Logistic Regression

Logistic model

다중선형회귀는 연속형 변수 Y와 수치형 설명변수 X 간의 관계가 선형임을 가정하고 오차제곱합이 최소가 되게 모수를 추정하는 것이다.

또한 선형성을 만족하지 못할때, 가법성은 유지해주면서 모델은 만드는 일반화 가법 모형 등과 같은 것도 많이 있다. 위의 모델들은 계수를 구하기 위해 최소 제곱법을 사용한다.(선형회귀에 대해선 뒤쪽에서 설명하겠다)

하지만 Y가 범주형 변수일때는 Logistic model을 사용 한다. 로지스틱은 계수를 추정하기 위해 MLE를 사용하는데 최대 우도 추정 방법 부터 살펴 보자.

최대우도법

최대우도법은 어떤 확률변수에서 표집한 값들을 토대로 그 확률변수의 모수를 구하는 방법이다. 모수가 주어 졌을 때, 원하는 값들이 나올 가능성을 최대로 만드는 모수를 선택하는 방법이다.(여기서 표본을 추출한 확률밀도함수를 알고 있음을 전제로 한다. 그럴때 최대우도법을 사용할 수 있다.)

예시로 정규분포에서 표본을 추출했을 때, 모분포의 평균과 분산을 추정해보자.

우리가 흔히 알고 있는 Consistency,..(수통이 기억이 나지 않네요)등등 을 만족하는 추정값은 아래와 같다.

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$$

최대우도법

$$f_{\mu,\sigma^2}(x_i)=\frac{1}{\sigma\sqrt{2\pi}}exp(-\frac{(x_i-\mu)^2}{2\sigma^2})$$

$$L^*(\theta)=\sum_{i=1}^{n} log\frac{1}{\sigma\sqrt{2\pi}}exp(-\frac{(x_i-\mu)^2}{2\sigma^2})$$

$$=\sum_{i=1}^{n}\{log(exp(-\frac{(x_i-\mu)^2}{2\sigma^2}))-(log(\sigma\sqrt{2\pi}))\}$$

$$=\sum_{i=1}^{n}\{-\frac{(x_i-\mu)^2}{2\sigma^2}-log(\sigma)-log(\sqrt{2\pi})\}$$

우도함수에 로그를 취하면

오른쪽과 같은 모습이다.

해당 함수에서 u,세타에 대해

각각 편미분을 취해 0이 나오는 값이 곧 추정값이다.

최대우도법

$$\frac{\partial L^*(\theta)}{\partial p} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial}{\partial \mu}(x_i^2 - 2x_i\mu + \mu^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(-2x_i + 2\mu)$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}x_i - n\mu\right) = 0$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n}x_i$$

$$\frac{\partial L^*(\theta)}{\partial \sigma} = -\frac{n}{\sigma} - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\frac{\partial}{\partial \sigma}\left(\frac{1}{\sigma^2}\right)$$

$$= -\frac{n}{\sigma} + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 \times 2\frac{1}{\sigma}\left(-\frac{1}{\sigma^2}\right)$$

$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

파라미터 추정

그렇다면, 왜 이 함수가 로지스틱이라는 이름을 갖게 되었고, 계수를 어떻게 추정하는지 알아보자 .

$$P(Y=1|X=\vec{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$
$$= \vec{\beta}^T \vec{x}$$

위와 같이 식을 형성하면 좌변과 우변의 범위가 맞지 않다.

$$\frac{P(Y=1|X=\vec{x})}{1-P(Y=1|X=\vec{x})} = \vec{\beta}^T \vec{x}$$

위의 식(오즈비)로 하더라도 좌변은 0~무한, 우변은 -무한~무한이다.

파라미터 추정

$$\log\left(\frac{P(Y=1|X=\vec{x})}{1-P(Y=1|X=\vec{x})}\right)=\vec{\beta}^{T}\vec{x}$$

하지만 위와 같이 log를 취해주면 좌변과 우변의 범위가 같게 된다.

$$\therefore P(Y=1|X=\vec{x})=\frac{1}{1+e^{-\vec{\beta}^{T}\vec{x}}}$$

결국 P(Y=1|X=x)에 대해 정리하면, 이 값은 로지스틱 함수를 따르게 된다. 그래서 로지스틱이라는 이름이 붙게 되었고, 확률값이 로지스틱 함수로부터 추출되었다고 가정 함으로서 최대우도법을 통해 계수를 구할 수 있는 것이다 .

파라미터 추정

그렇다면 계수를 구하기 위한 우도함수는 다음과 같이 정의할 수 있다

$$L = \prod_i \sigma(\beta^T \overrightarrow{x_i})^{y_i} \left\{1 - \sigma(\beta^T \overrightarrow{x_i})\right\}^{1-y_i} \qquad \ln L = \sum_i y_i \ln\left\{\sigma(\beta^T \overrightarrow{x_i})\right\} + \sum_i (1-y_i) \ln\left\{1 - \sigma(\beta^T \overrightarrow{x_i})\right\}$$

Y는 0,1의 값을 가지므로 베르누이분포를 따르고, P 즉 확률값은 로지스틱 함수를 따른다고 보면 우도 함수는 위와 같고, 앞에서 말했던 최대우도법을 통해 위 우도함수의 계수를 구할 수 있다.

하지만 위 로그 우도함수는 추정 대상 계수에 대해 비선형이기 때문에 선형회귀와 같이 명시적인 해가 존재하지 않음으로 경사하강법이나, 뉴턴 랩슨?, SGD(Stochastic Gradient Descent)등을 사용하게 된다.

파라미터 추정

하지만 위의 방법으로 모델이 수렴하지 않을 수 있다. 이는 반복 처리로써 적합한 해를 찾을 수 없기에 계수가 중요한 의미를 지니지 않음을 시사한다. 수렴에 실패하는 대표적인 이유는 사건에 매우 큰 영향력을 미치는 예측변수의 사용, 다중공선성, 희소성, 완분성 들이 있다.

https://ko.wikipedia.org/wiki/%EB%A1%9C%EC%A7%80%EC%8A%A4%ED%8B%B1_%ED%9A%8C%EA%B7%80#%EB%8F%85%EB%A6%BD_%EB%B3%80%EC%88%98(위키피디아)

독립변수와 종속변수

독립변수

독립 변수는 실제 값, 이진 값, 범주 등 어떤 형태가 올 수 있다.(너무 비대칭 정도가 심하면 log를 취해주긴 하더라)

종속변수

$$Y_i|x_{1,i},\ldots,x_{m,i}\ Bernoulli(p_i)$$

$$\mathbb{E}[Y_i|x_{1,i},\ldots,x_{m,i}] = p_i$$

$$\Pr(Y_i = y_i \mid \mathbf{X}_i) = {p_i}^{y_i}(1-p_i)^{1-y_i} = \left(\frac{1}{1+e^{-\beta\cdot\mathbf{X}_i}}\right)^{y_i}\left(1-\frac{1}{1+e^{-\beta\cdot\mathbf{X}_i}}\right)^{1-y_i}$$

즉 특정 독립변수가 주어졌을 때, 종속 변수는 위의 확률값으로 변환해서 나올 수 있다. (ln(오즈비) = BX지만, 이를 변형 하면 확률값은 위와 같다는 의미)

Logistic을 사용한 예시

독립변수로 모두 범주형 자료만 있을 때.

		CA		
Gender	ECG	Disease	No Disease	Total
Female	< 0.1	4	11	15
Female	≥ 0.1	8	10	18
Male	< 0.1	9	9	18
Male	≥ 0.1	21	6	27

1.우선 오즈비를 확인

Marginal OR between CA and ECG = 0.3586

Conditional ORs between CA and ECG at male = 0.285714

Conditional ORs between CA and ECG at female = 0.4545

Logistic을 사용한 예시

2. Pearson 카이 스퀘어 테스트로 CA와 ECG의 독립성을 확인

그 이전에, Breslow-Day 검정으로 Gender가 주어졌을 때, CA와 ECG의 공통오즈비가 동질연광성을 가지고 있다는 것을 검정한다.

```
        Breslow-Day test on Homogeneity of Odds Ratios

data:  count
X-squared = 0.21549, df = 1, p-value = 0.6425
```

귀무가설을 기각하지 못하므로, 각 성별에서 오즈비가 유의하게 동일하다고 볼 수 있다.

Logistic을 사용한 예시

3. CMH test를 통해 Gender를 제어했을 때, CA와 ECG가 서로 독립임을 test한다.

```
data:  count
Mantel-Haenszel X-squared = 4.5026, df = 1, p-value = 0.03384
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.1328392 0.9289241
sample estimates:
common odds ratio
        0.3512798
```

CA와 ECG가 서로 독립이라는 귀무가설을 기가한다. 따라서 두 변수 CA와 ECG간에 Gender를 제어했을 때 유의한 상관관계가 존재한다.

Logistic을 사용한 예시

4.단순히 카이 스퀘어테스트를 하는 것보단, CMH test를 통해 검정하는 것이 성별에 대한 정보도 누락시키지 않고 더 좋다. 대신, Breslow-Day test를 통해 동질연관성이 있음을 확인한 후에 CMH test를 하는 것이 바람직하다.

5.logistic regression model을 data fitting해서 관계를 살펴보자.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5947     0.3114   1.910   0.0562 .
ECG<0.1      -1.0255     0.4732  -2.167   0.0302 *
---
```

p-value가 유의함으로 ECG가>0.1일 때보다 <0.1일때, CA가 있을 확률이 exp(-1.02)배 감소함을 알 수 있다.

Logistic을 사용한 예시

6. Gender과의 상호작용 변수를 넣어서 ECG와 CA가 homoegeneous association 이 있는지를 볼 수 있다.

```
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.2528     0.4629   2.706   0.0068 **
SexFemale          -1.4759     0.6628  -2.227   0.0260 *
ECG<0.1            -1.2528     0.6607  -1.896   0.0579 .
SexFemale:ECG<0.1   0.4643     1.0012   0.464   0.6428
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.1983e+01  on 3  degrees of freedom
Residual deviance: 4.8850e-15  on 0  degrees of freedom
AIC: 21.095
```

상호작용 변수는 유의하지 않다. 즉 ECG와 CA는 동질연관성을 갖고 있다.

Logistic을 사용한 예시

7. 최종적으로 ECG와 gender변수를 넣고 확인을 해보면 된다.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1568     0.4036   2.866  0.00415 **
SexFemale    -1.2770     0.4980  -2.564  0.01034 *
ECG<0.1      -1.0545     0.4980  -2.118  0.03421 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Gender를 제어 했을 때, ECG<0.1일때, >0.1일 때보다. CA가 있을 확률이 exp(-1.0545)배다.

Logistic을 사용한 예시

독립변수에 연속형변수가 함께 있을 때

	color	spine	width	satell	weight	y	color.factor	spine.factor	color.factor2
1	3	3	28.3	8	3050	1	3	3	1
2	4	3	22.5	0	1550	0	4	3	1
3	2	1	26.0	9	2300	1	2	1	1
4	4	3	24.8	0	2100	0	4	3	1
5	4	3	26.0	4	2600	1	4	3	1
6	3	3	23.8	0	2100	0	3	3	1

1.우선, step wise를 통해 유의하지 않은 변수들 제거(이 부분은 임의로..)

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -12.7151     2.7617  -4.604 4.14e-06 ***
color.factor2   1.3299     0.8525   1.560   0.1188
color.factor3   1.4023     0.5484   2.557   0.0106 *
color.factor4   1.1061     0.5921   1.868   0.0617 .
width           0.4680     0.1055   4.434 9.26e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic을 사용한 예시

2. width 변수가 연속형이기 때문에, 적합도 검정을 위해 자료를 그룹화하고 Hosmer and Lemeshow검정을 사용한다.

Partition for the Hosmer and Lemeshow Test

		y = 1		y = 0	
Group	Total	Observed	Expected	Observed	Expected
1	17	5	3.75	12	13.25
2	17	6	6.33	11	10.67
3	17	7	8.25	10	8.75
4	17	9	9.88	8	7.12
5	19	14	12.19	5	6.81
6	17	11	11.95	6	5.05
7	17	12	12.90	5	4.10
8	18	16	15.01	2	2.99
9	17	14	14.86	3	2.14
10	17	17	15.88	0	1.12

```
        Hosmer and Lemeshow test (binary model)

data:  crabs$y, fitted(fit)
X-squared = 6.4078, df = 8, p-value = 0.6016
```

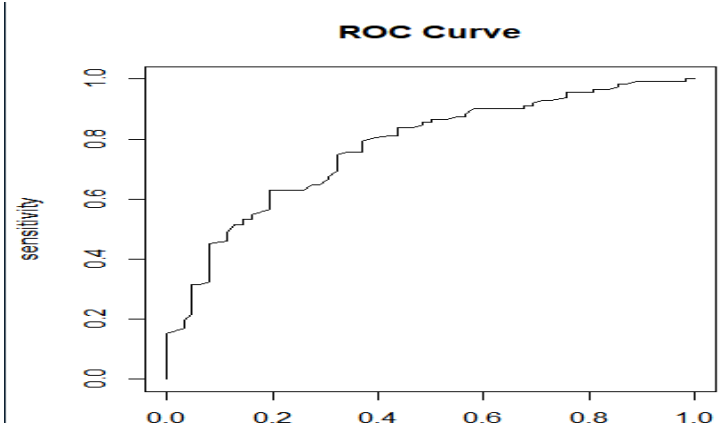
귀무가설을 기각하지 못한다. 즉 이모형은 적합함을 알 수 있다.

Logistic을 사용한 예시

3. 잔차확인을 통해 영향력이 있는 개체들을 확인할 수 있다.

4. AUC를 통해서도 모델의 성능을 확인할 수 있다.

```
[[1]]
[1] 0.7713601
```



Logistic을 사용한 예시

Proposed Nomogram Predicting the Individual Risk of Malignancy in the Patients With Branch Duct Type Intraductal Papillary Mucinous Neoplasms of the Pancreas

해당 논문은 AUC를 최대로 하는 logistic model을 만들고, Hosmer lemeshow를 calibration과 묶어 모델을 검정하였다.

해당 논문을 따라해본 R-code와 예시는 (논문 요약 및 코드)에 첨부 되어있다.

Logitstic을 사용한 예시

Y가 다범주 일때

	Happiness		
Income	Not	Pretty	Very
Below average	6	43	75
Average	6	113	178
Above average	6	57	117

```
fit = vglm(y ~ income, family=cumulative(parallel=TRUE), data=happy)
summary(fit)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept):1 -2.55518    0.72560  -3.521 0.000429 ***
(Intercept):2 -0.35129    0.26837  -1.309 0.190554
income:1      -0.22751    0.34120  -0.667 0.504907
income:2      -0.09615    0.12202  -0.788 0.430694
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic을 사용한 예시

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1 -2.55518    0.72560  -3.521 0.000429 ***
(Intercept):2 -0.35129    0.26837  -1.309 0.190554
income:1      -0.22751    0.34120  -0.667 0.504907
income:2      -0.09615    0.12202  -0.788 0.430694
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

해석:

기준을 very happy로 했기 때문에, income이 한단위 상승 했을 때, not veryhappy일 확률이 very happy일때 보다 exp(-0.22751)배 감소한다.

income이 한단위 상승 했을 때, prettyhappy일 확률이 very happy일때 보다 exp(-0.09615)배 감소한다.

Logistic을 사용한 예시

로그 라이클리후드 테스트를 통해 계수의 유의성을 확인한다.

```
fit0 = vglm(y ~ 1, family=multinomial(refLevel = "not"), data=happy)
LR.stat = -2*(logLik(fit0) - logLik(fit)); LR.stat
df = summary(fit0)@df.residual - summary(fit)@df.residual; df
p.value = 1 - pchisq(LR.stat, df=df); p.value
```

계산 결과 p-value는 `[1] 0.6237937` 로서 귀무가설을 기각하지 못한다.

즉 income은 유의하지 않다고 할 수 있다.

앞에서 말한것 처럼, 이 모델을 통해 각 범주에 속할 확률을 구하고, 가장 확률이 높은 범주에 Y가 배정된다.

Logistic을 사용한 예시

Y가 다범주 순서형 일때

proportional odds assumption을 확인한다.

```
fit.new = vglm(y ~ income, family=cumulative(parallel=FALSE), data=happy)
df = summary(fit)@df.residual - summary(fit.new)@df.residual
deviance(fit) - deviance(fit.new)
p.value = 1 - pchisq(deviance(fit) - deviance(fit.new), df=df)
p.value
```

이를 만족한다면,

```
fit = vglm(y ~ income, family=cumulative(parallel=TRUE), data=happy)
summary(fit)
```

cumulative(parallel=TRUE)옵션을 통해 모델을 만들 수 있다.

일반화 가법 모형을 이용한 Logistic

일반화 가법 모형은 쉽게 말하면, 선형성은 완화화고, 가법성은 유지해주는 모형이다. 비선형성이기에 예측력이 더 좋을 수 있다. 물론, 상호작용 변수를 넣어 줌으로서, 가법성 또한 완화시킬 수 있다.

Y를 질적변수로 하는 일반화 가법 모형의 개념은 따로 정리를 하고 있다. 하지만 그 이전에 Smoothing spline에 대한 개념이 필요한데 natural cubic spline까진 했지만 이부분은 하지 않았다. 다음에 또 보도록 하자.