CSC 591, Homework Generalized Linear Models
Student Id: 200261202
Mkhan8@ncsu.edu

Solution 1:

Let X_h be the predictor with the highest estimate (in terms of its absolute value) for its regression coefficient.

Regression coefficient X_h will highest coefficient is dummy_EverythingElse with a coefficient value of -2.127e+00. So, we can say that $X_h \rightarrow dummy\_EverythingElse$.

Using X_h as a single predictor for $fit.single$, the value of response variable

$$z = 0.12281 + (-2.32004 * \ dummy\_EverythingElse)$$

a) $Prob(Y = Yes \mid X_h = x) = \frac{1}{(1+e^{-(0.12281+(-2.32004* \ dummy\_EverythingElse))}\ )}$

b) $odds = \left(\frac{p}{1-p}\right) = e^z = e^{0.12281+(-2.32004* \ dummy\_EverythingElse)}$

c) $logit(p) = log\left(\frac{p}{1-p}\right) = \beta 0 + \beta 1 * x1 + \ldots + \beta k * xk = z$

$$logit(p) = 0.12281 + (-2.32004 * \ dummy\_EverythingElse)$$

Solution 2:

The top four coefficients of fit.all model are:

- dummy_EverythingElse (2.126912)
- dummy_GBP (1.739977)
- dummy_Health/Beauty (1.608681)
- dummy_Coins/Stamps (1.359195)

$$z = -0.6115 - 2.126912 * \text{dummy\_EverythingElse} + 1.739977 * \text{dummy\_GBP} - 1.608681 * \text{dummy\_Health/Beauty} - 1.359195 * \text{dummy\_Coins/Stamps}$$

a) $logit(p) = z = -0.6115 - 2.126912 * \text{dummy\_EverythingElse} + 1.739977 * \text{dummy\_GBP} - 1.608681 * \text{dummy\_Health/Beauty} - 1.359195 * \text{dummy\_Coins/Stamps}$

b) $odds = \left(\frac{p}{1-p}\right) = e^z = e^{-0.6115-2.126912*\text{dummy\_EverythingElse} +1.739977* \text{dummy\_GBP} -1.608681*\text{dummy\_Health/Beauty} -1.359195 *\text{dummy\_Coins/Stamps}}$

c) $Prob = \frac{1}{(1+e^{-z})} = \frac{1}{(1+e^{-(-0.6115-2.126912*dummy\_EverythingElse + 1.739977*dummy\_GBP - 1.608681*dummy\_Health/Beauty - 1.359195*dummy\_Coins/Stamps)})}$

Solution 3:

The highest coefficient X_h of the fit.all model is dummy_EverythingElse with a value of -2.126912. If we increase the value of X_h by 1 keeping all the other co-efficient constant, the value of log odd will increase by $e^{coeffient\ of\ X_h}$.

We know the odds is given by $odds(X_1, X_2, \dots, X_q) = e^z$ where $Z = \beta_0 + \beta_1 X_1 + \cdots . + \beta_q X_q$

If we increase the value of X_h by 1 → $X_h = X_h + 1$,

$Z_1 = \beta_0 + \beta_1 X_1 + \cdots + \beta_h (X_h + 1) + \cdots . + \beta_q X_q = \beta_h + \beta_0 + \beta_1 X_1 + \cdots + \beta_h X_h + \cdots . + \beta_q X_q = \beta_h z$

Where $\beta_h is\ the\ coefficient\ of\ X_h$

$$\frac{odds(X_1 + 1, X_2, \dots, X_q)}{odds(X_1, X_2, \dots, X_q)} = \frac{e^{z_1}}{e^z} = \frac{e^{\beta_h * z}}{e^z} = e^{\beta_h}$$

We know that $\beta_h = -2.126912$

$$\frac{odds(X_1 + 1, X_2, \dots, X_q)}{odds(X_1, X_2, \dots, X_q)} = e^{-2.126912} = 0.1192$$

In case of linear regression: $odds\left(X_1, X_2, \dots, X_q\right) = z = \beta_0 + \beta_1 X_1 + \cdots . + \beta_q X_q$

And $odds\left(X_1 + 1, X_2, \dots, X_q\right) = Z_1 = \beta_h z$

$\frac{odds(X_1+1, X_2, \dots, X_q)}{odds(X_1, X_2, \dots, X_q)} = \frac{(\beta_h z)}{z} = \beta_h$

In this case, the value will decrease by the value of the coefficient i.e. 2.126912

Solution 4:

The output of summary statistics of fit.all model:

```
Call:
glm(formula = competitive ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.5201  -0.9007   0.0001   0.8609   2.0405

Coefficients: (4 not defined because of singularities)
                                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                                          -6.115e-01  3.327e-01  -1.838  0.06603 .
sellerrating                                         -2.667e-05  1.467e-05  -1.818  0.06902 .
closeprice                                            1.295e-01  1.305e-02   9.924  < 2e-16 ***
openprice                                            -1.388e-01  1.372e-02 -10.114  < 2e-16 ***
dummy_SportingGoods                                  -4.162e-01  4.467e-01  -0.932  0.35153
dummy_Electronics                                     5.529e-01  6.323e-01   0.874  0.38189
dummy_EverythingElse                                 -2.127e+00  1.096e+00  -1.940  0.05232 .
`dummy_Coins/Stamps`                                 -1.359e+00  5.873e-01  -2.314  0.02064 *
`dummy_Health/Beauty`                                -1.609e+00  5.386e-01  -2.987  0.00282 **
dummy_Photography                                     2.480e-01  1.400e+00   0.177  0.85944
dummy_GBP                                             1.740e+00  5.451e-01   3.192  0.00141 **
dummy_5                                               6.544e-01  2.527e-01   2.590  0.00961 **
dummy_Mon                                             4.926e-01  2.302e-01   2.140  0.03236 *
dummy_Tue                                            -7.021e-02  3.021e-01  -0.232  0.81622
dummy_Thu                                            -7.904e-01  5.068e-01  -1.560  0.11886
`merged_Pottery/Glass_Automotive_Jewelry`            -8.154e-01  3.389e-01  -2.406  0.01614 *
`merged_Books_Clothing/Accessories_Toys/Hobbies`     -2.901e-01  3.067e-01  -0.946  0.34413
`merged_Antique/Art/Craft_Collectibles_Music/Movie/Game` -6.566e-02  2.836e-01  -0.232  0.81689
`merged_Home/Garden_Business/Industrial_Computer`            NA         NA      NA       NA
merged_US_EUR                                                NA         NA      NA       NA
merged_Sat_Fri                                        2.033e-01  2.026e-01   1.003  0.31566
merged_Wed_Sun                                               NA         NA      NA       NA
merged_3_7                                            1.527e-01  2.096e-01   0.728  0.46631
merged_1_10                                                  NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1637.9  on 1183  degrees of freedom
Residual deviance: 1188.1  on 1164  degrees of freedom
AIC: 1228.1

Number of Fisher Scoring iterations: 8
```

The significant coefficients from the summary statistics of fit.all model are:

- Openprice
- Closeprice
- dummy_Health/Beauty
- dummy_GBP
- dummy_5
- dummy_Mon
- merged_Pottery/Glass_Automotive_Jewelry
- dummy_Coins/Stamps

Model Comparison:

Accuracy:

Fit.all model: 80.5%

Fit.reduced: 80.2%

Chisquare test:

anova(fit.reduced, fit.all, test='Chisq') → 0.04615

- at $\alpha = 0.01$ and 0.001: the value 0.046 is non-significant. The null hypothesis is accepted, and the alternate hypothesis is ignored. Therefore, we can say that the models are equivalent.
- at $\alpha = 0.05$: the value 0.046 is significant. The null hypothesis is ignored. Therefore, we can say that the models are not equivalent.

Solution 5:

We know that,

For a well-fitting model: Residual Deviance ≈ Residual d.f. or we can say that

$$\frac{\text{Residual Deviance}}{Residual\ d.f} \approx 1$$

In case of fit.reduced model , we know the from summary statistic of the model that :

```
Residual deviance: 1188.1  on 1164  degrees of freedom
```

As Residual deviance ≈ residual degree of freedom, so we can say that the model is not over-dispersed and is a well fitted model.