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 WEIGHT_CISGENDER_TRANSPOP Q88 HINC_I        AGE RACE_RECODE_CAT5
## 1 0         1               0.022039215   0     11  0.6959294                1
## 2 1         1               0.008485489   0      7 -0.9078519                1
## 3 2         1               0.015764496   1      9 -1.6800429                1
## 4 3         1               0.035655390   0     11 -2.0958380                1
## 5 4         1               0.041801889   0      8 -1.3830463                3
## 6 5         1               0.021335387   0      4  0.6365301                4
##   Q93 SEX GENDER_IDENTITY POVERTYCAT_I Q99 Q200 Q205_I GEDUCATION HINC_I_means
## 1   3   0               4           4   0    0      2          4    0.2619827
## 2   4   1               3           3   0    0      2          5   -0.7376871
## 3   4   1               3           4   1    0      2          3   -0.3167735
```

```
## 4   3   1              3            4   0    0      4          1    0.2619827
## 5   4   1              3            4   0    0      1          4   -0.5272303
## 6   4   1              3            2   0    1      1          4   -1.1059865
##   HINC_I_strat CURRENT_SEX
## 1            1           1
## 2            0           0
## 3            1           0
## 4            1           0
## 5            0           0
## 6            0           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
## -0.7933  -0.4719  -0.4719  -0.4719   2.1215
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.13905    0.09816 -21.792  < 2e-16 ***
## TRANS_CIS    1.14424    0.17008   6.728 1.73e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1090.3  on 1363  degrees of freedom
## Residual deviance: 1048.2  on 1362  degrees of freedom
## AIC: 1052.2
##
## Number of Fisher Scoring iterations: 4
```

```
BIC(q88logitv1)
```

```
## [1] 1062.625
```

```
confint(q88logitv1, level=0.9)
```

```
## Waiting for profiling to be done...
```

```
##                    5 %      95 %
## (Intercept) -2.3040556 -1.980920
## TRANS_CIS    0.8626819  1.422719
```

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
## -1.1577  -0.6146  -0.4176  -0.2008   2.7977
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.46317    0.52583  -2.783 0.005393 **
## TRANS_CIS    0.90242    0.17954   5.026    5e-07 ***
## HINC_I1     -0.01474    0.67323  -0.022 0.982534
## HINC_I2      0.11089    0.63344   0.175 0.861030
## HINC_I3      0.07765    0.58740   0.132 0.894837
## HINC_I4      0.51414    0.58247   0.883 0.377405
## HINC_I5     -0.10753    0.61552  -0.175 0.861312
## HINC_I6      0.10058    0.61430   0.164 0.869941
## HINC_I7     -0.31851    0.58115  -0.548 0.583647
## HINC_I8     -0.48206    0.58975  -0.817 0.413700
## HINC_I9     -0.93267    0.59980  -1.555 0.119953
## HINC_I10    -0.84535    0.59459  -1.422 0.155099
## HINC_I11    -1.00224    0.58497  -1.713 0.086654 .
## HINC_I12    -1.36466    0.59629  -2.289 0.022103 *
## HINC_I13    -2.43032    0.72770  -3.340 0.000839 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1090.27  on 1363  degrees of freedom
## Residual deviance:  976.27  on 1349  degrees of freedom
## AIC: 1006.3
##
## Number of Fisher Scoring iterations: 6
```

```
vif(q88logitv2a)
```

```
##               GVIF Df GVIF^(1/(2*Df))
## TRANS_CIS 1.040753  1        1.020173
## HINC_I    1.040753 13        1.001538
```

```
BIC(q88logitv2a)
```

```
## [1] 1084.547
```

```
confint(q88logitv2a, level=0.9)
```

```
## Waiting for profiling to be done...
##                    5 %        95 %
## (Intercept) -2.4027097 -0.64726334
## TRANS_CIS    0.6051514  1.19629022
## HINC_I1     -1.1109528  1.12759176
## HINC_I2     -0.9085954  1.19801611
## HINC_I3     -0.8543801  1.10104547
## HINC_I4     -0.4072967  1.53153340
## HINC_I5     -1.0944114  0.95391719
## HINC_I6     -0.8826269  1.16109803
## HINC_I7     -1.2396268  0.69578496
## HINC_I8     -1.4201976  0.54362857
## HINC_I9     -1.8918253  0.10573665
## HINC_I10    -1.7935678  0.18656968
## HINC_I11    -1.9318811  0.01662191
## HINC_I12    -2.3174335 -0.33115727
## HINC_I13    -3.6724054 -1.23307404
```

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
## -1.0674  -0.6050  -0.4493  -0.2372   2.6777
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.2831     0.1098 -20.794  < 2e-16 ***
## TRANS_CIS      0.9026     0.1765   5.113 3.16e-07 ***
## HINC_I_means  -0.7570     0.1050  -7.207 5.73e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1090.27  on 1363  degrees of freedom
## Residual deviance:  984.51  on 1361  degrees of freedom
## AIC: 990.51
##
## Number of Fisher Scoring iterations: 5
```

```
vif(q88logitv2b)
```

```
##    TRANS_CIS HINC_I_means
##     1.017131     1.017131
```

```
BIC(q88logitv2b)
```

```
## [1] 1006.16
```

```
confint(q88logitv2b, level=0.9)
```

```
## Waiting for profiling to be done...
```

```
##                     5 %      95 %
## (Intercept)  -2.4687307 -2.107177
## TRANS_CIS     0.6103289  1.191497
## HINC_I_means -0.9343414 -0.588411
```

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
## -0.9451  -0.6286  -0.4152  -0.2814   2.5490
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.5211     0.1210 -12.568  < 2e-16 ***
## TRANS_CIS       0.9466     0.1756   5.391 6.99e-08 ***
## HINC_I_strat1  -0.8864     0.1925  -4.604 4.14e-06 ***
## HINC_I_strat2  -1.6879     0.2705  -6.241 4.36e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1090.3  on 1363  degrees of freedom
## Residual deviance:  989.2  on 1360  degrees of freedom
## AIC: 997.2
##
## Number of Fisher Scoring iterations: 5
```

```
vif(q88logitv2c)
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## TRANS_CIS    1.010662  1        1.005317
## HINC_I_strat 1.010662  2        1.002655
```

```
BIC(q88logitv2c)
```

```
## [1] 1018.078
```

```
confint(q88logitv2c, level=0.9)
```

```
## Waiting for profiling to be done...
```

```
##                       5 %       95 %
## (Intercept)   -1.7239695 -1.3255454
## TRANS_CIS      0.6559279  1.2340252
## HINC_I_strat1 -1.2092651 -0.5749387
## HINC_I_strat2 -2.1556712 -1.2620068
```

(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
## -1.0640  -0.6055  -0.4490  -0.2362   2.6807
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -2.28479    0.11355 -20.122  < 2e-16 ***
## TRANS_CIS               0.91069    0.22232   4.096 4.20e-05 ***
## HINC_I_means           -0.76093    0.12427  -6.123 9.17e-10 ***
## TRANS_CIS:HINC_I_means  0.01383    0.23250   0.059    0.953
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1090.3  on 1363  degrees of freedom
## Residual deviance:  984.5  on 1360  degrees of freedom
## AIC: 992.5
##
## 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
##               1.614710               1.422856               2.129968
```

```
BIC(q88logitv2d)
```

```
## [1] 1013.375
```

```
confint(q88logitv2d)
```

```
## Waiting for profiling to be done...
```

```
##                             2.5 %     97.5 %
## (Intercept)            -2.5167349 -2.0704796
## TRANS_CIS               0.4574148  1.3339559
## HINC_I_means           -1.0136396 -0.5253883
## TRANS_CIS:HINC_I_means -0.4575972  0.4580038
```

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.

```
q88logitv3a = glm(Q88 ~ TRANS_CIS + AGE, data=q88, family="binomial")
summary(q88logitv3a)
```

```
##
## Call:
## glm(formula = Q88 ~ TRANS_CIS + AGE, family = "binomial", data = q88)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0146  -0.5588  -0.4169  -0.3590   2.3554
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.12836    0.10036 -21.207  < 2e-16 ***
## TRANS_CIS    0.63722    0.19013   3.351 0.000804 ***
## AGE         -0.52264    0.08504  -6.146 7.95e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1090.3  on 1363  degrees of freedom
## Residual deviance: 1009.9  on 1361  degrees of freedom
## AIC: 1015.9
##
## Number of Fisher Scoring iterations: 5
```

```
vif(q88logitv3a)
```

```
## TRANS_CIS       AGE
##  1.201074  1.201074
```

```
BIC(q88logitv3a)
```

```
## [1] 1031.522
```

```
confint(q88logitv3a)
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %    97.5 %
## (Intercept) -2.3303367 -1.936546
## TRANS_CIS    0.2613685  1.007399
## AGE         -0.6904421 -0.356730
```

This model is clearly worse than the model with just TRANS_CIS and HINC_I. It also has a higher p-value than that of the HINC_I_means covariate Therefore, we will add HINC_I_means to the final model.

(V3b) Household Income Imputated, Means + Age

$Y \sim \beta_0 + \beta_1 G + \gamma_1 I + \gamma_2 A$

We investigate whether adding age builds a better model.

```
q88logitv3b = glm(Q88 ~ TRANS_CIS + HINC_I_means + AGE, data=q88, family="binomial")
summary(q88logitv3b)
```

```
##
## Call:
## glm(formula = Q88 ~ TRANS_CIS + HINC_I_means + AGE, family = "binomial",
##     data = q88)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3193  -0.5412  -0.3862  -0.2435   2.7704
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.26213    0.11158 -20.274  < 2e-16 ***
## TRANS_CIS     0.35481    0.19873   1.785   0.0742 .
## HINC_I_means -0.77626    0.10583  -7.335 2.22e-13 ***
## AGE          -0.53978    0.08549  -6.314 2.71e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1090.27  on 1363  degrees of freedom
## Residual deviance:  943.87  on 1360  degrees of freedom
## AIC: 951.87
##
## Number of Fisher Scoring iterations: 5
```

```
vif(q88logitv3b)
```

```
##    TRANS_CIS HINC_I_means          AGE
##     1.231042     1.022147     1.208025
```

```
BIC(q88logitv3b)
```

```
## [1] 972.7464
```

```
confint(q88logitv3b)
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %     97.5 %
## (Intercept)  -2.48813575 -2.0501232
## TRANS_CIS    -0.03863903  0.7411574
## HINC_I_means -0.99023881 -0.5747225
## AGE          -0.70862801 -0.3731409
```

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
V3a	0.68039	$[0.3116, \infty)$	0.000267	1045.5	1061.237	1.195264
V3b						