14기 정규세션

ToBig's 13기 이재빈

회귀분석 추가자료

- Q. 수학/통계 베이스가 없어서 너무 어려웠어요 ㅜㅜ

- 당연합니다 ㅜㅜ 회귀분석은 학부생은 1년동안, 원생은 한학기동안 배우는 내용이예요..
  너무 어려우셨던 분들은 세세하게 하나하나 증명하는 방법을 공부하시기 보다는,
  제가 알려드린 내용과 강의자료를 통해 흐름을 우선적으로 잡으시는 것을 추천드려요 !!

- Q. 수학/통계 베이스를 쌓을 만한 자료 추천해주세요 !

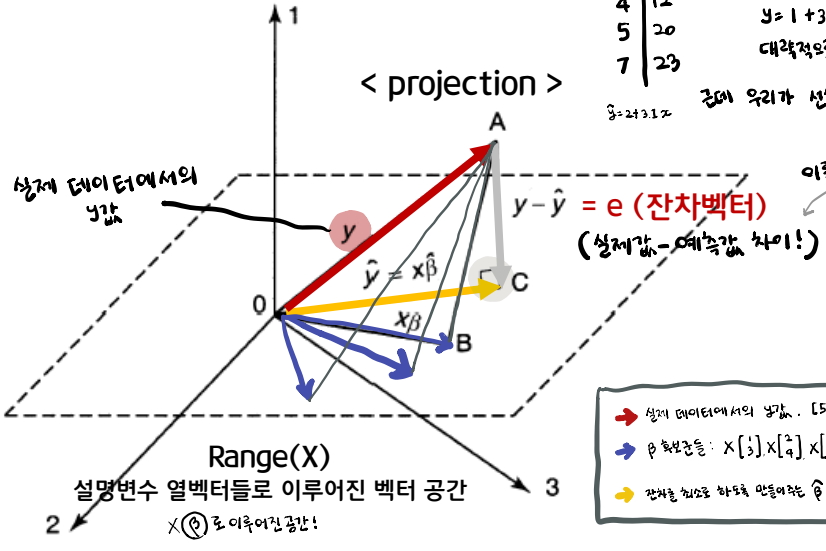
- 인공지능을 위한 수학 : https://book.naver.com/bookdb/book_detail.nhn?bid=14200387
  책 표지가 쫌 많이 글킨 한데 내용이 정말 괜찮다고 하더라구요 !!!
  AI를 위해 딱 필요한 수학만 쉽게 설명해 놓은 책이라고 합니다
- 투빅스 정보공유 게시판에 좋은 글이 적혀 있으니 한번 참고해 보시면 좋을 것 같습니다 ~!!
  http://www.datamarket.kr/xe/board_fpbt85/66140

Unit 02 | 선형 회귀분석

- Q. 정규방정식 기하학적 의미로 유도하기

x	y
1	5
3	12
4	12
5	20
7	23

$\hat{y} = 2 + 3.1x$



옆의 그림에서의 파란선

$y = X\beta + \varepsilon$ 에서, $\boxed{\beta\text{는 여러 개가 있을수 있어요!}}$

가령 옆의 데이터의 예시에서,

$y = 1 + 3x$, $y = 2 + 4x$, $y = 3 + 3x$ 등

대략적으로 보고 다양하게 추세를 예측할 수 있겠죠..?

근데 우리가 선형회귀분석을 하는 목적은, 우리가 가지고 있는 데이터를 통해 데이터의 특성을 반영할 수 있는 적절한 식을 예측하는 것이고,

이를 위해 $\boxed{(\text{실제값} - \text{예측값})}$을 최소화 하는 방향으로 적절한 $\boxed{\text{모수}(\hat{\beta})}$를 갖는 식을 도출하게 됩니다 ~~!

잔차벡터 $e = y - X\hat{\beta}$ 와 Range(X)의 거리가 최소가 되도록 하려면?

-> 두 벡터가 수직이어야 함 !

∵ 수직일 때 길이 최소

- 실제 데이터에서의 y값. $[5, 12, 12, 20, 23]^T$
- β 후보군들: $X\begin{bmatrix}1\\3\end{bmatrix}, X\begin{bmatrix}2\\4\end{bmatrix}, X\begin{bmatrix}3\\3\end{bmatrix}$
- 잔차를 최소로 하도록 만들어주는 $\hat{\beta}$: $X\begin{bmatrix}2\\3.1\end{bmatrix}$

$$X'(y - X\beta) = 0$$
$$X'y - X'X\beta = 0$$
$$\therefore \hat{\beta} = (X'X)^{-1}X'y$$

정규방정식 유도

Unit 02 | 선형 회귀분석

- Q. 회귀분석 표 해석

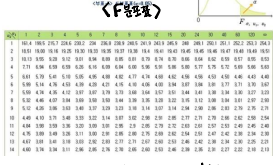
	자유도 (df)	제곱합 (SS)	제곱평균 (MS)	F
회귀 (Regression) SSR	p	SSR	MSR = SSR / p	F = MSR / MSE
잔차 (Residual) SSE	n-(p+1)	SSE	MSE = SSE / (n-p-1)	
총 (Total) SST	n-1	SST = SSR+SSE		

$H0: y = \beta_0$ (우리가 예측한 회귀식이 유의미하지 않다)

$H1: y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ (우리가 예측한 회귀식이 괜찮다!)

$H0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$

F > F(α, p, n-p-1) 이면 H0 기각



F(α, p, n-p-1) 은 여기서 찾습니다..

$F = \frac{MSR}{MSE}$ 값이 특정 값을 넘으면, H0를 기각하므로 우리가 예측한 회귀식이 괜찮다는 것을 의미합니다!

MSE = 회귀식이 설명하지 못하는 부분
-> MSE 값이 작을수록 좋음

SSE=잔차제곱합 = 회귀식으로 설명하지 못하는부분

(자유도) 현재 내가 가지고 있는 데이터 (n) 에서 얼만큼 자유로운지 = 제약이 없는지를 의미합니다

(cf) 단순회귀 제약조건 2개 = 1+1개
1. $\sum$ 잔차 = 0 2. $\sum$ 잔차 * x_i = 0

WHY? from LSE

$$\begin{cases} \frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0 \end{cases}$$

LSE에서 각각 설명변수들에 대해 미분해서 나온 식입니다 (2개) = n-2

설명변수가 $x_1, \ldots, x_p$ 까지 있다면, $\frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1}, \ldots, \frac{\partial L}{\partial \beta_p}$ 까지 해야 하므로 (p+1) 개가 생기겠고? ⇒ 그래서 자유도: n-(p+1) 됩니당

Unit 03 | 모형 진단

- Q. 교호작용에서 왜 연속형*연속형 범주는 고려하지 않죠..??

- 범주형*범주형과 범주형*연속형의 경우 교호작용 고려 시 선형회귀식으로 표현이 가능해져서 해석이 직관적이지만, 연속형*연속형의 경우 이차다항 항이 나오기 때문에 해석이 모호해지기 때문이예요 !

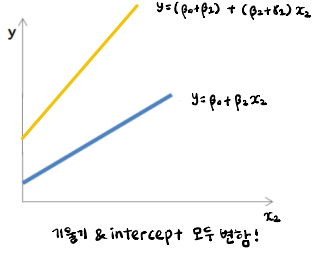
(ex) X_1: 음주 여부 (1/0) , X_1과 X_2의 교호작용 고려 : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 X_1 X_2$

음주를 하면 ⇔ $X_1 = 1$: $y = \beta_0 + \beta_1 + \beta_2 X_2 + \gamma_1 X_2$

$= (\beta_0 + \beta_1) + (\beta_2 + \gamma_1) X_2$

음주를 하지 않으면 ⇔ $X_1 = 0$: $y = \beta_0 + \beta_2 X_2$

→ 둘 다 선형회귀식으로 표현 가능!



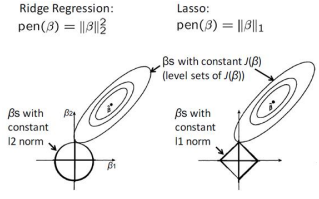
But.. X_1 X_2 모두 연속형이라면?

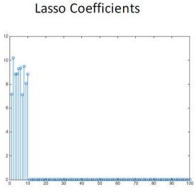
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 X_1 X_2$ 에서 γ_1을 어떻게 해석해야 할지 모호해집니다

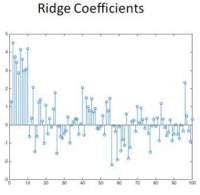
Unit 03 | 모형 진단

- Q. Ridge / Lasso 차이가 뭐예요 ?

- https://medium.com/mighty-data-science-bootcamp/linear-regression-ridge-lasso-elastic-net-fb8115c0a635 , https://brunch.co.kr/@itschloe1/11 두 개 글 한번 읽어보세요!!

- 





Ridge	Lasso
모든 변수의 계수의 값을 줄인다	중요한 몇 개 변수만 제외하고, 나머지 변수의 계수들은 0으로 줄인다
모델 성능 저하에 영향을 미칠 수 있음	정보가 손실됨에 따라 정확성 떨어질 수 있음

Ridge vs. Lasso

Ridge	Lasso
L_2-norm regularization	L_1-norm regularization
변수 선택 불가능	변수 선택 가능
Closed form solution 존재 (미분으로 구함)	Closed form solution이 존재하지 않음 (numerical optimization 이용)
변수 간 상관관계가 높은 상황 (collinearity)에서 좋은 예측 성능	변수 간 상관관계가 높은 상황에서 ridge에 비해 상대적으로 예측 성능이 떨어짐
크기가 큰 변수를 우선적으로 줄이는 경향이 있음	

ㄴ 고려대학교 김성범 교수님 강의자료

Unit 04 | 로지스틱 회귀분석

- Q. Likelihood 의 개념이 잘 와닿지 않습니다 ..

- 유투브 : https://www.youtube.com/watch?v=pYxNSUDSFH4 , 유투브 정리 :https://jjangjjong.tistory.com/41
- Probability vs Likelihood : https://blog.naver.com/sw4r/221361565730

- 위 세 개 링크를 보시면 좀 더 이해가 쉬우실 것 같아요 !

- $L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta)$
- Likelihood : 주어진 특정한 관찰된 데이터(우리가 갖고 있는 xi = 우리가 갖고 있는 데이터) 를 기반으로 하여, 모델의 파라미터들이 될 수 있는 가능성 (모델의 파라미터에 대한 확률!)
- MLE : 각각 값(xi)에 대한 Likelihood를 모두 곱해서, 그 결과값 L을 최대로 하는 추정값을 구하는 방법

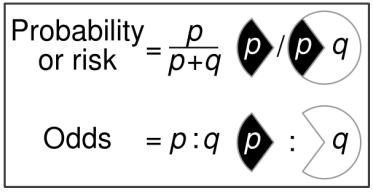
Unit 04 | 로지스틱 회귀분석

- Q. Odds 부분이 쫌 어려워요 ㅠㅠ

- https://no17.tistory.com/202 , https://gentlej90.tistory.com/69

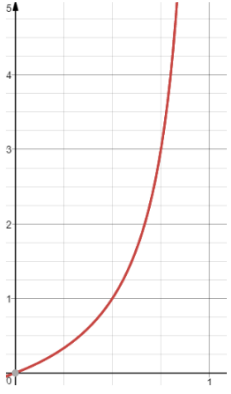
- Odds = $\frac{p}{1-p}$ 에서, p값이 0이면 Odds =0, p값이 1에 가까워지면 Odds = +∞ 값이 될 거예요

- 이를 일반적인 선형 회귀식으로 설명 가능하게끔 바꾸어 주려면, Odds 값에 log를 취해 선형 회귀식과 범위가 일치하게끔 만들어 주면 되는데, 이를 로짓변환 이라고 합니다 ~!
  log(Odds) = log($\frac{p}{1-p}$) : 범위 (-∞, +∞)



- 이를 통해 우리는 0과 1 사이의 값을 갖는 확률을 (-∞, +∞) 값을 갖는 값으로 변환시켰고, 일반 선형 회귀식과 범위를 일치시킴으로써, 적합된 회귀식을 통해 p값을 알아낼 수 있는 거예요 !
  $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$ (이게 로지스틱 "회귀"분석인 이유입니다..)

Odds



<참고> p

Unit 04 | 로지스틱 회귀분석

- Q. 생략된 수식 궁금해요 !

$$\log(Odds) = \log\left(\frac{p}{1-p}\right) = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \cdots + \widehat{\beta_k}x_k$$

$$\Rightarrow \quad \frac{p}{1-p} = e^{\widehat{\beta_0} + \widehat{\beta_1}x_1 + \cdots + \widehat{\beta_k}x_k}$$

$$p = (1-p)e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k} \rightarrow p(1+e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k}) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k} \rightarrow p = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k}}{1+e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k}}$$

양 변을 $e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k}$로 나누면

$$\Rightarrow \quad p = \frac{1}{1+e^{-(\widehat{\beta_0} + \widehat{\beta_1}x_1 + \cdots + \widehat{\beta_k}x_k)}} = \sigma(x, \beta)$$

Unit 04 | 로지스틱 회귀분석

- **Q. 생략된 수식 궁금해요 ! 2**

- **관측값 y_i에서의 확률 분포 :** $f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, i = 1,2,\ldots,n$

베르누이 시행

$$\begin{cases} y_i = 1 : f_i(y_i) = \pi_i & (\text{성공}) \\ y_i = 0 : f_i(y_i) = 1-\pi_i & (\text{실패}) \end{cases}$$

- Likelihood Function : $L(\mathbf{y}, \beta) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

- log-likelihood : $\ln L(\mathbf{y}, \beta) = \ln \prod_{i=1}^{n} f_i(y_i)$

$$= \sum_{i=1}^{n}\left[y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)\right] = \sum_{i=1}^{n}\left[y_i\left(\ln \pi_i - \ln(1-\pi_i)\right) + \ln(1-\pi_i)\right]$$

$$= \sum_{i=1}^{n}\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n} \ln(1-\pi_i)$$

Unit 04 | 로지스틱 회귀분석

- Q. 생략된 수식 궁금해요 ! 2

- log-likelihood :

$$\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n} \ln(1-\pi_i)$$

- $\log(Odds) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$
- $\pi_i = \frac{1}{1+e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}}$
- $(1-\pi_i) = \frac{e^{-(\beta_0 + \cdots + \beta_k X_k)}}{1+e^{-(\beta_0 + \cdots + \beta_k X_k)}} = \frac{1}{e^{\beta_0 + \cdots + \beta_k X_k} + 1}$

$$\ln L = \sum y_i(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k) - \sum \ln(1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k})$$

PPT P.63에 오타 있습니다ㅠㅠ + 아니고 − 임.. ㅠ 죄송해요

Q & A

들어주셔서 감사합니다.