GLM Project: Multinomial and Binary Models of Asteroid and Orbit Data

Kyeong A Yang, Taeho Yoon, Dongeun Lee

June 11, 2023

Contents

1 Introduction

소행성은 진화론적 관점에서 행성으로 이루어지지 못하고 남겨진 행성 잔여물을 일컫는다. 이는 혜성이랑 구분되어지는데, 꼬리가 있는 혜성과 달리 소행성은 꼬리가 없는게 특징이다. 지구가 있는 태양계 안에서는 태양계가 형성 될때 행성이 되지 못하고 남은 소행성들이 떠다니고 충돌하여 오늘날까지 이르게 되었다. 태양계내 대부분의 소행성은 화성과 목성 사이에 위치해 있으며 태양을 중심으로 공전한다.

지구 근처에도 수많은 소행성들이 존재하고 이들의 궤도는 매우 다양하다. 흔히 알려진 지구 근처의 궤도에 위치한 대표적인 소행성군(집단)은 Apollo 소행성군, Amor 소행성군, Aten 소행성군이다. 최근까지 NASA의 Near Earth Object(이하 NEO) Program을 통해 지구의 근처에 있는 소행성들의 궤도에 대한 다양한 정보가 기록되어져 왔다. 본 연구에서는 Jet Propulsion Laboratory(California Institute of Technology) 연구소에서 수집한 지구 근처에 있는 소행성들의 궤도에 관한 다양한 데이터들을 통해 지구 근처에 떠도는 크게 3가지의 소행성군을 분류해내는 모델을 제안하고자 한다. 이와 더불어, 지구 주변에 위치한 소행성은 지구와 충돌할 잠재적인 위험이 있다는 사실에 기반하여 소행성들의 궤도에 대한 데이터들을 토대로 각 소행성들이 잠재적인 위험을 지닌 소행성인지에 대한 유무를 판별하는 모델을 함께 제안하고자 한다.

우선 지구 근처에 위치한 3가지의 소행성군군(Apollo, Amor, Aten)을 분류해내는데 사용할 모형은 Baseline category logit model이다. 변수 선택 방법을 통해서 Baseline category logit model에서 가장 좋은 성능을 낼 수 있는 최적의 변수 조합을 찾아 모델링을 진행해보고 선택된 모델로 결과를 해석한다. 그 다음 지구에 잠재적인 위험이 있는 소행성인지를 분류하는데 사용할 모델은 이항분포(Binary distribution)를 가정한 일반화선형모형(generalized linear model(이하 glm))이다. 모델의 연결함수(link function)로는 로짓(logit), 프로빗(probit)을 적용해 볼 것이고 마찬가지로 각 연결함수(link function)에 변수선택방법을 적용하여 최적의 모형이 어떤 형태인지 살펴볼 것이다. 그리고 마지막으로 이 두 모델을 비교 및 해석하는 것을 통해 보고서를 마무리 하고자 한다.

2 Dataset

본 연구에 사용하고자 하는 데이터는 NASA의 Near Earth Object Program의 Jet Propulsion Laboratory(California Institute of Technology) 연구소에서 수집된 데이터이고 총 12개의 변수를 가진 15619개의 데이터이다. 이 데이터의 변수에 대한 설명은 다음과 같다. 설명변수(Explanatory variable)는 11개이고 소행성과 그것의 궤도와 관련된 변수들이다. 11개 변수 모두 연속형 변수이며, 단위는 AU, 도, 년으로 구성되어 있다.

1. Object.Classification (Y): Orbit classification
2. Orbit.Axis..AU. (X_1): Semi-major axis of the orbit in AU (궤도의 장반축, AU)
3. Orbit.Eccentricity (X_2): Eccentricity of the orbit (궤도의 이심률)
4. Orbit.Inclination..deg. (X_3): Inclination of the orbit with respect to the eclip-

tic plane and the equinox of J2000 (J2000-Ecliptic) in degrees (황도면과 J2000의 분점에 대한 궤도 경사, 도)
5. Perihelion.Argument..deg. (X_4): Argument of perihelion (J2000-Ecliptic) in degrees (근점 편각, 도)
6. Node.Longitude..deg. (X_5): Longitude of the ascending node (J2000-Ecliptic) in degrees (승교점 경도, 도)
7. Mean.Anomoly..deg. (X_6): Mean anomoly at epoch in degrees (평균 근점 이각, 도)
8. Perihelion.Distance..AU. (X_7): Perihelion distance of the orbit in AU (궤도의 근일점 거리, AU)
9. Aphelion.Distance..AU. (X_8): Aphelion distance of the orbit in AU (AU)(궤도의 원일점 거리, AU)
10. Orbital.Period..yr. (X_9): Orbital period in Julian years (율리우스년의 공전주기, 년)
11. Minimum.Orbit.Intersection.Distance..AU. (X_{10}): Minimum orbit intersection distance in AU (the minimum distance between the osculating orbits of the NEO and the Earth) (최소 궤도 교차 거리(NEO와 지구의 진동 궤도 사이의 최소 거리), AU)
12. Asteroid.Magnitude (X_{11}): Absolute V-magnitude (절대등급, V)

∗ AU(Astronomical Unit): 천문단위로 지구에서 태양까지 이르는 평균거리(1AU), 약 1억 5천만km.

반응변수(Response variable)인 Y는 아래 Table 1과 같다.

Category	Count
Amor Asteroid	5918
Amor Asteroid (Hazard)	99
Apollo Asteroid	6940
Apollo Asteroid (Hazard)	1520
Aten Asteroid	987
Aten Asteroid (Hazard)	155

Table 1: Object.Classification Counts

Table 1에 나타난 반응변수(Reponse variable)를 살펴보면 크게 3개의 소행성군(Amor, Apollo, Aten)으로 분류되어져 있고, Apollo 소행성의 개수가 가장 많고 Aten 소행성이 가장 적은 것을 확인할 수 있다. 세부적으로는 크게 분류되어진 3개의 소행성군이 잠재적인 위험이 있는 소행성군인지 아닌지(Hazard or not)에 따라 한번 더 분류되어 있다.

아래의 Table 2는 11개의 설명변수(explanatory variable)들의 기초통계량(평균, 분산, 표준편차, 최솟값, 최댓값)들을 정리해 놓은 표이다. Perihelion.Argument..deg.(X_4), Node.Longitude..deg.(X_5), Mean.Anomoly..deg.(X_6)은 평균과 분산(표준편차)가 다른 변수보다 눈에 띄게 크고 최댓값과 최솟값의 차이가 큰 것으로 보아 단위도 다르고 값의 범위도 넓은 변수라고 생각된다. 이와 같이 변수끼리의

Variable	mean	variance	standard deviation	min	max
Orbit.Axis..AU.	1.7821	0.3630395	8.535	0.5798	21.3954
Orbit.Eccentricity	0.4498	0.03097964	0.1760103	0.0044	0.9695
Orbit.Inclination..deg.	12.9349	127.7286	11.30171	0.0147	154.3751
Perihelion.Argument..deg.	181.4619	10786.54	103.8583	0.0081	359.9942
Node.Longitude..deg.	172.681	10684.74	103.367	0.007	359.998
Mean.Anomoly..deg.	172.8232	13490.38	116.1481	0.0031	359.9982
Perihelion.Distance..AU.	0.9154	0.05753963	0.2398742	0.0707	1.3000
Aphelion.Distance..AU.	2.649	1.304093	1.141969	0.990	41.540
Orbital.Period..yr.	2.475	2.30184	1.517182	0.440	98.970
Minimum.Orbit.Intersection.Distance..AU.	0.1019	0.0112304	0.1059736	0.0000	0.7069
Asteroid.Magnitude	22.29	9.093577	3.015556	9.45	33.20

Table 2: Basic Description of the Explanatory Variables

단위와 범위가 달라서 생기는 문제를 줄이기 위해 모든 설명변수에 표준화를 하여 분석을 진행한다.

이 데이터에 대한 결측치는 Astroid.Magnitude(X_{11})변수에서 1개의 자료만 NA값으로 missing 되었는데, 총 15619개의 데이터 중 1개라 크게 영향을 미칠것 같지 않아 이에 대한 데이터는 삭제하기로 한다. 그러므로 최종적으로 15618개의 데이터, 11개의 설명변수($X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}$), 그리고 Object.Classification(Y)를 반응변수로 사용한다.

3 Model

3.1 Multinomial model

(Part 1)

먼저 11개의 설명변수를 토대로 어떤 소행성군(1 = Amor, 2 = Apollo, 3 = Aten)에 속하는지 판별하는 모형을 만들기 위해 반응변수를 Object.Classification(Y)으로 설정하고 나머지 11개의 설명변수에 baseline category logit model을 적합한다. 적합된 결과는 아래의 Table 3와 같다.

→ vglm(formula = y ∼ ., family = multinomial, data = data)

- Null deviance: 27825.9 on 31234 degrees of freedom
- Residual deviance: 1509.053 on 31212 degrees of freedom
- AIC: 1557.053

결과를 살펴보면 Residual deviance는 1509.053, AIC는 1557.053으로 나왔다. 또한 Orbit.Eccentricity(X_2), Perihelion.Distance..AU.(X_7), Orbital.Period..yr.(X_9) 총 3개의 변수가 유의수준 0.05하에서 유의확률 작게 나와 유의한 변수라 할 수 있다.

(Part 2)

Part 1에서는 11개의 설명변수를 모두 다 포함시켜 완전모형으로 적합시켰다. 이번에는 R 'VGAM' 패키지 안에 있는'step4vglm' 함수를 사용하여 변수 선택을 진행한다. 이 패키지는 AIC 기준으로 후진선택법으로 변수가 선택된다. 'step4vglm' 함수를 사용하여 변수 선택한 결과는 앞선 baseline category logit model에서 완전모형을 적합시켜서 유의하게 나온 변수(X_2, X_7, X_9)와 똑같은 변수가 선택되었다. 그래서 선택된 3개의 변수인 Orbit.Eccentricity(X_2), Perihelion.Distance..AU.(X_7), Orbital.Period..yr.(X_9)를 설명변수로 설정하여 다시 Baseline category logit model을 적합한다. 적합된 결과는 아래 Table 4와 같다.

→ step4glm()

- Residual deviance: 1507.743 on 31228 degrees of freedom

- AIC: 1523.743

Variable	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept):1	8.23528	0.96494	8.535	$< 2 \times 10^{-16}$ ***
(Intercept):2	20.88443	0.85346	24.470	$< 2 \times 10^{-16}$ ***
Orbit.Axis..AU.:1	-8.95805	40.66484	-0.220	0.825646
Orbit.Axis..AU.:2	1.14257	32.51003	0.035	0.971964
Orbit.Eccentricity:1	9.52533	0.77193	12.340	$< 2 \times 10^{-16}$ ***
Orbit.Eccentricity:2	9.49210	0.72706	13.055	$< 2 \times 10^{-16}$ ***
Orbit.Inclination..deg.:1	-0.03755	0.13411	-0.280	0.779477
Orbit.Inclination..deg.:2	0.06535	0.09966	0.656	0.512025
Perihelion.Argument..deg.:1	-0.07592	0.09675	-0.785	0.432603
Perihelion.Argument..deg.:2	-0.04854	0.07870	-0.617	0.537367
Node.Longitude..deg.:1	-0.09426	0.09719	-0.970	0.332149
Node.Longitude..deg.:2	-0.05447	0.07816	-0.697	0.485866
Mean.Anomoly..deg.:1	-0.03181	0.10297	-0.309	0.757352
Mean.Anomoly..deg.:2	0.01685	0.08657	0.195	0.845664
Perihelion.Distance..AU.:1	45.49870	8.22153	5.534	3.13×10^{-8} ***
Perihelion.Distance..AU.:2	13.26337	6.53989	2.028	0.042553 ***
Aphelion.Distance..AU.:1	11.65259	38.45861	0.303	0.761897
Aphelion.Distance..AU.:2	1.61679	30.72823	0.053	0.958038
Orbital.Period..yr.:1	-3.12416	0.86793	-3.600	0.000319 ***
Orbital.Period..yr.:2	-2.54354	0.91159	-2.790	0.005267 ***
Minimum.Orbit.Intersection.Distance..AU.:1	0.22766	0.20248	1.124	0.260863
Minimum.Orbit.Intersection.Distance..AU.:2	-0.05436	0.13719	-0.396	0.691926
Asteroid.Magnitude:1	-0.04341	0.13653	-0.318	0.750523
Asteroid.Magnitude:2	0.01130	0.11027	0.102	0.918402

Table 3: Coefficient Estimates (Part 1)

최종적으로 적합된 결과를 보면 Orbit.Eccentricity(X_2)와 Perihelion.Distance..AU(X_7)가 유의한 변수로 나왔다. 또한 residual deviance는 1507.743이 나왔고 AIC는 1523.743이 나왔다. Part 1 모델에서는 residual deviance가 1509.053이 나왔고 AIC는 1557.053이 나왔다. 이를 통해 변수 선택을 한 후의 선택된 변수로 모형을 적합시켰을 때의 성능이 residual deviance와 AIC 기준에서 Part 1보다 개선됨을 확인할 수 있다. Part 2 모델을 데이터에 적합시켰을 때 최적의 모형이라 결론 내리고 이를 다음과 같이 해석해보고자 한다.

Variable	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
(Intercept):1	7.2878	0.9398	7.755	8.83×10^{-15} ***
(Intercept):2	20.5020	0.8270	24.792	$< 2 \times 10^{-16}$***
Orbit.Eccentricity:1	10.3833	0.5143	20.188	$< 2 \times 10^{-16}$ ***
Orbit.Eccentricity:2	10.2610	0.4721	21.733	$< 2 \times 10^{-16}$ ***
Perihelion.Distance..AU.:1	45.1189	1.2302	36.677	$< 2 \times 10^{-16}$ ***
Perihelion.Distance..AU.:2	13.9332	0.6289	22.154	$< 2 \times 10^{-16}$ ***
Orbital.Period..yr.:1	-0.8533	0.5862	-1.456	0.145
Orbital.Period..yr.:2	-0.7863	0.5173	-1.520	0.129

Table 4: Summary of a Variable Selection (Part2)

1)
$log(\hat{\pi_1}/\hat{\pi_3}) = 7.288 + 10.383 \times Orbit.Eccentricity(X_2) + 45.119 \times Perihelion.Distance..AU.(X_7) - 0.853 \times Orbital.Period..yr(X_9)$

- X_2가 1단위 증가할때 소행성군이 Aten보다 Amor일 오즈는 exp(10.383)배와 같다.
- X_7가 1단위 증가할 때 소행성군이 Aten보다 Amor일 오즈는 exp(45.119)배와 같다.
- X_9가 1단위 증가할 때 소행성군이 Aten보다 Amor일 오즈는 exp(0.835)배와 같다.

2)
$log(\hat{\pi_2}/\hat{\pi_3}) = 20.502 + 10.261 \times Orbit.Eccentricity(X_2) + 13.933 \times Perihelion.Distance..AU.(X_7) - 0.786 \times Orbital.Period..yr(X_9)$

- X_2가 1단위 증가할때 소행성군이 Aten보다 Amor일 오즈는 약 exp(10.261)배와 같다.
- X_7가 1단위 증가할 때 소행성군이 Aten보다 Amor일 오즈는 exp(13.933)배와 같다.
- X_9가 1단위 증가할 때 소행성군이 Aten보다 Amor일 오즈는 exp(0.786)배와 같다.

3)
$log(\hat{\pi_1}/\hat{\pi_2}) = -13.214 + 0.122 \times Orbit.Eccentricity(X_2) + 31.186 \times Perihelion.Distance..AU.(X_7) - 0.067 \times Orbital.Period..y(X_9)$

- X_2가 1단위 증가할때 소행성군이 Apollo보다 Amor 일 오즈는 exp(0.122)배와 같다.
- X_7가 1단위 증가할 때 소행성군이 Apollo보다 Amor일 오즈는 exp(31.186)배와 같다.
- X_9가 1단위 증가할 때 소행성군이 Apollo보다 Amor일 오즈는 exp(0.067)배와 같다.

3.2 Binary model

3.2.1 Logit link

(Part 1)

다음은 지구 근처 있는 3가지의 소행성군들 중에 지구에 충돌할 잠재적인 위험이 있는지 없는지(0 = 없음, 1 = 있음) 판별해내는 모델을 만들기 위해 이항분포(binary distribution)을 가정한 일반화선형모형(glm)을 추정하고자 한다. 따라서 모델을 추정하기 위해 잠재적인 위험이 있으면 1, 없으면 0을 나타내는 새로운 Y 변수를 만들었다.

이번에는 이 Y변수를 반응변수로 가정하고 11개의 모든 변수를 설명변수에 포함하여 연결함수(link function)를 로짓(logit)으로 설정하여 이항모형(binary model)을 추정하였다. 결과는 아래의 Table 5와 같다.

→ glm(formula = y ∼ ., family = binomial(link = "logit"), data = data)

- Null deviance: 11056.0 on 15617 degrees of freedom
- Residual deviance: 3191.7 on 15606 degrees of freedom
- AIC: 3215.7

Variable	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-9.83244	0.26230	-37.485	$< 2 \times 10^{-16}$ ***
Orbit.Axis..AU.	-22.30170	19.15396	-1.164	0.244
Orbit.Eccentricity	0.06553	0.16460	0.398	0.691
Orbit.Inclination..deg.	0.01783	0.05568	0.320	0.749
Perihelion.Argument..deg.	-0.03901	0.04770	-0.818	0.413
Node.Longitude..deg.	-0.03274	0.04599	-0.712	0.477
Mean.Anomoly..deg.	-0.01507	0.04752	-0.317	0.751
Perihelion.Distance..AU.	4.40524	3.81333	1.155	0.248
Aphelion.Distance..AU.	20.94892	18.15457	1.154	0.249
Orbital.Period..yr.	0.03445	0.23065	0.149	0.881
Minimum.Orbit.Intersection.Distance..AU.	-11.46858	0.33985	-33.746	$< 2 \times 10^{-16}$ ***
Asteroid.Magnitude	-4.31854	0.11413	-37.840	$< 2 \times 10^{-16}$ ***

Table 5: Coefficient Estimates for Logit Link (Part1)

결과를 보면 Minimum.Orbit.Intersection.Distance(X_{10}), Asteroid.Magnitude (X_{11}) 두 변수가 유의하게 나왔다. Residual deviance는 3197.7, AIC는 3215.7임을 확인할 수 있다.

(Part 2)

Binary model을 다른 변수 조합으로 추정하기 위하여 R 내장 함수인 step function을 사용하였다. 이 함수는 AIC를 기준으로 변수가 선택된다. 전진 선택법, 후진제거법, 그리고 단계적 선택법을 이용하여 변수선택을 진행하였을때, 모두 결과가 같았다. 따라서 후진제거 변수선택 방법을 이용했을때 분석 과정의 처음과 마지막은 아래의 Table 6과 같다.

Start: AIC=3215.72
$y \sim .$

Variable	Df	Deviance	AIC
-Orbital.Period..yr.	1	3191.7	3213.7
-Mean.Anomoly..deg.	1	3191.8	3213.8
-Orbit.Inclination..deg.	1	3191.8	3213.8
-Orbit.Eccentricity	1	3191.9	3213.9
-Node.Longitude..deg.	1	3192.2	3214.2
-Perihelion.Argument..deg.	1	3192.4	3214.4
-Aphelion.Distance..AU.	1	3193.0	3215.0
-Perihelion.Distance..AU.	1	3193.1	3215.1
-Orbit.Axis..AU.	1	3193.1	3215.1
-Asteroid.Magnitude	1	7976.0	7998.0
-Minimum.Orbit.Intersection.Distance..AU.	1	8683.9	8705.9
(none)		3191.7	3215.7

Start: AIC=3203.37
$y \sim Orbit.Axis..AU. + Minimum.Orbit.Intersection.Distance..AU. + Asteroid.Magnitude$

	Df	Deviance	AIC
(none)	3195.4	3203.4	
- Orbit.Axis..AU.	1	3201.4	3207.4
- Asteroid.Magnitude	1	8978.1	8984.1
- Minimum.Orbit.Intersection.Distance..AU.	1	9647.2	9653.2

Table 6: Comparison of Models (Part 2)

변수 선택 과정 처음 부분을 보면 Asteroid.Magnitude(X_{11})의 변수가 빠지면 deviance는 3191.7에서 7976으로, Minimum.Orbit.Intersection.Distance(X_{10})의 변수가 빠지면 8683.9로 크게 증가하는 것을 확인할 수 있다. AIC 또한 마찬가지다. Asteroid.Magnitude(X_{11})의 변수가 빠지면 AIC는 3215.7에서 7998으로, Minimum.Orbit.Intersection.Distance..AU(X_{10})의 변수가 빠지면 8705.9로 크게 증가하는 것을 확인할 수 있다.

결국 최종 선택된 변수는 총 3가지로 orbit axis, Asteroid.Magnitude(X_{11}), Minimum.Orbit.Intersection.Distance(X_{10})인 것을 확인할 수 있다. 여기서 Orbit axix의 변수가 모델에서 빠진다고 가정했을 때 나머지 두 변수에 비해서 deviance와 AIC 기준으로는 변화량이 상대적으로 매우 적은 것으로 보인다. 하지만 선택되지 않은 다른 9개의 변수들에 비해서 AIC의 변화량이 상대적으로 크기 때문에 유의한 변수로 선택되었다고 할 수 있다.

3.2.2 Probit link

(Part 1)

3.2.2절은 Y가 이항분포를 따를 때 연결함수(link function)를 로짓(logit)이 아닌 프로빗(probit)으로 선택하여 11개의 변수 모두 포함한 완전모형을 적합한다. 적합한 결과는 아래의 Table 7과 같다.

→ glm(formula = y ∼ ., family = binomial(link = "probit"), data = data)

- Null deviance: 11056 on 15617 degrees of freedom
- Residual deviance: 3183 on 15606 degrees of freedom
- AIC: 3207

결과를 보면 Residual deviance는 3183, AIC는 3207이 나왔으며 Minimum.Orbit.Intersection.Distance..AU.(X_{10})와 Asteroid.Magnitude(X_{11}) 총 2개의 변수가 유의수준 0.05하에서 유의확률이 0.05보다 작게 나와 유의한 변수라는 것을 Table 7에서 확인할 수 있다.

(Part 2)

이제 3.2.1절과 마찬가지로 R의 step function을 활용하여 AIC 기준으로 전진선택법, 후진제거법, 그리고 단계적 선택법을 이용하여 변수선택을 진행하였다. 모두 결과가 같은 이유로 후진제거법의 결과를 정리하였고, 그 결과는 아래의 Table 8과 같다.

변수 선택 과정 처음 부분을 보면 Asteroid.Magnitude(X_{11})의 변수가 빠지면 deviance는 3183에서 7935.2으로, Minimum.Orbit.Intersection.Distance..AU.(X_{10})의 변수가 빠지면 8553.6으로 크게 증가하는 것을 확인할 수 있다. 앞서 본 logit link와 마찬가지로 AIC 또한 같은 양상을 보였다. Asteroid.Magnitude(X_{11})의 변수가 빠지면 AIC는 3207에서 7957.2으로, Minimum.Orbit.Intersection.Distance..AU.(X_{10})의 변수가 빠지면 8575.6로 크게 증가하는 것을 확인할 수 있다. 결국 최종 선택된 변수는 총 3가지로 Orbit Axis..AU.(X_1), Minimum.Orbit.Intersection.Distance..AU.(X_{10}), Asteroid.Magnitude(X_{11})인 것을 확인할 수 있다. Logit link와 마찬가지로 변수선택의 마지막 단계에서 앞의 두 변수와 Orbit axix(X_1)로 구성된 모델에서 Orbit axix(X_1)가 빠진다고 가정했을 때 나머지 두 변수에 비해서 deviance와 AIC 변화량이 상대적으로 매우 적은 것으로 보인다. 하지만 선택되지 않은 다른 9개의 변수들에 비해서 AIC의 변화량이 상대적으로 크기 때문에 최종적인 모델에서는 포함된 것으로 생각할 수 있다. 즉, 로닛(logit)과 프로빗(probit) 모형 모두 최종적인 모델로 같은 변수를 선택되었다.

Variable	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-5.396758	0.132256	-40.805	$< 2 \times 10^{-16}$ ***
Orbit.Axis..AU.	-11.788082	10.420114	-1.131	0.258
Orbit.Eccentricity	0.021434	0.090241	0.238	0.812
Orbit.Inclination..deg.	0.012206	0.030365	0.402	0.688
Perihelion.Argument..deg.	-0.021407	0.025962	-0.825	0.410
Node.Longitude..deg.	-0.024091	0.025080	-0.961	0.337
Mean.Anomoly..deg.	-0.001175	0.025773	-0.046	0.964
Perihelion.Distance..AU.	2.312874	2.074880	1.115	0.265
Aphelion.Distance..AU.	11.091009	9.875502	1.123	0.261
Orbital.Period..yr.	0.014026	0.125911	0.111	0.911
Minimum.Orbit.Intersection.Distance..AU.	-6.278959	0.173807	-36.126	$< 2 \times 10^{-16}$ ***
Asteroid.Magnitude	-2.327851	0.056302	-41.346	$< 2 \times 10^{-16}$ ***

Table 7: Coefficient Estimates for Probit Link (Part 1)

Start: AIC=3206.95
$y \sim .$

Variable	Df	Deviance	AIC
- Mean.Anomoly..deg.	1	3183.0	3205.0
- Orbital.Period..yr.	1	3183.0	3205.0
- Orbit.Eccentricity	1	3183.0	3205.0
- Orbit.Inclination..deg.	1	3183.1	3205.1
- Perihelion.Argument..deg.	1	3183.6	3205.6
- Node.Longitude..deg.	1	3183.9	3205.9
- Perihelion.Distance..AU.	1	3184.2	3206.2
- Aphelion.Distance..AU.	1	3184.2	3206.2
- Orbit.Axis..AU.	1	3184.2	3206.2
- Asteroid.Magnitude	1	7935.2	7957.2
- Minimum.Orbit.Intersection.Distance..AU.	1	8553.6	8575.6
(none)		3183.0	3207.0

Start: AIC=3195.3
y ~ Orbit.Axis..AU. + Minimum.Orbit.Intersection.Distance..AU. + Asteroid.Magnitude

	Df	Deviance	AIC
(none)	3195.4	3203.4	
- Orbit.Axis..AU.	1	3201.4	3207.4
- Asteroid.Magnitude	1	8978.1	8984.1
- Minimum.Orbit.Intersection.Distance..AU.	1	9647.2	9653.2

Table 8: Comparison of Models (Part 2)

3.2.3 Model comparison(Logit vs Probit)

Logit Model

Variable	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-9.8160	0.2609	-37.621	$< 2 \times 10^{-16}$ ***
Orbit.Axis..AU.	-0.1195	0.0481	-2.484	0.013 *
Asteroid.Magnitude	-4.3307	0.1113	-38.896	$< 2 \times 10^{-16}$ ***
Minimum.Orbit.Intersection.Distance..AU.	-11.4684	0.3340	-34.340	$< 2 \times 10^{-16}$ ***

- Null deviance: 11056.0 on 15617 degrees of freedom
- Residual deviance: 3191.7 on 15606 degrees of freedom
- AIC: 3203.4

연결함수(Link function)를 로짓(logit)으로 설정하고 변수선택의 결과로 선택된 3가지 변수를 설명변수로 설정하여 모형을 적합시킨 결과, residual deviance는 3191.7, AIC는 3203.4가 나온 것을 확인할 수 있다. 또한, 3개의 변수 모두 유의확률이 낮게 나와 유의수준 0.05하에 유의한 변수임을 확인할 수 있다.

Probit Model

Variable	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-5.38568	0.13154	-40.944	$< 2 \times 10^{-16}$ ***
Orbit.Axis..AU.	-0.06237	0.02587	-2.411	0.0159 *
Asteroid.Magnitude	-2.33622	0.05471	-42.702	$< 2 \times 10^{-16}$ ***
Minimum.Orbit.Intersection.Distance..AU.	-6.28228	0.17090	-36.759	$< 2 \times 10^{-16}$ ***

- Null deviance: 11056.0 on 15617 degrees of freedom
- Residual deviance: 3187.3 on 15614 degrees of freedom
- AIC: 3195.3

연결함수(Link function)를 프로빗(probit)으로 설정하고 변수선택의 결과로 선택된 3가지 변수를 설명변수로 설정하여 모형을 적합시킨 결과, residual deviance는 3187.3, AIC는 3195.3이 나온 것을 확인할 수 있다. 또한, 3개의 변수 모두 유의확률이 낮게 나와 유의수준 0.05하에 유의한 변수임을 확인할 수 있다.

두 모델을 비교해보면 연결함수(link function)를 프로빗(probit)으로 설정하였을 때의 residual deviance와 AIC가 더 낮은 값을 가진다. 그러므로 연결함수(link function)를 프로빗(probit)으로 설정하고 Orbit.Axis..AU(X_1), Asteroid.Magnitude(X_{11}), Minimum.Orbit.Intersection.Distance..AU(X_{10}) 총 3개의 변수를 설명변수로 모형 적합시켰을 때 이항모형(binary model)의 deviance와 AIC 측면에서 가장 성능이 좋다고 할 수 있다.

최종 선택된 binary model의 결과를 해석하면 다음과 같다.
잠재변수 $Y_i^* = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0,1)$ 를 가정한다.

- Orbit.Axis..AU(X_1)가 한 단위 증가할 때마다 $E(Y^*)$는 0.06237만큼 감소한다.

- Asteroid.Magnitude(X_{11})가 한 단위 증가할 때마다 $E(Y^*)$는 2.33622만큼 감소한다.

- Minimum.Orbit.Intersection.Distance..AU(X_{10})가 한 단위 증가할 때마다 $E(Y^*)$는 6.28228만큼 감소한다.

4 Conclusion

태양계에는 수 많은 소행성들이 존재하고 지구 근처에 위치한 대표적인 소행성군으로는 Amor, Apollo, Aten 소행성군이 있다. 본 분석에서는 NASA의 Near Earth Object Program의 Jet Propulsion Laboratory(California Institute of Technology)에서 수집한 데이터를 통해 지구 근처에 위치한 소행성군에 대해 분석해보고자하였다.

이 데이터에서 주어진 소행성들의 궤도 정보를 통해 Amor, Apollo, Aten 소행성군을 분류해내는 Baseline category logit model(1 = Amor, 2 = Apollo, 3 = Aten)을 11개의 모든 설명변수(explanatory variable)를 모두 포함하여 적합시켰고 이후 R 패키지의 'step4vglm' 패키지를 이용하여 AIC 기준으로 변수선택한 모형과 비교해보았다. 모델의 성능은 deviance와 AIC 측면에서 보았을 때 변수선택을 하여 선택된 변수로 적합시킨 모형의 성능이 더 좋았다. 이후 최종 선택된 모형을 해석을 해보았고, 더 나아가 R 패키지의 'vglm' 함수는 baseline category가 Aten일때의 결과를 보여주기 때문에 baseline category가 Aten이 아닌 Amor와 Apollo인 경우에 대해서도 모형을 추가적으로 해석해보았다.

그 다음으로 지구 근처에 있는 소행성들이 지구와 잠재적인 충돌 위험이 있는지 없는지(0 = 없다, 1 = 있다) 판별해내는 모형을 만들기 위해 이항분포(Binary distribution)을 따르는 일반화선형모형(glm)을 적합시켰다. 마찬가지로 Binary 모모형에 11개의 모든 설명변수(explanatory variable)를 포함하여 완전모형으로 적합시켰고 R 패키지의 'step' 함수를 이용하여 AIC 기준으로 변수선택한 후 선택된 변수들로 모형을 적합시킨 후 두 모형을 비교하였다. 이때 일반화선형모형(glm)의 연결함수(link function)로 로짓(logit)을 사용했을 때와 프로빗(probit)을 사용했을 때를 직접 비교하였다. 모델의 성능은 deviance와 AIC 측면에서 프로빗 연결함수(probit link)를 사용하여 AIC 기준으로 변수선택한 모형의 성능이 가장 좋았다. 최종 선택된 모델을 통해 설명변수(explanatory variable)가 변할 때 잠재변수 $E(y*)$가 어떻게 변하는지 해석하는 것으로 분석을 마무리하였다.

Multinomial logit model과 두 binary model이 결과적으로 유의하다고 본 변수는 다른데 multinomial model은 최종적으로 Orbit.Eccentricity(X_2), Perihelion. Disance..AU.(X_7), Orbital.Period..yr.(X_9)를 Object.Classification(Y)에 유의한 변수로 보았고 두 binary model은 Orbit.Axis..AU.(X_1), Minimum.Orbit.Intersection.Distance..AU.(X_{10}), Asteroid.Magnitude(X_{11}) 를 유의한 변수로 보았다는 점에서 차이가 있음을 추가적으로 확인할 수 있었다.