Q88: Logistic Regression

Candus Shi

2023-05-07

Q88: Was there a time in the past 12 months when you needed to see a doctor but could not because of the cost?

We use a logistic regression model to test whether more transgenders answer "Yes" to this question than cisgenders.

H_0: $\beta_1 \neq 0$

H_1: $\beta_1 > 0$

where β_1 is the coefficient for the TRANS_CIS variable.

To test whether to add a new feature and/or possible interaction terms, we choose the term with the lowest p-value and update the model. We repeat this until the lowest p-value $> \alpha = 0.05$

Index of Variables

- Y (response variable): answer to Q88; binomial; No (0), Yes (1)
- G (variable of interest): TRANS_CIS; binomial; Cisgender (0), Transgender (1)
- I: HINC_I (household income imputated); continuous quantitative (standardized)
- A: AGE; discrete quantitative (standardized)

Setup

```
library(car)
```

```
## Loading required package: carData
```

```
library(flexmix)
```

```
## Loading required package: lattice
```

```
q88 = read.csv("q88.csv", header=T)
head(q88)
```

```
##   X TRANS_CIS Q88 HINC_I        AGE HINC_I_means HINC_I_strat
## 1 0         1   0     11  0.6959294    0.2619827            1
## 2 1         1   0      7 -0.9078519   -0.7376871            0
## 3 2         1   1      9 -1.6800429   -0.3167735            1
## 4 3         1   0     11 -2.0958380    0.2619827            1
## 5 4         1   0      8 -1.3830463   -0.5272303            0
## 6 5         1   0      4  0.6365301   -1.1059865            0
```

(V1) Baseline Model

$Y \sim \beta_0 + \beta_1 G$

```
q88logitv1 = glm(Q88 ~ TRANS_CIS, data=q88, family="binomial")
summary(q88logitv1)
```

```
##
## Call:
## glm(formula = Q88 ~ TRANS_CIS, family = "binomial", data = q88)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8021  -0.4711  -0.4711  -0.4711   2.1230
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.1425     0.0969  -22.11  < 2e-16 ***
## TRANS_CIS     1.1736     0.1674    7.01 2.39e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1123.5  on 1401  degrees of freedom
## Residual deviance: 1077.9  on 1400  degrees of freedom
## AIC: 1081.9
##
## Number of Fisher Scoring iterations: 4
```

```
BIC(q88logitv1)
```

```
## [1] 1092.347
```

There is no multicollinearity to consider because there is only one feature. With this model, we reject H_0 because $\beta_1 = 1.1736$ with a p-value of 2.39×10^{-12}. Given a standard error of 0.1674, its 95% CI is $[0.8430526, \infty)$.

(V2a) Household Income Imputated, 14 Ordinal Categories

$Y \sim \beta_0 + \beta_1 G + \gamma_1 I_1 + \gamma_2 I_2 + ... + \gamma_{13} I_{13}$

```
q88$HINC_I = factor(q88$HINC_I)
q88logitv2a = glm(Q88 ~ TRANS_CIS + HINC_I, data=q88, family="binomial")
summary(q88logitv2a)
```

```
##
## Call:
## glm(formula = Q88 ~ TRANS_CIS + HINC_I, family = "binomial",
##     data = q88)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1484  -0.6131  -0.4300  -0.1969   2.8116
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.47244    0.52627  -2.798 0.005144 **
## TRANS_CIS    0.92693    0.17644   5.253 1.49e-07 ***
## HINC_I1     -0.05303    0.67201  -0.079 0.937103
## HINC_I2      0.15433    0.61926   0.249 0.803192
## HINC_I3      0.12145    0.58575   0.207 0.835744
## HINC_I4      0.47689    0.58215   0.819 0.412676
## HINC_I5     -0.10352    0.61607  -0.168 0.866561
## HINC_I6      0.08138    0.61437   0.132 0.894625
## HINC_I7     -0.28157    0.57931  -0.486 0.626933
## HINC_I8     -0.51013    0.59007  -0.865 0.387295
## HINC_I9     -0.86227    0.59565  -1.448 0.147723
## HINC_I10    -0.84110    0.59508  -1.413 0.157531
## HINC_I11    -1.02593    0.58531  -1.753 0.079640 .
## HINC_I12    -1.32824    0.59195  -2.244 0.024844 *
## HINC_I13    -2.46073    0.72795  -3.380 0.000724 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1123.5  on 1401  degrees of freedom
## Residual deviance: 1004.3  on 1387  degrees of freedom
## AIC: 1034.3
##
## Number of Fisher Scoring iterations: 6
```

```
vif(q88logitv2a)
```

```
##               GVIF Df GVIF^(1/(2*Df))
## TRANS_CIS 1.038895  1        1.019262
## HINC_I    1.038895 13        1.001469
```

```
BIC(q88logitv2a)
```

```
## [1] 1112.986
```

As shown by the VIFs, there is no significant multicollinearity. With this model, we reject H_0 because $\beta_1 = 0.92693$ with a p-value of 1.49×10^{-7}. Given a standard error of 0.17644, its 95% CI is $[0.5785093, \infty)$.

(V2b) Household Income Imputated, Means of Bins

$Y \sim \beta_0 + \beta_1 G + \gamma I$

Since the v2a model had split the income variable into 14 categories, this led to its BIC score increasing, while its AIC decreased. We try to resolve this by taking the mean of each income category and standardizing to create a continuous quantitative income feature. (Suraj's idea)

```
q88logitv2b = glm(Q88 ~ TRANS_CIS + HINC_I_means, data=q88, family="binomial")
summary(q88logitv2b)
```

```
##
## Call:
## glm(formula = Q88 ~ TRANS_CIS + HINC_I_means, family = "binomial",
##     data = q88)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0757  -0.6045  -0.4494  -0.2378   2.6759
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.2826     0.1082 -21.100  < 2e-16 ***
## TRANS_CIS      0.9264     0.1738   5.330 9.84e-08 ***
## HINC_I_means  -0.7544     0.1029  -7.331 2.28e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1123.5  on 1401  degrees of freedom
## Residual deviance: 1012.1  on 1399  degrees of freedom
## AIC: 1018.1
##
## Number of Fisher Scoring iterations: 5
```

```
vif(q88logitv2b)
```

```
##    TRANS_CIS HINC_I_means
##      1.01758      1.01758
```

```
BIC(q88logitv2b)
```

```
## [1] 1033.799
```

Again, we don't see multicollinearity as the VIF scores are near 1. We also reject H_0 because $\beta_1 = 0.9264$ with a p-value of 9.84×10^{-8}. Given a standard error of 0.1738, its 95% CI is $[0.5831262, \infty)$.

(V2c) Household Income Imputated, 3 Ordinal Categories

$Y \sim \beta_0 + \beta_1 G + \gamma_1 I_1 + \gamma_2 I_2$

We show another attempt to maintain HINC_I as an ordinal feature, but by reducing the number of categories by stratifying by low, middle and high income. Low income is defined to be less than $50,000; middle income is defined to be from $50,000 to $100,000 and high income is defined to be more than $100,000. Note that this survey was conducted in from 2016-2018.

```
q88$HINC_I_strat = factor(q88$HINC_I_strat)
q88logitv2c = glm(Q88 ~ TRANS_CIS + HINC_I_strat, data=q88, family="binomial")
summary(q88logitv2c)
```

```
##
## Call:
## glm(formula = Q88 ~ TRANS_CIS + HINC_I_strat, family = "binomial",
##     data = q88)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9522  -0.6274  -0.4161  -0.2833   2.5438
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.5254     0.1197 -12.746  < 2e-16 ***
## TRANS_CIS       0.9695     0.1729   5.608 2.05e-08 ***
## HINC_I_strat1  -0.8777     0.1904  -4.610 4.02e-06 ***
## HINC_I_strat2  -1.6699     0.2637  -6.333 2.40e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1123.5  on 1401  degrees of freedom
## Residual deviance: 1017.9  on 1398  degrees of freedom
## AIC: 1025.9
##
## Number of Fisher Scoring iterations: 5
```

```
vif(q88logitv2c)
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## TRANS_CIS    1.011777  1        1.005871
## HINC_I_strat 1.011777  2        1.002931
```

```
BIC(q88logitv2c)
```

```
## [1] 1046.891
```

(V2d) Household Income Imputated, Means + Interaction

$Y \sim \beta_0 + \beta_1 G + \gamma I + \delta GI$

Since the means of the imputated household income was the best way to represent household income in terms of models scoring, we investigate whether we should add the interaction between TRANS_CIS and HINC_I.

```
q88logitv2d = glm(Q88 ~ TRANS_CIS + HINC_I_means + TRANS_CIS:HINC_I_means, data=q88, family="binomial")
summary(q88logitv2d)
```

```
##
## Call:
## glm(formula = Q88 ~ TRANS_CIS + HINC_I_means + TRANS_CIS:HINC_I_means,
##     family = "binomial", data = q88)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0827  -0.6035  -0.4442  -0.2398   2.6697
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.27917    0.11126 -20.486  < 2e-16 ***
## TRANS_CIS                 0.90943    0.22076   4.120 3.80e-05 ***
## HINC_I_means             -0.74628    0.12103  -6.166 7.01e-10 ***
## TRANS_CIS:HINC_I_means -0.02896    0.22996  -0.126      0.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1123.5  on 1401  degrees of freedom
## Residual deviance: 1012.0  on 1398  degrees of freedom
## AIC: 1020
##
## Number of Fisher Scoring iterations: 5
```

```
vif(q88logitv2d)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##              TRANS_CIS           HINC_I_means TRANS_CIS:HINC_I_means
##               1.638341               1.409899               2.144995
```

```
BIC(q88logitv2d)
```

```
## [1] 1041.028
```

Since the p-value of the interaction term is 0.9, we do not need to include it in the following iterations of the model. As a result, we determine that for the income variable, V2b is the best model.

(V3a) Age

$Y \sim \beta_0 + \beta_1 G + \gamma A$

First, we do a baseline model with just our variable of interest and the new AGE feature.

(V3b) Household Income Imputated, Means + Age

Model Metrics

Model	β_1 Estimate	β_1 95% CI	p-value	AIC	BIC	Highest VIF
V1	1.1736	$[0.8431, \infty)$	2.39×10^{-12}	1081.9	1092.347	N/A
V2a	0.92693	$[0.5785, \infty)$	1.49×10^{-7}	1034.3	1112.986	1.038895
V2b	0.9264	$[0.5831, \infty)$	9.84×10^{-8}	1018.1	1033.799	1.01758
V2c	0.9695	$[0.6281, \infty)$	2.05×10^{-8}	1025.9	1046.891	1.011777
V2d	0.90943	$[0.4589, \infty)$	3.08×10^{-5}	1020	1041.028	1.638341