes))\right) = B_o + B_1 X_1$$

$$P(Y) = 1/\left(1 + e^{B_0 + B_1 x}\right)$$

$$P(Y = yes) = 1/\left(1 + e^{(-0.5 - 2.072 \cdot \text{categorycoins/stamps})}\right)$$

b)

$$\log\left(P(Y = yes)/(1 - P(Y = yes))\right) = B_o + B_1 X_1$$

$$\log(odds) = -0.507 - 2.072 \cdot \text{categorycoins/stamps}$$

$$odds = e^{-0.5 - 2.072 \cdot \text{categorycoins/stamps}}$$

c)

$$\text{logit} = \log(odds) = -0.5 - 2.072 \cdot \text{categorycoins/stamps}$$

Question 2)

The top four predictors for model_all model are: Category_Coins/Stamps, Category_Health/Beauty, Category_Clothing/Accessories, currency_GBP. Based on these predictors, we can express the different equations as follows:

a)

```
Logit = log(odds) = -0.5 - 2.072*Category_Coins/Stamps -
1.64*Category_Health/Beauty - 1.42*Category_Clothing/Accessories +
1.26*currency_GBP.
```

b)

```
Odds = e^(-0.5 - 2.072*Category_Coins/Stamps -
1.64*Category_Health/Beauty - 1.42*Category_Clothing/Accessories +
1.26*currency_GBP)
```

c)

```
P(Y=yes/X) = 1/(1+e^logit) = 1/(1+e^(-0.5 - 2.072*Category_Coins/Stamps
- 1.64*Category_Health/Beauty - 1.42*Category_Clothing/Accessories +
1.26*currency_GBP))
```

Question 3)

The highest predictor for model_all is Category_Coins/Stamps. The generalised equation for odds can be given as:

```
  Odds = e^(-0.5 - 2.072*Category_Coins/Stamps -
1.64*Category_Health/Beauty - 1.42*Category_Clothing/Accessories +
1.26*currency_GBP)
```

Now, if we increase the value of Category_Coins/Stamps by 1 and keep the value of coefficients constant, the corresponding equation for odds is:

```
Odds’ = e^(-0.5 - 2.072*(Category_Coins/Stamps+1) -
1.64*Category_Health/Beauty - 1.42*Category_Clothing/Accessories +
1.26*currency_GBP)
```

Hence the odds ratio is:

```
Odds/Odds' = e^(-0.5 - 2.072*Category_Coins/Stamps -
1.64*Category_Health/Beauty - 1.42*Category_Clothing/Accessories +
1.26*currency_GBP) / e^(-0.5 - 2.072*(Category_Coins/Stamps+1) -
1.64*Category_Health/Beauty - 1.42*Category_Clothing/Accessories +
1.26*currency_GBP)
```

odds'/odds = e^-2.072

Therefore, if the value of Category_Coins/Stamps increases by 1, the value of the response variable changes by a factor of e^-2.072. If it was linear regression, then the value of response would change by the factor of 2.072 (coefficient) times. Since, the value of logistic regression gives us the output of the logit function, we have to find the value of e . Whereas in linear regression the value 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 not. 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 ϕ = $\gg$ 1 . In this case, Residual deviance Residual df ϕ = 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.