

연속형자료 벌점화 회귀모형

벌점화 회귀모형의 기본적인 원리는 잔차에 벌점항 (Penalty) 을 더하여 회귀계수를 축소하는 것입니다. Lasso 회귀는 벌점항으로 회귀계수의 절대값의 합을 사용하고, Ridge 회귀는 벌점항으로 회귀계수의 제곱의 합을 사용합니다. Ridge는 입력변수가 전반적으로 비슷한 수준으로 출력변수에 영향을 미치는 경우에 사용하고, Lasso는 출력변수에 미치는 입력변수의 영향력 편차가 큰 경우에 사용합니다. Elastic Net 회귀는 Lasso의 변수축소 속성과 Ridge의 정규화 속성을 둘 다 갖는 모델입니다. 다수의 변수 간에 상관관계가 존재할 때 Elastic Net 회귀를 많이 사용합니다.

메뉴 호출하기

- 고급분석 > 벌점화 회귀모형>연속형자료벌점화회귀모형



• 변수설정 탭

연속형자료별점화회귀모형

변수설정 | 분석옵션 | 자료분할 | 출력옵션

데이터

전체변수

① 종속변수(필수)

② 질적변수(선택-1개이상가능)

③ 양적변수(선택-1개이상가능)

④ ▼주효과 ⑤ ▼교호작용

⑥ 최종모형

⑦ ☐ 상수항 포함하지 않음

⑧ ☒ 설명변수 표준화

삭제

도움말 | 재설정 | 확인 | 취소

메뉴 요소	설명
① 종속변수	모형화하고자 하는 종속변수를 전체변수로부터 선택합니다. 반드시 한 개의 양적 변수를 선택해야 합니다.
② 질적변수	설명변수에 포함된 변수들의 유형을 지정해줍니다. 질적변수로 지정된 변수는 문자로 인식되어 분석에 사용됩니다. 엑셀 시트 상에 가장 먼저 등장하는 수준이 기저범주(reference)로 인식됩니다. 질적변수와 양적 변수 중에서 적어도 하나의 변수를 선택하여 최종모형에 추가해야 분석이 가능합니다.
③ 양적변수	설명변수에 포함된 변수들의 유형을 지정해줍니다. 문자형 변수는 선택 될 수 없으며, 선택된 경우 분석에서 제외됩니다. 질적변수에 지정된 변수와 중복되어 선택될 수 없습니다.
④ 주효과	[질적변수]와 [양적변수]에서 유형이 지정된 변수를 1개 이상 선택한 상태에서 [주효과] 버튼을 클릭하면, 해당 변수들이 최종모형에 각각 주효과로 포함됩니다.
⑤ 교호작용	[질적변수]와 [양적변수]에서 유형이 지정된 변수를 2개 이상 선택한 상태에서 [교호작용] 버튼을 클릭하면, 해당 변수들의 교호작용이 최종모형에 포함됩니다.
⑥ 최종모형	주효과 또는 교호작용으로 정의된 변수들이 설명변수로 간주되어 모형에 포함됩니다. 포함된 주효과 또는 교호작용 중 삭제하고자 하는 항목이 있는 경우, 해당 항목을 선택한 뒤 [삭제] 버튼을 클릭하면 최종 모형에서 제외됩니다. 한 개 이상의 최종모형이 있어야 분석이 가능합니다.

• 변수설정 탭

연속형자료 별점화 회귀모형

변수설정

분석옵션

자료분할

출력옵션

데이터

전체변수

id

bweight

lowbw

gestwks

preterm

matage

hyp

sex

① 종속변수(필수)

>

<

변수 유형

② 질적변수(선택-1개이상가능)

>

<

③ 양적변수(선택-1개이상가능)

>

<

④ ▼주효과

⑤ ▼교호작용

⑥ 최종모형

삭제

⑦ ☐ 상수항 포함하지 않음

⑧ ☒ 설명변수 표준화

도움말

재설정

확인

취소

메뉴 요소	설명
⑦ 상수항 포함하지 않음	모형에 상수항(intercept)을 포함하지 않으려면 이 옵션을 선택합니다. 최종모형에 1개 이상의 설명변수가 포함되지 않은 경우, 이 옵션을 선택할 수 없습니다.
⑧ 설명변수 표준화	설명변수를 표준화 합니다.

• 분석옵션 탭

연속형자료 별점화 회귀모형

변수설정
분석옵션
자료분할
출력옵션

1
별점 방법

☒ Ridge
☐ Lasso
☐ Elastic net
Elastic net 별점 ($0 < \alpha < 1$)

2
탐색방법

☒ 격자 탐색 (Grid search)
개수
☐ 사용자 정의 (실험표로 구분)

3
교차검증

☒ K-fold 교차검증
K
☒ 교차검증 예측값을 훈련자료 적합값으로 설정

4
정확도 지표

☒ 평균표준오차(MSE) ☐ 평균절대오차(MAE)

도움말
재설정

확인
취소

메뉴 요소	설명
① 별점 방법	<p>별점 부여 방법 3가지 중 하나를 선택합니다.</p> <ul style="list-style-type: none"> Ridge : $\lambda \sum_{j=1}^p \beta_j^2$ 을 별점항으로 사용합니다. 변수 간 상관관계가 높은 상황에서 좋은 예측 성능을 보입니다. Lasso : $\lambda \sum_{j=1}^p \beta_j$ 를 별점항으로 사용합니다. 변수 간 상관관계가 높은 상황에서 Ridge에 비해 상대적으로 예측 성능이 떨어집니다. 하지만 Lasso에서는 Ridge에서 불가능한 변수선택이 가능합니다. Elastic net : $\lambda \left((1 - \alpha) \sum_{j=1}^p \beta_j^2 / 2 + \alpha \sum_{j=1}^p \beta_j \right)$ 을 별점항으로 사용합니다. Ridge와 Lasso속성에 대한 강도를 조정하는 방식입니다. 보통 성능이 좋다고 알려져 있습니다. Elastic net 별점 : 별점의 범위(α)를 직접 입력합니다. 0에서 1 사이의 값을 입력할 수 있으며, Default는 0.5입니다.
② 탐색방법	<p>최적의 λ를 찾기 위한 방법을 설정합니다.</p> <ul style="list-style-type: none"> 격자 탐색 : 격자 탐색 방법으로 최적의 λ를 찾습니다. 개수 : 후보군의 개수를 직접 지정합니다. 2 이상의 정수만 입력 가능하며, Default는 100입니다. 사용자 정의 : λ 값의 후보를 사용자가 직접 입력합니다. '사용자 지정'을 선택할 경우 하단의 박스가 활성화됩니다. 하단의 박스에, 사용할 후보값을 실험표로 구분하여 입력합니다.
③ 교차검증	<p>K-fold 교차검증을 시행합니다.</p> <ul style="list-style-type: none"> K : K 값을 직접 지정합니다. 2 이상의 정수만 입력 가능하며, Default는 10입니다. 교차검증 예측값을 훈련자료 적합값으로 설정 :

• 분석옵션 탭

연속형자료 별점화 회귀모형

변수설정
분석옵션
자료분할
출력옵션

1
별점 방법

☒ Ridge
☐ Lasso
☐ Elastic net
Elastic net 별점 ($0 < \alpha < 1$)

2
탐색 방법

☒ 격자 탐색 (Grid search)
개수
☐ 사용자 정의 (실패로 구분)

3
교차검증

☒ K-fold 교차검증
K
☒ 교차검증 예측값을 훈련자료 적합값으로 설정

4
정확도 지표

☒ 평균표준오차(MSE) ☐ 평균절대오차(MAE)

도움말
재설정
확인
취소

메뉴 요소	설명
④ 정확도 지표	<p>정확도에 사용할 지표 두 가지 중 하나를 선택합니다.</p> <ul style="list-style-type: none"> 평균표준오차 (MSE) : 오차의 제곱에 대해 평균을 취한 값을 사용합니다. 식은 다음과 같습니다. $MSE = \frac{1}{n} \sum_{i=1}^n (\hat{x}_i - x)^2$. 가장 많이 사용되는 지표입니다. 평균절대오차 (MAE) : 모든 절대 오차의 평균을 사용합니다. 식은 다음과 같습니다. $MAE = \frac{1}{n} \sum_{i=1}^n x_i - \hat{x}_i$

• 자료분할 탭

연속형자료별점화회귀모형

변수설정 분석옵션 자료분할 출력옵션

변수목록

- id
- bweight
- lowbw
- gestwks
- preterm
- matage
- hyp
- sex

① 훈련 및 시험

☒ 모든 데이터를 훈련에 이용

☐ 비율에 따라 임의로 분할

훈련(train) 자료 %

시험(test) 자료 %

☐ 변수로 분할

분할변수(1-훈련, 2-시험)

>

<

도움말 재설정 확인 취소

메뉴 요소

설명

① 훈련 및 시험

자료분할 방법 3가지 중 1개를 선택할 수 있습니다.

- 모든 데이터를 훈련에 이용 (Default) : 시험자료 없이 모든 개체를 회귀모형 적합에 사용합니다.
- 비율에 따라 임의로 분할 : 훈련자료와 시험자료의 비율을 설정하여 임의로 분할하는 방식입니다. Default 값은 훈련자료 70%, 시험자료가 30% 입니다. 사용자는 훈련자료에 0~100을 입력할 수 있으며, 시험자료에는 100에서 입력한 값을 뺀 수치가 자동으로 입력됩니다. 임의로 분할된 개체들 중 훈련자료와 시험자료의 인덱스를 저장하려면 [출력옵션]-[저장]-[자료분할지표]를 선택합니다.
- 변수로 분할 : 훈련자료와 시험자료로 사용될 개체가 결정되어 있는 경우 이 옵션을 선택합니다. 이때, 훈련자료에 해당하는 개체는 1, 시험자료에 해당하는 개체는 2의 값을 갖는 인덱스 변수를 분할변수로 지정해주어야 합니다.

• 출력옵션 탭

연속형자료 별점화 회귀모형

변수설정 분석옵션 자료분할 출력옵션

출력

① ☐ 조율모수에 따른 해의 변화

② ☒ 최적모형의 회귀분석 결과 출력

⑥ 회귀계수

☒ 신뢰구간
신뢰수준

☐ 분산팽창지수(VIF)

③ ☒ 분산분석표

제공합유형
☐ Type I ☐ Type II ☒ Type III

④ ☐ 적합도검정

⑤ ☐ 잔차진단그래프

저장

⑦ 별점화 모형

☐ 적합값 ☐ 예측값

⑧ ☐ 최적모형의 회귀분석 결과 저장

⑨ 훈련자료

☐ 적합값 ☐ 비표준화 잔차 ☐ 록의 거리
☐ 신뢰구간 ☐ 표준화 잔차 ☐ 해트 행렬의 대각원소
☐ 예측구간 ☐ 스튜던트화 잔차
신뢰수준

⑩ 시험자료

☐ 예측값
☐ 신뢰구간
☐ 예측구간
신뢰수준

⑪ ☐ 자료분할지표

도움말 재설정 확인 취소

메뉴 요소	설명
① 조율모수에 따른 해의 변화	조율모수(tuning parameter)에 따라 변화하는 해를 함께 출력합니다.
② 최적모형의 회귀분석 결과 출력	최적의 조율모수를 갖는 모형의 회귀분석 결과를 출력합니다. 이 옵션을 선택할 경우 [출력]-[회귀계수]와 '분산분석표', '적합도검정', '잔차진단그래프'가 활성화됩니다.
③ 분산분석표	[출력]-[최적모형의 회귀분석 결과 출력]을 선택할 경우 활성화됩니다. 회귀모형식에 대한 분산분석표(ANOVA table)이 출력됩니다. 종속변수의 변동량에 대한 회귀모형의 설명력을 판단하기 위한 제공합 계산 방식으로 다음의 3가지 옵션 중 1개를 선택할 수 있습니다. <ul style="list-style-type: none"> • Type I : 설명변수를 순차적으로 하나씩 추가하면서 제공합의 증가량을 계산하는 방식입니다. • Type II : 전체 회귀모형에서 주효과를 하나씩 제거하면서 제공합의 감소량을 계산하는 방식입니다. 최종모형이 주효과만으로 구성된 경우에 적합합니다. • Type III (Default) : 전체 회귀모형에서 주효과와 교호작용을 포함한 모든 효과를 하나씩 제거하면서 제공합의 감소량을 계산하는 방식입니다. 최종모형에 교호작용도 포함된 경우 적합합니다.

출력옵션 탭

연속형자료 별점화 회귀모형

변수설정 분석옵션 자료분할 출력옵션

출력

① ☐ 조출모수에 따른 해의 변화

② ☒ 최적모형의 회귀분석 결과 출력

⑥ 회귀계수

☒ 신뢰구간
신뢰수준

☐ 분산팽창지수(VIF)

③ ☒ 분산분석표
제공합유형
☐ Type I ☐ Type II ☒ Type III

④ ☐ 적합도검정

⑤ ☐ 잔차진단그래프

저장

⑦ 별점화 모형

☐ 적합값 ☐ 예측값

⑧ ☐ 최적모형의 회귀분석 결과 저장

⑨ 훈련자료

☐ 적합값 ☐ 비표준화 잔차 ☐ 쿡의 거리
☐ 신뢰구간 ☐ 표준화 잔차 ☐ 해트 행렬의 대각원소
☐ 예측구간 ☐ 스튜던트화 잔차
신뢰수준

⑩ 시험자료

☐ 예측값
☐ 신뢰구간
☐ 예측구간
신뢰수준

⑪ ☐ 자료분할지표

도움말 재설정 확인 취소

메뉴 요소	설명
④ 적합도검정	주어진 데이터에 대해 최종모형이 얼마나 적합한지 확인하기 위한 적합도 통계량을 계산하고, 적합도 검정을 수행합니다. <ul style="list-style-type: none"> 적합도 통계량 : Deviance, Pearson's chi-square, $-2(\log\text{-likelihood})$, AIC, BIC (2개 이상의 모형에 대해 통계량이 작을수록 적합도가 더 좋다고 판단합니다) 적합도 검정 : 설명변수가 없는 NULL 모형 대비 현재의 최종모형이 통계적으로 유의한 차이를 보이는지 검정하는 Likelihood Ratio test (LRT)를 수행합니다.
⑤ 잔차진단그래프	회귀진단을 위한 잔차진단 그래프 6가지가 출력됩니다. <ul style="list-style-type: none"> Residuals vs Fitted : 적합값 대비 잔차의 그래프 Normal Q-Q : 잔차의 QQ그림 Scale-Location : 적합값 대비 표준화잔차의 그래프 Cook's distance : 개체별 쿡의 거리 그래프 Constant Leverage : 요인별 표준화잔차의 그래프 Cook's distance vs Leverage : 쿡의 거리와 지렛값의 산점도
⑥ 회귀계수	<ul style="list-style-type: none"> 신뢰구간 : 회귀계수 추정값의 신뢰구간을 출력합니다. 신뢰수준 : [신뢰구간]을 선택할 경우 활성화됩니다. 0과 1 사이의 값을 입력할 수 있으며, Default는 0.95입니다. 분산팽창지수 (VIF) : 설명변수들의 다중공선성(multicollinearity)을 진단하는 지표인 분산팽창지수(variance inflation factor, VIF)를 출력합니다.

출력옵션 탭

연속형자료 별점화 회귀모형

변수설정 분석옵션 자료분할 출력옵션

출력

① ☐ 조율모수에 따른 해의 변화

② ☒ 최적모형의 회귀분석 결과 출력

⑥ 회귀계수

☒ 신뢰구간
신뢰수준

☐ 분산팽창지수(VIF)

③ ☒ 분산분석표
제공합유형
☐ Type I ☐ Type II ☒ Type III

④ ☐ 적합도검정

⑤ ☐ 잔차진단그래프

저장

⑦ **별점화 모형**

☐ 적합값 ☐ 예측값

⑧ ☐ 최적모형의 회귀분석 결과 저장

⑨ **훈련자료**

☐ 적합값 ☐ 비표준화 잔차 ☐ 쿡의 거리
☐ 신뢰구간 ☐ 표준화 잔차 ☐ 해트 행렬의 대각원소
☐ 예측구간 ☐ 스튜던트화 잔차

신뢰수준

⑩ **시험자료**

☐ 예측값
☐ 신뢰구간
☐ 예측구간

신뢰수준

⑪ ☐ 자료분할지표

도움말 재설정 확인 취소

메뉴 요소	설명
⑦ 저장 > 별점화 모형	<ul style="list-style-type: none"> 적합값 예측값
⑧ 최적모형의 회귀분석 결과 저장	최적의 조율 모수를 사용한 회귀분석 모형의 결과를 저장합니다. 이 옵션을 선택할 경우 하단의 '훈련자료', '시험자료' 박스가 활성화됩니다.
⑨ 저장 > 훈련자료	<p>[출력]-[최적모형의 회귀분석 결과 출력] 옵션을 선택해야 정상적인 결과가 출력됩니다. 최종모형 적합에 사용된 훈련자료에 대하여 다음 중 선택되는 통계량을 엑셀시트에 저장합니다. 괄호 안에 표기된 변수명으로 저장됩니다.</p> <ul style="list-style-type: none"> 적합값 : 최종모형으로 예측된 적합값 (Fitted_train_PLM) 신뢰구간 : [적합값]이 선택된 경우 활성화. 적합값의 신뢰구간 (Fitted_95CI_Lower_train_PLM / Fitted_95CI_Upper_train_PLM) 예측구간 : [적합값]이 선택된 경우 활성화. 적합값의 예측구간 (Fitted_95PI_Lower_train_PLM / Fitted_95PI_Upper_train_PLM) 신뢰수준 : [적합값]-[신뢰구간] 혹은 [예측구간] 선택 시 활성화됩니다. 0과 1 사이의 사이 값을 입력할 수 있으며, Default는 0.95입니다. 비표준화 잔차 : 적합값과 실제값의 차이 (unstdResid_train_PLM) 표준화 잔차 : 잔차를 표준편차로 나눈 값 (stdResid_train_PLM) 스튜던트화 잔차 : 해당 개체를 제외한 상태에서 계산된 표준편차로 잔차를 나눈 값 (studResid_train_PLM) 쿡의 거리 : 개별 개체들이 모형에 미치는 영향력을 평가하기 위해, 잔차와 지렛값을 동시에 고려한 척도 (CookDist_train_PLM) 해트 행렬의 대각원소 : 해당 개체와 나머지 개체의 평균의 차이인 지렛값 (HatValue_train_PLM)

• 출력옵션 탭

연속형자료 별점화 회귀모형

변수설정 분석옵션 자료분할 출력옵션

출력

① ☐ 조출모수에 따른 해의 변화

② ☒ 최적모형의 회귀분석 결과 출력

⑥ 회귀계수

☒ 신뢰구간
신뢰수준

☐ 분산팽창지수(VIF)

③ ☒ 분산분석표
제공합유형
☐ Type I ☐ Type II ☒ Type III

④ ☐ 적합도검정

⑤ ☐ 잔차진단그래프

저장

⑦ 별점화 모형

☐ 적합값 ☐ 예측값

⑧ ☐ 최적모형의 회귀분석 결과 저장

⑨ 훈련자료

☐ 적합값 ☐ 비표준화 잔차 ☐ 록의 거리
☐ 신뢰구간 ☐ 표준화 잔차 ☐ 해트 행렬의 대각원소
☐ 예측구간 ☐ 스튜던트화 잔차

⑩ 시험자료

☐ 예측값
☐ 신뢰구간
☐ 예측구간
신뢰수준

⑪ ☐ 자료분할지표

도움말 재설정 확인 취소

메뉴 요소	설명
⑩ 시험자료	<p>[자료분할]-'비율에 따라 임의로 분할' 또는 '변수로 분할'이 선택된 경우 활성화됩니다. 시험자료에 대하여 다음 중 선택되는 통계량을 엑셀 시트에 저장합니다. 괄호 안에 표기된 변수명으로 저장됩니다.</p> <ul style="list-style-type: none"> 예측값 : 최종모형으로 예측된 값 (Predicted_testing_PLM) 신뢰구간 : [예측값]이 선택된 경우 활성화. 예측값의 신뢰구간 (Predicted_95CI_Lower_testing_PLM / Predicted_95CI_Upper_testing_PLM) 예측구간 : [예측값]이 선택된 경우 활성화. 예측값의 예측구간 (Predicted_95PI_Lower_testing_PLM / Predicted_95PI_Upper_testing_PLM) - 신뢰수준 : [예측값]-[신뢰구간] 혹은 [예측구간] 선택 시 활성화됩니다. 0과 1 사이의 사이 값을 입력할 수 있으며, Default는 0.95입니다.
⑪ 자료분할지표	<p>[자료분할] 탭에서 '변수로 분할'에서 분할된 개체들 중 훈련자료와 시험자료의 인덱스를 엑셀 시트에 저장합니다. 저장된 변수명은 "Partition_idx_PLM"으로 훈련자료의 경우 'training', 시험자료의 경우 'testing'의 인덱스를 갖습니다.</p>