6장 다중 선형 회귀분석

1. 서론

전통적인 통계학에서 추론을 목적으로 회귀모델을 적합 vs 적합시킨 모델을 사용.

예측이 목적 - 모델적합 vs 이의 사용에 대한 차이점.

↗ 전통적인 통계학에서 많이 사용.

다중선형회귀모델 - 정량적인 종속변수 Y (출력변수, 반응변수)
예측변수 $X_1, X_2, X_3 \cdots X_p$ (독립변수, 입력변수, 회귀변수, 공변량)
→ 사이의 관계를 적합시키기 위해 사용.

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$ 의 관계를 가정

$\beta_0, \beta_1, \beta_2 \cdots$ 는 회귀계수, ε은 잡음 (noise) 주어진 데이터는 회귀계수를 추정하고 잡음을 정량화하는데 사용.

회귀모델링 - 회귀계수 추정뿐만 아니라 어떤 예측변수를 어떤 형식으로 포함할지 선택하는 것을 의미.

↳ 다양한 예측 모델링 분야에 적용 가능.

2. 설명모델과 예측모델의 모델링.

출력에 대한 입력의 평균 효과에 대한 설명 또는 수량화 (각각, 설명적 혹은 기술적 과제)

주어진 입력값을 활용하여 새 레코드의 결과값을 예측 (예측과제)

↳ 이 2가지 목적은 통계에서의 회귀와 같음.

Ex) 의사결정에서 다른 모든 요소의 변화량을 고려하지 않을 때, X_2의 증가는 Y를 5포인트 증가 등과 같은 문장 생성.

↳ 이 관계가 인과관계로 알려진 경우 이를 설명 모델링.

인과관계 구조가 확실하지 않은 경우 입·출력 사이의 연관성을 정량화 → 기술모델링

예측분석에서는 생성된 모델을 활용하여 새로운 레코드에 대한 예측이 중점.

설명모델과 예측모델은 모델링 단계와 성능평가 방법이 다르므로 모델 선택의 목적이 설명적인지 예측적인지 구분 필요.

설명 및 기술모델링 - 평균레코드의 모델링에 초점 → 데이터에 가장 적합한 모델선택

예측모델링 - 새로운 개별 레코드에 대한 예측이 가장 뛰어난 모델 선택 → 기존 데이터를 잘 맞추는 모델이 예측력이 떨어질 수도 있음.

여기서는 예측에 초점.

3. 회귀식의 추정과 예측.

포함시킬 입력변수 및 형태가 결정되면 보통 최소 제곱법 (OLS; ordinary least squares)로 회귀계수 추정.

↳ 실제 값 (Y)과 예측 값 ($\hat{Y}$) 사이의 편차 제곱합을 최소화시키는 추정치 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 \cdots \hat{\beta}_p$를 찾음.

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$ → 이 식에 의한 예측이 가장 좋은 예측

가정
① 잡음 ε (또는 Y)는 표준 정규분포를 따른다. 이러 가정을 만족할 때

② 선택된 예측변수가 적절하고 선형성을 따른다.

③ 관측치들이 서로 독립적이다.

④ 각 예측변수에 대한 Y값의 변동성은 예측변수의 값에 상관없이 동일하다. (등분산성)

예측값이 불편값 (예측값의 평균 = 실제값) 이고, 평균제곱오차가 가장 작다.

예측이 목표인 경우 노이즈가 임의적인 분포를 따르더라도 예측의 추정값은 매우 좋은 결과.

↳ 위 식과 같이 정의된 선형 모델에서는 최소제곱 추정값이 가장 작은 평균 제곱 오차를 가짐.

4. 선형 회귀분석의 변수 선택.

· 예측변수의 수 줄이기.

변수의 수를 줄여야하는 이유

① 미래에 대한 예측을 할 때, 예측변수들 전부를 수집하는 것이 불가능하거나 비용↑

② 적은 수의 예측변수를 사용하면 더 정확한 추정 가능.

③ 예측 변수가 많을수록 결측치가 있을 확률도 증가. (결측치 처리가 힘들어짐)

④ 간결성은 좋은 모델의 중요한 성질.

⑤ 변수가 많은 경우 다중공선성(multicollinearity)로 회귀계수 추정치가 불안할 수 있음.

↳ 두 개 이상의 예측변수가 종속변수에 유사한 선형 관계 공유

⑥ 종속변수와 상관관계가 없는 예측변수를 사용하면 예측의 분산이 증가할 수 있다.

⑦ 종속변수와 실제 상관관계가 있는 예측변수를 누락시키면 예측의 평균 오차 혹은 bias가 증가할 수 있다.

↳ 예측변수의 수가 너무 적거나 많은 경우에 대한 상충 관계 (trade-off)

약간의 bias를 허용하면 예측의 분산을 줄일 수 있다.

bias-variance trade-off는 예측변수가 많을 때 특히 중요.

· 예측변수의 수를 어떻게 줄일 것인가.

첫 번째 단계는 그 분야의 지식을 활용하는 것.

↳ 여러가지 예측변수들이 무엇을 측정하고 있는지, 왜 이 변수들이 종속변수의 반응예측에 적절한지.

⇒ 예측변수들이 분별력있는 예측변수들이 되도록 그 개수를 줄여나가야 함.

예측변수 삭제의 이유는 정보수집 비용, 부정확성, 다른 변수와의 높은 상관관계, 다수의 결측치, 복잡성성 등이 있음.

요약통계량과 그래프, 즉 빈도와 상관관계 테이블, 예측변수 등의 요약통계량, 산점도, 결측값의 개수

⇒ 잠재적인 예측변수를 조사하는데 유용.

두 번째 단계는 계산적과 통계적 유의성을 이용.

전역탐색 - 예측변수의 가능한 모든 조합으로부터 회귀모델을 적합시켜 최적의 예측변수집합을 찾는 방법.

↳ 다소 비현실적이어서 많이 쓰이지 않음.

적당한 P 값을 찾기 위한 부분집합의 수가 매우 많기 때문에 가장 가능성이 높은 부분집합이 어떤 것인지 검토하여 그로부터 예측변수를 선택하는 방법이 필요.

↳ 너무 단순한 모델(과소적합 모델)을 선택해서 중요한 변수들을 포함하지 않게 되거나,

너무 복잡한 모델(과적합 모델)을 선택하여 노이즈까지 학습하게 되는 경우에 주의.

모델들을 평가하고 비교하는 기준은 학습데이터에 대한 적합에 근거.

수정 결정계수 (R^2_{adj}, adjusted R^2) $\quad R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1-R^2)$ ← R^2은 모델에서 설명할 수 있는 변동성의 비율. (단일 예측변수를 가진 모델에서 이는 상관계수의 제곱에 해당)

R^2와 마찬가지로 R^2_{adj}의 값이 높으면 보다 나은 적합성을 가짐.

↳ 모델에 사용된 예측변수의 수를 고려하지 않는 R^2과는 달리, R^2_{adj}는 예측변수의 수에 대한 Penalty를 부여.

⇒ 정보량의 증가가 아니라 단순히 예측변수의 수만을 증가시켰을 경우에도 발생하는 R^2의 인위적인 증가를 배제하는 효과

↳ 부분집합을 고르기 위해 R^2_{adj}를 사용하는 것은 $\hat{\sigma}^2$을 최소화시키는 부분집합을 찾는 것과 동일.

두 번째 과소적합 / 과적합 균형 위한 기준 - AIC (Akaike Information Criterion)
BIC (Schwartz's Bayesian Information Criterion)
↳ 적합성뿐 아니라 매개변수의 수에 대한 벌점도 측정.

↗ 더 나은 모델일수록 두 지표의 값이 작음.

같은 데이터를 적합시킨 여러 모델을 비교할 때 사용. → 정보이론을 기반으로 예측에러를 측정.

부분집합을 고르기 위한 세번째 기준: 맬로우 C_p (Mallow's C_p)

↳ 모델에서 일부 예측변수를 제외하면 예측의 변동성이 줄어들수도 있지만, 모든 예측변수를 사용하는 완전모델에서는 bias가 없다고 가정. → 부분집합의 모델이 bias가 없다면 평균 C_p는 $P+1$ (P: 예측변수의 수 + 1)

약간의 bias가 있는 부분집합모델을 식별하기 위한 합리적인 접근방식. C_p가 대략 $P+1$ 정도의 값을 가지는 부분집합을 검사. → 좋은 모델이란 $P+1$에 근사한 C_p값을 가지고 있고 예측변수의 수 P가 작은 모델

$$C_p = \frac{SSE}{\hat{\sigma}^2_{Full}} + 2(P+1) - n$$

$\hat{\sigma}^2_{Full}$ - 모든 예측변수를 포함하는 완전모델의 σ^2에 대한 추정치. ← 여기의 신뢰성에 높은 의존. ⇒ 예측변수의 수에 비해 상대적으로 많은 관측치 필요.

- 선형회귀에서 표본의 수가 많다면 맬로우 C_p와 AIC는 동일.

- 부분집합의 크기가 고정되어있다면 R^2, R^2_{adj}, C_p, AIC, BIC는 모두 같은 집합 선택.

◦ 대표적인 부분집합선택 알고리즘.

모든 가능한 회귀모델로 이루어진 공간에 대해 부분적이고 반복적인 탐색 시행. ⇒ 결과는 하나의 부분 집합. 최적 예측변수

↳ 전역탐색에 비해 '좋은' 예측변수를 놓칠 위험이 있음. 예측변수의 수가 적당하다면 전역탐색 권장.

전방 선택 방법 (forward Selection)

↳ 처음에는 예측변수가 없는 상태에서 하나씩 추가. → 추가되는 변수는 R^2증가에 가장 큰 기여를 하는 변수추가.

예측변수의 기여도가 통계적으로 유의하지 않을 때 종료.

주된 단점.

⇒ 둘/그 이상의 예측변수와 함께 사용될 때는 효과적이나 단일변수로서 낮은 설명을 보이는 변수는 누락 가능성.

후방 소거법 (backward elimination)

모든 변수들을 포함해서 시작했다가 단계별로 가장 유용하지않은 통계적으로 유용하지 않은 변수들을 제거.

제거되지 않고 남아있는 변수들이 모두 유의하면 종료.

⇒ 모든 예측변수들을 포함하는 초기모델을 계산하는데 시간이 많이 걸리고 불안정.

단계적 선택방법 (stepwise selection)

후방 소거법에서처럼 각 단계별로 통계적으로 유의하지않은 변수들을 제거하는 것만 제외하면

전방 선택방법과 동일.

그 외에 주성분 분석이나 회귀나무 사용해서 예측변수의 수를 줄일 수 있음.