# http://www.dodomira.com/

# 데이터 분석의 유형 6가지 – 목적에 따라 달라지는 분석 방법



데이터 분석의 첫 단계는 “문제 정의”이다.

데이터 분석을 시작하기 전 단계에서 분석을 통해 알고 싶은 것이 “무엇”인지를 구체적으로 명확히 정의하지 않으면 “어떻게” 분석해야 할 지도 알기 어렵다.  
문제 정의 단계에서의 목표를 구체적으로 설정하는 것의 중요성은 많은 데이터 분석가들이 주지하고 있지만, 또 하나 이 단계에서 중요하게 생각해야 할 Task가 있다.  
바로 “문제 유형 알기”가 그것이다. 내가 해결하고자 하는 문제가 어떤 유형의 문제인지를 알고 있어야 어떤 분석 방법을 사용할 것인지, 어떤 방법은 적당하지 않은지를 알 수 있게 된다. 이 포스팅에서는 데이터 분석 유형에 대해 정리하고, 해결해야 할 유형을 오해하고 있을 때 발생하는 문제를 알아보고자 한다.

**1. Descriptive analysis**

가장 기본이 되는 분석으로, 주어진 데이터를 요약/집계하여 결과를 도출하는 것이 Descriptive Analysis의 목표가 된다.  
이번 달 매출액, 평균 세션 타임, 설문 응답자의 남/녀 비중 등이 이 유형의 분석 결과이다.  
이러한 지표들은 Business 관점에서 보면 Bottom line이고, 특히 경영자, 주주 및 여러 이해관계자들이 가장 눈여겨 보는 지표이기 때문에 회사 차원에서 가장 중요한 수치이지만,  
과거의 데이터를 단순 계산/집계하여 얻어진 Fact이기 때문에, 분석 결과를 따로 해석하는 과정(So what?, Can we apply this to other population?)을 거치지 않는다. (비즈니스가 아니라 분석 관점에서)

이 단계에서 주로 사용되는 그래프는 Pie chart, Box plot, Bar plot 등이 있으며, 요약 형식의 테이블도 주로 이러한 유형의 분석의 결과물 이다.

**2. Exploratory analysis**

EDA라고도 불리우는 Exploratory Analysis는 최근 Visualization이 데이터 분석의 화두로 떠오르면서 주목을 받게 된 분석 유형이다.  
Exploratory Analysis의 주요 목표는 여러 변수 간 트렌드나 패턴, 관계를 찾는 것인데,  
이 분석 역시 통계적 기법을 사용한 모델링이라기 보다는 그래프(주로 Plot)를 통한 사실 확인이 주된 작업이다.

주로 데이터 분석 프로젝트 초기에 가설을 수립하기 위해 수행되기도 한다.  
새로운 데이터를 접할 때 마다 여러 변수들 간의 관계를 이리 저리 그래프로 그리다 보면 흥미로운 사실들이 발견되면서 점점 빨려들어(?)가는데,  
아래 두가지를 염두하고 있으면 이 단계를 효율적으로 진행할 수 있을 것이다.

A. 분석의 목적을 항상 염두에 두자.

데이터 탐색은 기본적으로 끝이 없는 작업이다. 현실에서 분석하게 되는 데이터들은 다양한 변수들을 가지고 있는데,  
각 변수들의 조합, 범주형 변수가 있을 경우 범주 내에서의 조합, 범주형 변수 끼리의 조합 등등 목적없이 파게 되면 끝없이 팔 수 있는 것이 탐색적 데이터 분석이기 때문에, 분석의 목적이 가설 수립인지, 특정 변수들 간의 관계 파악인지, 트렌드 파악인지를 항상 머릿속에 가지고 분석을 진행하는 것을 추천한다.

B. Reproduciblity

변수의 갯수가 많을 경우 데이터 탐색을 위해 그리게 되는 Plot도 늘어나게 된다. 흥미로운 Plot을 발견한 경우 해당 그래프를 다시 생성할 수 있도록 어떤 조건에서 그러한 Plot을 그리게 됬는지를 확인하자.

이 단계에서 주로 사용되는 그래프는 Scatter plot, Bobble chart 시계열 차트 등이 있다.

**3. Inferential analysis**  
샘플 – 모집단 간의 관계를 탐구하는 것이 이 분석 유형의 목적이다.  
샘플에서 얻어낸 정보가 모집단에도 적용될 수 있는지를 검토하는 것인데, 이는 자원이 한정되어 있어 샘플 크기를 무한정 늘릴 수 없기 때문에 발생하는 일반적인 문제를 해결하기 위한 분석 유형이다.  
“95%의 확률로 모집단의 평균 점수는 80~85점 사이라고 말할 수 있다” , “그룹 A의 성적이 그룹 B보다 유의미하게 높았다”는 식의 결론이 Inferential analysis의 결과 이다.

**4. Predictive analysis**

머신 러닝(Machine learning), 의사결정나무(Decision Tree) 등 다양한 통계적 기법을 사용하여 미래 혹은 발생하지 않은 어떤 사건에 대한 예측을 하는 것이 Predictive analysis의 주요 목표이다.  
“왜” 보다는 “정확한 예측”을 하는 것이 주요 관심사이기 때문에, 특정한 설명 변수가 어떤 매커니즘으로 목표 변수에 영향을 미치는 가에 대한 설명력은 약하다.

예를 들어 금연 성공률과 여러가지 설명 변수들 간의 관계를 분석한다고 가정 하자.  
Causal analysis의 목표가 금연 성공률에 영향을 끼치는 변수는 금연 방법 (그냥 끊기, 약물치료, 상담치료, 약물/상담 병행) 및 흡연 기간이다라는 “관계”를 밝혀내는 것이라면,  
Predictive analysis의 목표는 직관적으로 금연과 관계 있는 요소 뿐 만 아니라 나이, 직업, 키, 몸무게, 거주 지역 등 사용가능한 변수들을 모두 활용하여 모델의 “예측력”을 높이는 것이다.

최근 Predictive analysis에서 좀 더 나아간 Prescriptive analysis도 주목 받고 있는데,  
이 유형의 분석은 예측 모델을 기반으로 어떤 의사결정을 해야 하는지를 “처방”하는 것이 주요 목적이다.  
예측 모델 속에 여러 의사 결정 옵션을 변수로 포함시켰다고 보면 이해가 쉬울 것 같다.  
Prescriptive model은 어떤 의사 결정을 하느냐에 따라 변하는 결과를 예측해 주기 때문에 이러한 분석을 통해 최적화된 의사 결정을 할 수 있다.

2000년대 초중반 까지 유행했던 Scenario analysis와 유사한 점도 있지만,  
Scenario analysis가 비즈니스 인싸이트에 기반하여 만든 정성적인 시나리오를 기반으로 미래를 예측하고 행동 옵션을 제시했다면,  
최근의 Prescriptive analytics는 철저하게 수학 모델에 기반하고 있다는 점, 행동 옵션도 모델에 포함시켜 미래를 예측 한다는 점이 차이라고 볼 수 있다.

**5. Causal analysis**

독립 변수와 종속 변수 간의 인과관계가 있는지 여부를 확인하기 위한 분석이다.  
Causal analysis의 특징 중 하나는 실험을 통해 수집된 데이터를 대상으로 한다는 점이다.  
주로 독립 변수를 실험대상에 랜덤하게 할당한 후, 그룹 간 실험 전/후의 종속 변수의 변화를 관찰하여 실험 데이터 수집이 이루어 진다.  
과거의 단순 관찰 데이터(Observational data)로는  Confounding variable이 있을 가능성을 배제할 수가 없으므로, 인과 관계 여부를 명확하게 밝혀 낼 수가 없기 때문이다.

최근 자주 언급되는 A/B test도 이러한 데이터 수집 방법의 하나라고 볼 수 있다.

선형Regression이 가장 많이 사용되는 분석 방법이며, 변수가 여러개일 경우 [Multi variable regression](http://www.dodomira.com/2016/01/31/r%EC%9D%84-%EC%82%AC%EC%9A%A9%ED%95%9C-%EB%8B%A4%EC%A4%91%ED%9A%8C%EA%B7%80%EB%B6%84%EC%84%9D-simple-multiple-regression-r/), 변수가 범주형일 경우 [Logistics regression](http://www.dodomira.com/2016/02/12/%EB%A1%9C%EC%A7%80%EC%8A%A4%ED%8B%B1-%ED%9A%8C%EA%B7%80%EB%B6%84%EC%84%9D-r/)을 사용한다.

**6. Mechanistic analysis**

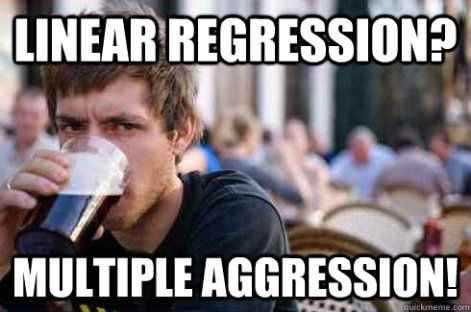
독립 변수가 어떤 매커니즘으로 종속 변수에 영향을 미치는 지를 분석하는 것으로, 위 6가지 유형 중 가장 노력이 많이 필요한 분석 유형이다.  
Causal analysis의 목적이 독립 -종속 변수 간의 “인과 관계 여부”를 밝혀내는 것이었다면,  
Mechanistic analysis의 목적은 이에 더해 어떠한 독립 변수가 어떠한 작용을 통해 독립 변수에 그러한 영향을 미치는 지를 이해하는 것이다.  
Causal analysis와 동일하게 실험 데이터를 분석 대상으로 한다,

# 1. R을 사용한 다중회귀분석 (Multiple regression in R)

**다중회귀분석(multiple regression)이란?**

명칭에서 짐작할 수 있듯이 다중회귀분석이란 설명 변수가 두개 이상인 회귀분석을 의미한다.  어떤 요소에 영향을 미치는 원인이 하나 이상이라는 감(?)이 올때 한번 돌려보면 된다.

“다중회귀분석”이라는 이름을 보고 어려울 것 같아 지레 겁먹을 수도 있지만,  본 포스팅을 읽다보면 쉽고 빠르게 다중회귀분석을 실시하고, 결과를 해석하는 방법을 알 수 있을 것이다.



**다중회귀분석의 순서**

R에서는 lm함수를 사용하여 간단히(?) 다중회귀분석을 실시할 수 있으며, predict 함수를 쓰면 새로운 데이터(설명 변수)를 사용하여 결과 변수를 예측할 수도 있다.

본젹적으로 분석을 실시하기 전에 아래 분석 순서를 기억해 두는 것이 멘탈을 잡는 것에 매우 큰 도움이 된다.

아래는 분석하고 싶은 설명/대상 변수가 머릿속에 있는 경우를 가정한 분석 프로세스이다.

**1단계.**lm함수를 사용하여 회귀분석 모델을 생성한다.

**2단계.**모델이 통계적으로 유의한지 여부를 확인한다.

A. F-test  
F-test는 회귀분석 모델 전체에 대해 이것이 통계적으로 의미가 있는지를 결정하기 위해 사용하며, 회귀분석 모델에서 F-Statistic의 p-value의 값이 0.05보다 작은 경우 회귀식 전체는 유의하다고 볼 수 있다.

B. p-value  
변수의 p-value는 각 변수가 대상 변수에 유의하게 영향을 미치는 지를 확인하기 위해 사용한다.  역시 0.05보다 작은 경우 각 변수는 유의하게 결과 변수를 설명한다고 불 수 있다.

C. adjusted-r 제곱  
모델이 대상 변수의 몇%를 설명하는지 확인한다.

**3단계.** 필요한 경우 partial F-test를 통해 추가할 새로운 변수가 삭제할 변수는 없는지 확인한다.

**4단계.** predict함수를 사용하여 새로운 데이터 셋에 대한 예측값을 구한다.

다중회귀분석, 생각보다 간단하다.

**R을 사용한 다중회귀분석 예제 (dataset: swiss)**

필요한 패키지 및 데이터를 로드한다. 본 예제에서는 R datasets 패키지에 내장되어 있는 Swiss 데이터 셋을 사용한다.

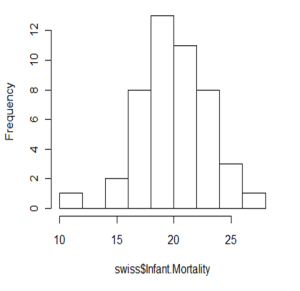
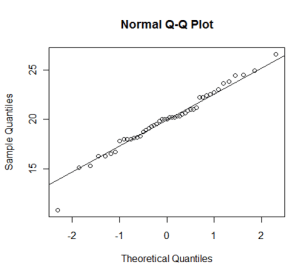
우선 아래 코드를 실행해서 데이터셋 swiss의 변수들을 확인한다.R

|  |  |
| --- | --- |
| 1  2  3  4 | require(datasets); require(ggplot2)  data(swiss)  str(swiss)  summary(swiss) |

데이터셋 swiss의 Infant.Mortality을 대상으로 분석할 것이므로 아래 코드를 통해 Infant.Mortality가 정규분포를 따르고 있는지를 확인한다.

R

|  |  |
| --- | --- |
| 1  2  3 | hist(swiss$Infant.Mortality)  qqnorm(swiss$Infant.Mortality)  qqline(swiss$Infant.Mortality) |

아래 Q-Q plot상의 직선에서 점들이 크게 벗어나 있지 않는다면 Infant.Mortality 변수는 정규 분포를 따른다고 볼 수 있다.

다음 단계는 회귀 분석을 위한 모델을 만드는 것이다.  
swiss 데이터셋의 다른 모든 변수와 Infant.Mortality 간의 관계를 살펴보자.

lm함수를 사용하는 문법은 간단하다.  
lm(대상변수~설명변수1+설명변수2, data= 사용할 데이터 셋 명)

R

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24 | model<-lm(Infant.Mortality~. ,data=swiss)  summary(model)    Call:  lm(formula = Infant.Mortality ~ ., data = swiss)    Residuals:      Min      1Q  Median      3Q     Max  -8.2512 -1.2860  0.1821  1.6914  6.0937    Coefficients:                Estimate Std. Error t value Pr(>|t|)  (Intercept)  8.667e+00  5.435e+00   1.595  0.11850  Fertility    1.510e-01  5.351e-02   2.822  0.00734 \*\*  Agriculture -1.175e-02  2.812e-02  -0.418  0.67827  Examination  3.695e-02  9.607e-02   0.385  0.70250  Education    6.099e-02  8.484e-02   0.719  0.47631  Catholic     6.711e-05  1.454e-02   0.005  0.99634  ---  Signif. codes:  0 ?\*\*?0.001 ?\*?0.01 ??0.05 ??0.1 ??1    Residual standard error: 2.683 on 41 degrees of freedom  Multiple R-squared:  0.2439, Adjusted R-squared:  0.1517  F-statistic: 2.645 on 5 and 41 DF,  p-value: 0.03665 |

lm함수를 사용해서 만든 모델에 summary함수를 적용하면 위와 같은 결과를 얻을 수 있다.

결과 아래쪽의 F-statistic 결과의 p-value를 보면 0.03665로 0.05보다 작아 이 모델은 유의하게 Infant.Fertility를 설명하는 데 사용할 수 있다고 판단 가능하다.  
두번째로는 Coefficients 항목인데,  여러 변수들의 p-value를 확인한 결과 Infant.Fertility와 유의한 관계가 있는 변수는 Fertility 하나이다. (p-value = 0.00734)

각 변수가 통계적으로 유의한 경우(=p-value가 0.05보다 작은 경우) p-value 뒤에 \* 마크가 붙으며, \*의 갯수가 늘어날 수록 통계적으로 유의할 확률이 커진다.

Fertility 변수를 제외한 다른 변수들을 모두 삭제했을 때 모델이 통계적으로 달라지는지를 확인하기 위해 Partial F-test를 해보자.  
Partial F-test에는 anova 함수를 사용한다.

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11 | model\_simple<-lm(Infant.Mortality~Fertility ,data=swiss)  anova(model, model\_simple)    Analysis of Variance Table    Model 1: Infant.Mortality ~ Fertility + Agriculture + Examination + Education +      Catholic  Model 2: Infant.Mortality ~ Fertility    Res.Df    RSS Df Sum of Sq      F Pr(>F)  1     41 295.07  2     45 322.54 -4   -27.472 0.9543 0.4427 |

Fertility와 Infant.Mortality 두개의 변수만을 사용하여 model\_simple 객체를 만든 후,  
anova 함수로 두 모델을 비교하면 된다.  
결과의 가장 아랫부분을 보면 p-value는 0.4427로 두 모델의 Infant.Mortality에 대한 설명력에는 차이가 없다고 볼 수 있다.

회귀 분석의 중요한 목적 중 하나는 새로운 변수가 주어졌을 때 결과 변수를 예측하는 것이니 만큼, predict함수를 사용하여 새로운 Fertility에 대한 Infant.Mortality를 예측해 보자.

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16 | new\_Fertility<-rnorm(10, mean=mean(swiss$Fertility), sd=sd(swiss$Fertility))  new\_Fertility<-as.data.frame(new\_Fertility)  colnames(new\_Fertility)<-c("Fertility")  predict(model\_simple, new\_Fertility, interval="prediction")            fit      lwr      upr  1  19.18853 13.71692 24.66015  2  20.10759 14.65725 25.55792  3  19.87284 14.42339 25.32230  4  18.34644 12.79772 23.89517  5  19.93479 14.48552 25.38405  6  21.87881 16.28379 27.47383  7  20.03850 14.58888 25.48813  8  19.59332 14.13926 25.04739  9  19.07952 13.60099 24.55805  10 20.05165 14.60192 25.50138 |

새로운 10개의 Fertility 변수에 대한 예측 값은 fit 컬럼에, 95% 신뢰구간은 lwr/up에서 확인하면 된다. 새로 입력하는 데이터 셋의 변수명은 모델에 사용한 변수명과 동일하게 지정해 주어야 하는 점을 기억하자.

주의 – 다중공선성 문제 (Multicollinearity)

회귀 모델의 설명 변수들 사이에 상관관계가 있는 경우가 있는데, 이것을 “다중공선성” 라고 한다.  (고객 만족도를 조사하면서 상품의 내구도와 불량률을 설명 변수로 동시에 포함시키는 경우 등이다.)

다중공선성이 존재할 경우 모델의 정확도가 하락하게 되므로 어떤 두 변수 간에 다중 공선성이 존재할 경우 설명력이 더 적은 변수를 제거하고 모델을 재구성 한다.

다중공선성을 판단하는 방법에는 여러가지가 있지만,  
R에서는 vif 함수를 사용해 VIF값을 간단히 구할 수 있으며, 보통 VIF 값이 4가 넘으면 다중공선성이 존재한다고 본다.

아래와 같이 car 패키지의 vif 함수를 사용하여 다중공선성 문제를 판단한다.

|  |  |
| --- | --- |
| 1  2  3  4  5 | require(car)  vif(model)    Fertility Agriculture Examination   Education    Catholic     2.855428    2.606925    3.754457    4.253857    2.349161 |

**2. R을 사용한 로지스틱 회귀분석 (Logistic regression in R)**

그림 출처: themegenerator.net

**로지스틱 회귀분석이란? (Logistic regression)**

로지스턱 회귀분석은 [선형 회귀분석](https://ko.wikipedia.org/wiki/%EC%84%A0%ED%98%95_%ED%9A%8C%EA%B7%80)과 달리 결과 변수(종속 변수)가 범주형 데이터인 경우에 사용되는 기법입니다.

선형 회귀분석 모델에서는 설명 변수를 입력하면 수치형 결과를 얻게 되는데요, 신장, 시험 성적, 연간 소득 따위를 예측하고 싶을 때는 선형 회귀분석을 사용하면 됩니다.

하지만 내가 예측하고자 하는 것이 수치화 하기 힘든 변수, 예를 들면 어떤 고객이 부도를 낼 것인지의 여부, 타이타닉 호에서 살아남을 것인지의 여부, 어떤 학생의 최종 학력 (초등학고, 중학교, 대학교, 대학원 이상) 일 경우에는 선형 회귀분석 대신 로지스틱 회귀분석을 사용하게 됩니다.

회귀분석이기 때문에 기본적으로 [지도 학습](https://ko.wikipedia.org/wiki/%EC%A7%80%EB%8F%84_%ED%95%99%EC%8A%B5) (Supervised Learning) 으로 분류 되며, 다양한 분야에서 분류 및 예측을 위한 기법으로 사용되고 있습니다.  제가 근무하고 있는 라이엇 게임즈에서도 플레이어 이탈을 예측하기 위한 모델링 과정에서 로지스틱 회귀 분석을 비롯한 다양한 지도 학습 기법을 사용한 바 있습니다.

**로지스틱 회귀분석 in R**

R을 사용하면 아주 간단히 로지스틱 회귀분석 모델을 만들고, 이 모델을 이용해서 신규 데이터가 주어졌을 때 결과 값을 예측하는 작업을 할 수 있습니다.

**1. 데이터셋 로드**

UCLA에서 제공하는 데이터 셋을 불러와서 데이터의 구조를 확인하면 아래와 같은 결과를 얻을 수 있습니다.

|  |  |
| --- | --- |
| 1  2  3 | data <- read.csv("http://www.ats.ucla.edu/stat/data/binary.csv")  str(data)  head(data) |

Assembly (x86)

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13 | 'data.frame': 400 obs. of  4 variables:  $ admit: int  0 1 1 1 0 1 1 0 1 0 ...  $ gre  : int  380 660 800 640 520 760 560 400 540 700 ...  $ gpa  : num  3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...  $ rank : int  3 3 1 4 4 2 1 2 3 2 ...  admit gre  gpa rank  1     0 380 3.61    3  2     1 660 3.67    3  3     1 800 4.00    1  4     1 640 3.19    4  5     0 520 2.93    4  6     1 760 3.00    2 |

admit 변수는 지원자의 입학 여부를 나타내며, 0 혹은 1의 값을 가집니다. gre, gpa 변수는 수치형 변수인 시험 점수이고, rank는 지원자의 학교 등급을 나타내며 1~4의 값을 가지는 범주형 변수입니다.

**2. 데이터 클린징 (Data cleansing)**

1번 단계에서 str 결과를 보면 rank 변수가 int로 설정되어 있는 것이 보이죠?  
분석에 들어가기 전 rank 변수를 factor 변수로 변경해 줍니다.

|  |  |
| --- | --- |
| 1  2 | data$rank <-as.factor(data$rank)  str(data) |

|  |  |
| --- | --- |
| 1  2  3  4  5 | data.frame': 400 obs. of  4 variables:  $ admit: int  0 1 1 1 0 1 1 0 1 0 ...  $ gre  : int  380 660 800 640 520 760 560 400 540 700 ...  $ gpa  : num  3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...  $ rank : Factor w/ 4 levels "1","2","3","4": 3 3 1 4 4 2 1 2 3 2 ... |

rank 변수의 유형이 Factor로 바뀌었습니다. 이제 모델 fitting을 해 볼까요?

**3.로지스틱 회귀분석 모델 작성 (MODEL FITTING)**

로지스틱 회귀분석 모델에서는 “glm” 함수를 사용하는데요, 선형 회귀분석에서의 “lm”과 유사한 함수라고 보시면 됩니다.

“glm”함수를 사용해서 모델을 만든 후 “summary” 함수로 모델 결과를 확인합니다.

|  |  |
| --- | --- |
| 1  2  3  4 | train<-data[1:200, ]  test<-data[201:400, ]  model <- glm(admit ~ gre + gpa + rank, data =train, family = "binomial")  summary(model) |

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26 | Call:  glm(formula = admit ~ gre + gpa + rank, family = "binomial",      data = train)    Deviance Residuals:      Min       1Q   Median       3Q      Max  -1.6154  -0.7872  -0.5325   0.8785   2.3242    Coefficients:               Estimate Std. Error z value Pr(>|z|)  (Intercept) -3.790221   1.750305  -2.165 0.030352 \*  gre          0.003483   0.001718   2.027 0.042640 \*  gpa          0.534007   0.519132   1.029 0.303642  rank2       -0.707194   0.476566  -1.484 0.137826  rank3       -1.803655   0.531352  -3.394 0.000688 \*\*\*  rank4       -1.959514   0.629189  -3.114 0.001844 \*\*  ---  Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1    (Dispersion parameter for binomial family taken to be 1)        Null deviance: 237.18  on 199  degrees of freedom  Residual deviance: 205.76  on 194  degrees of freedom  AIC: 217.76    Number of Fisher Scoring iterations: 4 |

**4. 모델 해석**

표의 가장 우측 열의 p-value가 0.05보다 작은 변수는 gre, rank3, rank4 이네요.  
이 세가지가 모두 통계적으로 유의하다고 해석 할 수 있습니다. rank3 의 p-value가 가장 낮은 것으로 보아 예측력이 좀더 강한 것으로 보이네요. Estimate 열의 부호가 마이너스(-) 이므로 랭크가 낮은 학교의 지원자의 합격 확률은 낮아진다고 볼 수 있겠네요.

(보다 정확히는 랭크 4 학교의 지원자는 그렇지 않은 지원자에 비해 합격 확률의 Log가 1.95만큼 감소하게 된다는 의미입니다.)

anova 분석을 해 볼까요?

|  |  |
| --- | --- |
| 1 | anova(model, test="Chisq") |

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16 | Analysis of Deviance Table    Model: binomial, link: logit    Response: admit    Terms added sequentially (first to last)           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)  NULL                   199     237.18  gre   1  12.8706       198     224.31 0.0003338 \*\*\*  gpa   1   0.5607       197     223.75 0.4539719  rank  3  17.9894       194     205.76 0.0004421 \*\*\*  ---  Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |

Null deviance와 residual deviance의 차이는 Null 모델과 비교하여 해당 변수가 모델에 포함 되었을 때 모델의 성능이 얼마나 나아지는지를 보여주는 수치입니다. rank 가 포함되었을 때 resid. Dev 의 값이 역시 가장 많이 하락하네요.

로지스틱 회귀분석에도 선형 회귀 분석에서의   R2 와 유사한 개념이 존재하는데요, Mcfadden R2으로 모델 fit을 확인할 수 있습니다.  
“pscl” 패키지의 pR2 함수를 사용하여 Mcfadden R2를 간단히 알아 볼 수 있는데요.  
R2값이 0.2인 것으로 보아 모델이 train 데이터 셋의 분산의 약 20% 정도 밖에 설명해 주지 못한다고 봐야 겠네요.

|  |  |
| --- | --- |
| 1  2  3  4 | library(pscl)  pR2(model)           llh      llhNull           G2     McFadden         r2ML         r2CU  -102.8802728 -118.5906635   31.4207814    0.1324758    0.1453847    0.2093280 |

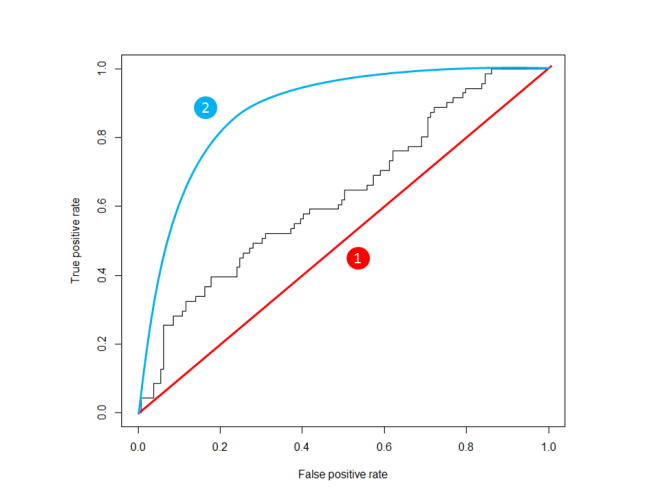
**5. 모델 평가**

로지스틱 회귀 모델의 평가를 하기 위해 [ROC 커브](https://en.wikipedia.org/wiki/Receiver_operating_characteristic)를 그려볼까요?

ROC 커브 아래의 영역 AUC가 1에 가까울 수록 모델의 예측 정확도가 높다고 할 수 있있습니다. “ROCR” 패키지를 사용하면 로지스틱 회귀 모델의 ROC 곡선을 그릴 수 있습니다.

|  |  |
| --- | --- |
| 1  2  3  4  5 | library(ROCR)  p <- predict(model, newdata=test, type="response")  pr <- prediction(p, test$admit)  prf <- performance(pr, measure = "tpr", x.measure = "fpr")  plot(prf) |

아래 그래프는 train 데이터 셋으로 만든 모델을 test 데이터 셋에 적용했을 경우의 ROC 곡선입니다.  ROC 곡선이 reference line인 1번 라인에서 2번 라인으로 가까워 질 수록 모델의 정확도는 향상되는데요, 그래프가 reference line에 가까운 것으로 보아 모델의 성능이 그다지 뛰어나지는 않은 것 같네요.



ROC 그래프 아래 영역의 넓이인  [AUC](https://en.wikipedia.org/wiki/Area_under_the_curve_(pharmacokinetics))를 구해볼까요?

ROC 곡선이 2번 방향을 향할 수록 AUC는 넓어지는데요,  
모든 case의 결과를 완벽하게 예측하는 모델인 경우 AUC는 1,  무작위로 예측한 모델과 다를 바가 없는 경우 (=reference line) AUC는 0.5의 값을 가지게 됩니다.

AUC를 판단하는 대략적인 기준은 아래와 같습니다.

excellent =  0.9~1  
good = 0.8~0.9  
fair = 0.7~0.8  
poor = 0.6~0.7  
fail = 0.5~0.6

예시 모델의 AUC를 볼까요? 0.632로 poor 한 수준으로 보입니다.

|  |  |
| --- | --- |
| 1  2  3  4 | auc <- performance(pr, measure = "auc")  auc <- auc@y.values[[1]]  auc  [1] 0.6323834 |

결과를 종합해 보면 모델은 유의하지만, 예측력은 그다지 높지는 않은 것으로 보이네요.  
설명 변수를 추가하고, 유의하지 않은 변수를 제거하고 여러 번 모델을 다시 fitting 하는 작업이 필요할 것 같습니다.

작업 하고 계신 모델의 성능이 좋지 않다고 너무 낙담하지는 마세요.  
모델의 성능을 높이기 위해서는 아래 세 가지 작업을 반복 수행해야 합니다.  
1. 적당한 설명 변수를 입력 하는 것  
2. 적절한 모델/알고리즘을 선택하는 것  
3. 위 1~2번을 반복

대부분의 머신 러닝을 사용한 예측 모델은 위 과정을 반복하면서 더욱 정교해지니까요!