**MTH 522 PROJECT 2: Logistic Regression and Its Application in R**

Student Name: Venkata Vamshi Krishanam Raju, Mudenuri

Course: Advance Mathematical Stats

Professor: Gary Davis

**Contents**

[DATA 1(College Students Completing a Preliminary Year) 3](#_Toc130222144)

[Issues 3](#_Toc130222145)

[Data Collection 3](#_Toc130222146)

[Data-Coding 3](#_Toc130222147)

[Method 6](#_Toc130222148)

[R-Code 6](#_Toc130222149)

[R-Output 6](#_Toc130222150)

[Findings/Discussion 7](#_Toc130222151)

[DATA 2(Heath Health Data) 8](#_Toc130222152)

[The issues 8](#_Toc130222153)

[Data Collection 8](#_Toc130222154)

[Method 8](#_Toc130222155)

[Case 1 8](#_Toc130222156)

[R Code 9](#_Toc130222157)

[R Output 9](#_Toc130222158)

[Case 2 11](#_Toc130222159)

[R Code 11](#_Toc130222160)

[R Output 11](#_Toc130222161)

[Findings/Discussion 12](#_Toc130222162)

[Comparison 13](#_Toc130222163)

# **DATA 1(College Students Completing a Preliminary Year)**

## **Issues**

The data set named preliminary college year has 199 factors or variables which are assumed to affect the success and failure of finishing a course by students taking a preliminary year. The value 1 indicates that a student completed a preliminary year successfully. If a student fails to complete a preliminary year, they are given the value 0. This research assumes that all the factors contribute to the outcome. However, some factors are more useful in predicting student outcomes than other variables. Does the question now become which of the variables is important as compared to others? This paper thus aims at finding out which factors are important in predicting the success or failure of a student taking a preliminary year by fitting a multiple logistic regression model using the r statistical software.

## **Data Collection**

A sample of 106 students taking the preliminary year was taken to check whether they successfully completed the preliminary year or not. The participants were asked to give their Gender, Sat score, high school gap score, whether they are athletes or not, resident or not, Educational stress, Receptivity to academic excellence, receptivity to career guidance, receptivity to social engagement, receptivity to finance guidance, receptivity to personal guidance, Desire to transfer, eligibility to Pell grant, the process to drop out, whether they attended an orientation or not, F17GPA, S18GPA, CUMGPA, whether they completed event requirement or not, whether they completed community service requirement, the number of peer mentor meeting attended, number of workshops attended and their desire to transfer. The responses were then recorded in an excel spreadsheet.

## **Data-Coding**

The data was coded as follows:

|  |  |  |
| --- | --- | --- |
| Name | New name | Value Code |
| High School GPA | HSGPA | value |
| SAT Score | SAT Score | value |
| Gender | Gender | (1=Female,0=Male) |
| Pell Grant Eligible? | PGE | (1=yes, 0=no) |
| Attended Orientation? | AED | (1=yes, 0=no) |
| Resident/Commuter | Resident Commuter | (1=resident, 0=commuter) |
| Athlete? | Athlete | (1=yes, 0=no) |
| Dropout Proneness | DP | Percentile before start of semester |
| Educational Stress | ES | Percentile before start of semester |
| Receptivity to Academic Assistance | RAA | Percentile before start of semester |
| Receptivity to Career Guidance | RCG | Percentile before start of semester |
| Receptivity to Financial Guidance | RFG | Percentile before start of semester |
| Completed Campus Event Requirement | CCER | (1=yes, 0=no) |
| Number of Peer Mentor Meetings Attended | NFAMA | Percentile before start of semester |
| Number of Workshops Attended | NWA | Percentile before start of semester |
| F17 GPA | F17GPA | value |
| S18 GPA | S18GPA | value |
| CUM GPA | CUMGPA | value |
| Number of Credits Earned | NCE | Percentile before start of semester |
| Completed Connect? | CCER | (1=yes, 0=no) |

## **Method**

This research incorporated the use of the r multiple logistic regression model.

## **R-Code**

glm (formula = CC ~ HSGPA + SAT Score + Gender + PGE + AO + Resident Commuter +

Athlete + DP + PAD + EI + RAA + RCG + RFG + CCER + NPMMA +

NWA + F17GPA + S18GPA + CUMGPA + NCE, data = Preliminary College Year Clean Data)

## **R-Output**

Deviance Residuals:

Min 1Q Median 3Q Max

-0.75370 -0.17321 0.01464 0.14787 0.40606

Coefficients:

Estimate Std. Error t value PR (>|t|)

(Intercept) -0.6100376 0.6414511 -0.951 0.344319

HSGPA -0.0103316 0.1183655 -0.087 0.930653

SAT Score 0.0003800 0.0003541 1.073 0.286259

Gender -0.0956251 0.0578587 -1.653 0.102118

PGE 0.0096833 0.0541709 0.179 0.858561

AO -0.2518371 0.1792207 -1.405 0.163656

Resident Commuter -0.0408109 0.0861289 -0.474 0.636846

Athlete -0.0218132 0.0629485 -0.347 0.729815

DP 0.0007980 0.0013996 0.570 0.570115

PAD -0.0006277 0.0014459 -0.434 0.665292

EI 0.0013045 0.0010459 1.247 0.215771

RAA 0.0012515 0.0011688 1.071 0.287353

RCG 0.0001483 0.0010233 0.145 0.885112

RFG -0.0014102 0.0011258 -1.253 0.213840

CCER 0.0792237 0.0655180 1.209 0.229983

NPMMA 0.0064306 0.0193314 0.333 0.740227

NWA 0.0462384 0.0251488 1.839 0.069510.

F17GPA 0.0100665 0.0399643 0.252 0.801743

S18GPA 0.0595533 0.0503830 1.182 0.240536

CUMGPA 0.1320531 0.0552123 2.392 0.019002 \*

NCE 0.0241736 0.0070020 3.452 0.000872 \*\*\*

---

Signify. Codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘’ 1

(Dispersion parameter for Gaussian family taken to be 0.05302632)

Null deviance: 22.2476 on 104 degrees of freedom

Residual deviance: 4.4542 on 84 degrees of freedom

(3 observations deleted due to missingness)

AIC: 10.165

Number of Fisher Scoring iterations: 2

## **Findings/Discussion**

The results from the multiple logistic regression model have indicated the three variables (number of credits earned, CUMGPA and number of workshops attended) have a p-value which is less than 0.05. This means that they have a bigger impact in predicting the complete success or failure of a student taking a preliminary year. However, the other variables have p-values which are bigger than 0.05 which shows that they are of less importance in predicting the research outcome.

# **DATA 2(Heath Health Data)**

## **The issues**

Depending on several factors, patients usually seek medication after different days of being ill. This research, therefore, works to find out which factors affect the number of days a person takes before undergoing medication using a multiple logistic regression model. The second data set analyses the factors which affect a person to sick medical treatment early or delays before seeking medical treatment using multiple logistic regression. This data set analysis includes two cases. In the first case, a person takes 2 days or less to seek medical treatment or takes longer than 2 days to seek medical treatment and in the second case, a person sacks treatment on or less than the cohort average delay days “1” or takes longer than the cohort average number of days to sick treatment. A logistic model is built, and a comparison is done based on the two cases.

## **Data Collection**

A sample of 406 patients suffering from heart health conditions participated in this survey... The participants were asked to give their age, gender, education, ethnicity, marital, living with, teaching, palpitations, orthopnea, chest pain, nausea, cough, fatigue, dyspnea, edema, PND, tight shoes, weight gain, DOE, and the number of delay days it took them to seek medication were. All the information was then recorded in an excel spreadsheet for analysis.

## **Method**

This research incorporated the use of the r multiple logistic regression model.

## **Case 1:**

In the first case, a participant takes 2 days or less to get treatment or takes longer than 2 days to go for treatment.

## **R Code**

Call:

glm (formula = delay days ~ ID + Age + Gender + Ethnicity + Marital +

Live with + Education + palpitations + orthopnea + chest pain +

nausea + cough + fatigue + dyspnea + edema + PND + tight shoes +

weight gain + DOE, family = binomial, data = Heart\_Health\_Data)

## **R Output**

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7696 -1.0946 -0.7423 1.1348 2.0418

Coefficients:

Estimate Std. Error z value PR (>|z|)

(Intercept) -1.3566854 1.1892796 -1.141 0.2540

ID 0.0003654 0.0010623 0.344 0.7309

Age -0.0078037 0.0092433 -0.844 0.3985

Gender 0.1639839 0.2144944 0.765 0.4446

Ethnicity 0.2333795 0.2022291 1.154 0.2485

Marital 0.0046471 0.1748522 0.027 0.9788

Live with 0.3063775 0.2606683 1.175 0.2399

Education -0.0136398 0.0773701 -0.176 0.8601

Palpitations -0.0714948 0.1251206 -0.571 0.5677

Orthopnea 0.0022697 0.1153733 0.020 0.9843

Chest pain -0.0936855 0.1254306 -0.747 0.4551

Nausea 0.1114939 0.1332388 0.837 0.4027

Cough 0.2214852 0.1114380 1.988 0.0469 \*

Fatigue 0.2215776 0.1367478 1.620 0.1052

Dyspnea -0.1013658 0.1338821 -0.757 0.4490

Edema 0.2146256 0.1204540 1.782 0.0748.

PND 0.2026615 0.1103181 1.837 0.0662.

Tight shoes -0.2160773 0.1277739 -1.691 0.0908.

Weight gain -0.1211072 0.1112706 -1.088 0.2764

DOE 0.1048633 0.1225494 0.856 0.3922

---

Signify. Codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 559.82 on 403 degrees of freedom

Residual deviance: 533.51 on 384 degrees of freedom

(2 observations deleted due to missingness)

AIC: 573.51

Number of Fisher Scoring iterations: 4

The results from the multiple logistic regression model have indicated that only one variable (Cough) has a p-value which is less than 0.05. This means that the variable is statistically significant (has a bigger impact) in predicting the number of days it takes a person to seek medical treatment. However, the other variables have p-values which are bigger than 0.05 which shows that they are of less importance in predicting the research outcome. Thus, the conclusion is that only one variable (cough) is of more importance in predicting the dependent variable compared to the other predictors in the first case.

## **Case 2:**

In the second case, a person goes for treatment on or less than the cohort average delay days “1” or takes longer than the cohort average number of days to get treatment.

## **R Code**

Call:

glm (formula = delay days ~ ID + Age + Gender + Ethnicity + Marital +

Live with + Education + palpitations + orthopnea + chest pain +

nausea + cough + fatigue + dyspnea + edema + PND + tight shoes +

weight gain + DOE, family = binomial, data = Heart\_Health\_Data)

## **R Output**

Deviance Residuals:

Min 1Q Median 3Q Max

-2.0737 -1.1432 0.6720 0.8965 1.6010

Coefficients:

Estimate Std. Error z value PR (>|z|)

(Intercept) 1.0462456 1.2642382 0.828 0.40791

ID 0.0003628 0.0011384 0.319 0.74995

Age -0.0157635 0.0101583 -1.552 0.12071

Gender -0.1070903 0.2303146 -0.465 0.64195

Ethnicity 0.1039926 0.2136078 0.487 0.62637

Marital -0.1543684 0.1900889 -0.812 0.41674

Live with 0.2526813 0.2751404 0.918 0.35842

Education -0.0763146 0.0812488 -0.939 0.34759

Palpitations -0.1092509 0.1320482 -0.827 0.40804

Orthopnea 0.3427347 0.1247521 2.747 0.00601 \*\*

Chest pain 0.1169276 0.1382011 0.846 0.39751

Nausea -0.0031744 0.1453773 -0.022 0.98258

Cough 0.3325670 0.1214992 2.737 0.00620 \*\*

Fatigue -0.1433455 0.1483275 -0.966 0.33384

Dyspnea -0.1583936 0.1426783 -1.110 0.26694

Edema 0.3465353 0.1343607 2.579 0.00990 \*\*

PND -0.1205484 0.1204278 -1.001 0.31683

Tight shoes -0.1305805 0.1406626 -0.928 0.35324

Weight gain -0.1396419 0.1212165 -1.152 0.24932

DOE 0.2725216 0.1310016 2.080 0.03750 \*

## **Findings/Discussion**

The table above represents multiple logistic models built by assuming that a person gets treatment on or less than the cohort average delay days of 1 or takes more than the average number of delay days to go for treatment. It is crystal clear that the variables orthopnea, cough, edema and DOE are statistically significant in predicting the outcome since their p values (0.00601, 0.00620, 0.00990 and 0.03750 respectively) are less than 0.05. On the other hand, the variables weight gain, tight shoes, PND, dyspnea, fatigue, nausea, chest pain, Age, Gender, Palpitations, Ethnicity, Marital, living with and Education have p values which are bigger than 0,05. This shows that they have no effect in predicting the number of days before a person seeks medical treatment.

## **Comparison**

In the first case (a participant takes 2 days or less to get treatment or takes longer than 2 days to go for treatment), only one variable (cough) is useful in predicting the dependent outcome while in the second case (a person goes for treatment on or less than the cohort average delay days “1” or takes longer than the cohort average number of days to get treatment.), four predictor variables (Orthopnea, Cough edema and DOE) are useful in predicting the outcome.