2.4 Exercises

텍스트, 스크린샷, 컴퓨터이(가) 표시된 사진

자동 생성된 설명텍스트, 메뉴, 문서이(가) 표시된 사진

자동 생성된 설명

3장 approach

선형 회귀에 대한 얘기

제품 판매량이 TV, 라디오, 신문 매체들의 광고비에 따라 어떻게 변하는지 설명

선형회귀에 빗대어서.

3.1

X 변수를 바탕으로, Y를 예측.

선형 관계가 있다고 하면,

 -> 이렇게 표현 가능.

B\_0는 절편, B\_1은 기울기 뜻함.

이걸 어떤 실제 데이터에 대입하여 생각해보면,

 이렇게 구성할 수 있고, sales가 판매량, 이게 Y값. TV가 광고비. 이게 X로 들어가고,

여기서 B\_0, B\_1은 파라미터라고도 부름.

3.1.1

 여기서 B^들이 의미하는게 예측 값인데, 이걸 잘 맞춰서 실제 Y값 잘 맞추는게 관건.

폰트, 타이포그래피, 텍스트, 화이트이(가) 표시된 사진

자동 생성된 설명 -> 이게 잔차 e\_i = y\_i – y^\_i로, 실제 값이랑 예측값 차이 나타냄. 이걸 다 더한게 RSS # RSS 최소화 해야함.

폰트, 텍스트, 친필, 화이트이(가) 표시된 사진

자동 생성된 설명이거 최소화 하는 B^\_0, B^\_1 찾는게 이거.

3.1.2

편향, 표준오차 설명 가능.

모집단의 평균에 추정된 평균이 가까워 지면 편향이 없다는 것.

폰트, 텍스트, 라인, 화이트이(가) 표시된 사진

자동 생성된 설명

각각 표준오차(standard error) 척도. 실제 값이랑 추정 값이 얼마나 벗어났는지 볼 수 있는 거.

B^\_0는 절편 표준오차, B^\_1은 기울기 표준오차 나타내는 값. #시그마 제곱 봐볼 것.

 => 신뢰구간 95% 의미. 이게 표준오차 2배 만큼 떨어진 범위 내에 실제값이 존재할 가능성이 95% 이다는 의미.

폰트, 텍스트, 스크린샷, 디자인이(가) 표시된 사진

자동 생성된 설명 -> 귀무가설 내용(처음부터 버릴 것)

여기서 B\_1 = 0이 X랑 Y의 관계가 아무것도 없음을

B\_1 =/ 0 이 X랑 Y의 관계가 있다. # 이게 귀무가설 기각할 수 있음 의미.

3.1.3

RSE, 결정계수(R^2) 얘기

RSE (Residual Standard Error) 잔차 표준 오차:

폰트, 텍스트, 라인, 도표이(가) 표시된 사진

자동 생성된 설명

모델이 데이터에 얼마나 잘 맞는지 나타내는 척도.

Y^에서 일정한 오차 값 갖는 다는 것을 의미.

예측 오차 무슨 퍼센트 형태로 나타나고.

R^2:

회귀 모형 적합도 나타내는 척도.

항상 0~1 사이에.

1에 가까울수록 데이터 잘 설명한 다는 것.

텍스트, 폰트, 라인, 스크린샷이(가) 표시된 사진

자동 생성된 설명

TSS(총 제곱합) 출력 변수 Y의 총 변동성 나타냄.

RSS는 설명되지 않은 변동성.

만약 R^2값이 0.61이 나오면, 61% 변동을 설명할 수 있다는 것 의미.

텍스트, 폰트, 라인, 스크린샷이(가) 표시된 사진

자동 생성된 설명

R^2가 상관계수 제곱과 동일한데, 상관계수는 두 변수간 선형 관계 나타내는 척도.

다중회귀 에서는 R^2을 많이 사용한다고 함.

3.2

다중선형회귀 얘기

단순 선형 회귀는 여러 개의 예측 변수를 한꺼번에 고려할 수 없다.

그래서 다중 선형회귀가 나온거고,



이런식으로, 여러 별도의 기울기 B\_p 두어서 여러 개 예측 변수를 하나의 식에 때려박아서 모델화 한다.

3.2.1

최종 Prediction 값은 당연히 다중 선형 회귀라서

폰트, 텍스트, 타이포그래피, 화이트이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 화이트, 라인이(가) 표시된 사진

자동 생성된 설명

RSS 값도 위의 단순 선형 회귀랑 같은데, 파라미터 값 추가됨.

라인, 스크린샷이(가) 표시된 사진

자동 생성된 설명 다중 선형 회귀는 이런식으로 3차원으로 예측 한다는 걸 시각화 한 것.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

이 표를 위의 다중회귀 식에 넣으면,

Sales = 2.939 + 0.046 \* TV + 0.198 \* radio – 0.001 \* newpaper로,

단순 회귀는 newspaper가 큰 영향을 끼쳤지만, 여기 다중 회귀에서는 거의 영향이 없다.

그래서 상관 관계를 봐봐야 한다.

3.2.2

## 여기서부터 순차적 접근.

예측변수(Predictors)가 응답변수(response)와 관련 있는가? :

다중 선형 회귀에서는 이걸 보기 위해서 계수들이 0인지 먼저 봐야함. #여기서는 Hypothesis(검정)한다라고 표현.

이 식이 나왔는데, 이게 계수들이 0인지 먼저 본다는 것.

예측 변수가 응답 변수와 관련이 없다 라는 것을 의미.

 계수 중 0이 아닌 값 있다는 것 의미.

이걸 측정 하는 것은,

폰트, 텍스트, 라인, 화이트이(가) 표시된 사진

자동 생성된 설명이 F 통계량을 봐야 하는데,

이 F 통계량이 1에 가까우면 H\_0가 참일 가능성이 높지만,

1보다 크면 H\_a 채택 가능. 그 말이 계수 중에 0 아닌 값이 있다는 것 의미.

중요 변수 결정: (모든 변수가 쓸모 있는지, 아님 일부만 쓸모 있는지 내용)

F-통계량 보고, H\_0가 아니라면, 각 변수가 응답 변수(Y)에 얼만큼의 영향 미치는지 판단해야 함.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

이를 위해서 각 변수값들 t-검정 하고, p-value 통해서 변수 쓸 모 있는지 보면 되는데,

p-value값 낮으면 그 변수가 응답 변수와 관련 있다는 것 의미.

# 근데 이것만 보면 변수들 많을 때, 좀 보기 어려움. 그래서 전체 F-통계량 보는게 좋다고 말함.

모든 예측 변수가 응답 변수랑 관련 있는 건 드물어서

변수 selection 하는게 필요함.

그 방법.

1. Forward selection(전진 선택법): 변수 하나씩 넣으면서 RSS 최소화 하는 변수 선택
2. Backward selection(후진 제거법): 모든 변수 포함해서 넣어보고, p-value값 가장 큰 변수 제거
3. Mixed selection(혼합 선택법): 둘다 썪은 방식. 추가된 변수가 의미 없게 되면 제거.

Mixed selection 방식: 초기 모델은 아무 변수도 포함하지 않은 null model(기본 모델, 절편값만 들어있는 모델)로 시작.

그 다음 forward selection 방식 적용 🡺 p개의 모든 예측 변수를 고려해서 각 변수 하나씩 추가해 봄. 이 추가한 변수 중에서 RSS값 가장 많이 감소시키는 변수 모델에 추가하고,

그 다음 backward selection 방식 적용 🡺 변수 추가되면, 기존 변수의 P-value 값 확인. 그리고 이 p-value 값이 정해놓은 임계값(ex.0.5) 보다 크면 제거.

이걸 계속 반복.

모델 적합도(모델이 데이터를 얼마나 잘 설명하는지?):

R^2값이랑 RSE값 확인하면 됨.

R^2는 모델이 응답변수의 변동성 얼마나 잘 설명하는지. 그래서 이게 1에 가까우면 모델이 데이터 잘 설명한다는 것 의미.

RSE는 예측값과 실제값의 차이. 값 작을수록 정확도 높음 의미.

예측(예측값은 무엇이고, 예측이 얼마나 정확한가?):

신뢰구간(confidence interval), 예측 구간(prediction interval) 보면 알 수 있음.

신뢰구간은 예측된 ‘평균’ response값의 불확실성 나타냄. 실험 여러 번 해서 실제값이 이 구간에 포함될 확률 나타냄.

예측구간은 ‘개별’ 관측치(데이터)에 대한 불확실성. 신뢰구간보다 항상 넓다. 예측 오차랑 개별 데이터의 변동성까지 같이 고려해서.

3.3

3.3.1

질적변수(qualitative predictors): 숫자 아닌 것.(집 소유 여부, 학생 여부)

이를 사용하기 위해선 0 or 1로 변환(이를 더미 변수라고도)해서 써야 함.

이게 변수 두개인 경우,

변수 3개 넘어가면 n-1 개의 더미 변수 생성 필요.

이렇게 생성한 더미변수들이 쓸모 있는지 없는지는 p-value 측정해서 평가.

3.3.2

Additivity: 예측 변수 x\_j가 응답변수 Y에 미치는 영향이 독립적이라는 것 의미.

상호작용 효과: 한 x값이 다른 x값에 영향 끼치는 것.

## 현실 세계에서는 additivity 한것이 거의 없기 때문에 거의 대부분이 상호작용 효과를 가지고 있다.

 상호작용 항 추가한 것.

X\_1X\_2가 상호작용 항 의미.

이걸 실제 데이터에 적용해보면,



로 나오는데, R^2값 확인해보면, 상호작용항 있는게 값이 더 큼.

상호작용 항 있는게 모델 더 적합하다는 의미.

상호작용 항 껴 넣어주면, 이 변수도 같이 모델에 당연히 들어가야 하고,

Qualitative 변수(학생 여부)랑 Quantitative 변수(소득)랑 같이 썪어서 쓸 수 있음.

현실 세계에서는 보통 데이터가 비선형적으로 나타나서



와 같이, 다항회귀 모델을 쓴다.

3.3.3

# 선형 회귀에서의 문제 발생 시, 해결책.

비선형적 데이터:

텍스트, 도표, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명

Residual plot로 시각화 해서 봐보고, non linearity 하게 나오면

X\_j(예측변수)값을 로그,루트,제곱 해서 하면서 변환해서 쓰거나,

모델 자체를 그냥 다항회귀로 떄려 박음 됨.

오차항 correlation:

선형 회귀는 오차 € 가 상관관계가 없어야 함.

이거 자체도 학습 되어 있다고 생각하면 과적합이기 때문.

이때는 동일하게 residual plot 써서 직접 시각화 된거 봐보거나,

뭐 다른 모델을 갖다가 써라. 라고 대에충 애매 하게 말함.

오차 값의 분산이 일정하지 않다면?:

선형회귀 모델은 오차값의 분산이 일정해야 한다는 가정 가지고 있음.

근데 이 오차값들이 Y값에 따라 달라지는 경우 많음. 이때를 heteroscedasticity(이분산성)이라고 함.

텍스트, 도표, 라인, 지도이(가) 표시된 사진

자동 생성된 설명

이 문제 직면하면 이렇게 깔때기 모양처럼 나타남.

#뭐 위의 방식 써서 로그 씌우든 루트 씌우든 제곱 씌우든 해서 하면 좀 완화 됨.

이상치:

텍스트, 도표, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

이상치 볼 때 studentized residual 시각화 해서 본다는데, -3~3 범위 값에 안 들어가있는건 이상치로 간주.

Outlier를 제거하거나, 변수 추가해보거나, 모델 자체를 갈아 끼우는 방식 써서 해결.

High leverage(영향력) 점:

도표, 라인, 그래프, 스크린샷이(가) 표시된 사진

자동 생성된 설명

여기서 빨간 점들이 high leverage 점들인데, 위의 그래프처럼 시각화 해서 직접 보고, 이 점 자체를 제거하는 방식 사용해라.

Collinearity(다중공선성):

X\_j값들 간에 매우 강한 상관관계 있을 때 발생하는 문제.

이는 모델 정확도 저하 시키고,

이 문제 발생했다 하면,(뭐 상관관계나, VIF 봐보고 판단)

변수 자체를 없애거나, collinearity 한게 보이는 변수들 결합해서 새로운 파생 변수 만들거나,

하면 됨.

텍스트, 폰트, 라인, 번호이(가) 표시된 사진

자동 생성된 설명이게 VIF 식인데, 이 값이 5나 10 값 넘어가면 다중공선성 문제 해결해야 함.