TOPIC 2: Generalized Linear Models

1. X_h : predictor with the highest estimated (absolute value) of regression coeffs.

(a) Probabilities : Prob $(Y = Yes \mid X_h = x)$

→ $g(\mu_Y) = \text{logit}(P) = \log_e \left| \frac{P}{1-P} \right| = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_q X_q$

$$P = \frac{1}{1 + e^{-z}}$$

where $z = \beta_0 + \beta_1 x$.

In our case,

$x =$ Categoryeverythingelse : a dummy variable

or

$x =$ Category

So,

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 (\text{Category}^*))}}$$

(b) Odds : $\frac{\mu_Y}{1-\mu_Y} = \frac{P}{1-P} = e^{\beta_0 + \beta_1 (\text{Category}^*)}$

(c) Logit : $\log_e \left(\frac{P}{1-P} \right) = \beta_0 + \beta_1 (\text{Category}^*)$.

* Here we can also use Categoryeverythingelse as well. But since it is a dummy category

2. fit. all model with the top 4 predictors in terms of absolute value estimates

The four predictors with the highest absolute value estimates are:-

X_1: Category → Categoryeverythingelse

X_2: currency → currencyGBP

X_3: endDay ~~Sat-Wed~~ → endDay Sat_Wed

X_4: openPrice

(i) Probability, $P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)}}$

(ii) Odds: $e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}$

(iii) Logit: $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$.

3. X_n: predictor with highest estimate in fit.all

X_n = Category everything else or Category

Since its a categorical predictor it can take only 2 values 0 or 1.

So assuming $X_n = 1$ originally, we change to 0 by doing a single unit increase.

$$\frac{odds(X_n = 1)}{odds(X_n = 0)} = \frac{e^{\beta_0 + \beta_1 X_n + \dots \beta_q X_q}}{e^{\beta_0 + \beta_1 X_n + \dots \beta_q X_q}}$$

$$\Rightarrow \frac{\cancel{e^{\beta_0}} \cdot e^{\beta_1 X_n} \cdot \cancel{e^{\beta_2 X_2}} \dots \cancel{e^{\beta_q X_q}}}{\cancel{e^{\beta_0}} \cdot e^{\beta_1 X_n} \cdot \cancel{e^{\beta_2 X_2}} \dots \cancel{e^{\beta_q X_q}}}$$

$$\Rightarrow \frac{e^{\beta_1 \times 1}}{e^{\beta_1 \times 0}} \Rightarrow \frac{e^{\beta_1}}{1} \Rightarrow e^{\beta_1}$$

β_1 for category everything else $= -2.041$

So odds ratio $= e^{-2.041} = 0.1298$

4. We built a reduced model with 6 significant Predictors (everything except "Duration").

We then carried the anova(), chi-square version of the test. The test gave us a non-significant chi-squale value of 0.2783. which suggests that fit.reduced fits as well as fit.all.

5. We can test for overdispersion by computing the mean variance.

$$\text{Mean Variance} = \frac{\text{Residual Deviance}}{\text{Residual Degree of Freedom}} = \frac{1189.6}{1167} = 1.019$$

Since mean variance is close to 1, the model is well fitted.

THE END