강의 계획 및 평가 방법

주차	날짜	강의 내용	과제	대면/비대면	평가
1	03/06	강의 소개		Online	
2	03/13	데이터 마이닝 절차	과제 1 (10%)	A704	
3	03/20	데이터 탐색 및 시각화		B224	
4	03/27	차원 축소		Online	
5	04/03	예측성능 평가		Online	
6	**04/10**	**다중 선형 회귀분석**		**A704**	
7	04/17	중간 프로젝트 발표		A704	30%
8	04/24	k-최근접이웃 알고리즘 나이브 베이즈 분류	과제 2 (10%)	Online	
9	05/01	분류와 회귀 나무		Online	
10	05/08	로지스틱 회귀분석		Online	
11	05/15	신경망		A704	
12	05/22	판별 분석		Online	
13	05/29	연관 규칙		Online	
14	06/05	군집 분석		A704	
15	06/12	기말 프로젝트 발표		A704	40%

Data Mining for Business Analytics

Ch. 06 Multiple Linear Regression

2023.04.10.

Contents

6.1 Introduction

Multiple Linear Regression

- 변수와 변수 사이의 관계를 알아보기 위한 통계적 분석방법
- 독립변수(independent variable)와 종속변수(dependent variable)간의 관계 규명/예측

- Y: 출력, 반응변수, 종속변수(outcome, response/dependent variable)
- $X_1, X_2, \cdots, X_p$: 독립변수, 입력변수, 회귀변수, 공변량(independent/input variable, regressors, covariates)
- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$
  - ✓ $\beta_0, \cdots, \beta_p$: 회귀계수(coefficients)
  - ✓ ϵ : 잡음(noise or unexplained part)

- 기업의 순이익, 내부유보(사내유보), 매출액, 부채비율을 이용한 해당 기업의 주가 예측
- 기업의 광고가 판매량에 미치는 영향
- 기계의 온도에 따른 플라스틱 제품의 견고도
- 독립변수 : 소득수준, 순이익, 광고, 온도
- 종속변수: 식료품비, 주가, 판매량, 플라스틱의 견고 정도

- 선형성: 응답 변수가 예측 변수와 선형 회귀 계수의 선형 조합으로 표현 가능함을 의미
- 다음과 같은 모형도 다중회귀모형에 속함에 유의
  - ✓ $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \epsilon$
  - ✓ $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$

6.2 Explanatory vs. Predictive Modeling

Explanatory(or Descriptive) Model

- 출력에 대한 입력의 평균 효과에 대한 설명 또는 수량화
- 결과: "다른 모든 요소($X_2, X_3, \cdots, X_p$)의 변화량을 고려하지 않았을 때, 서비스 속도(X_1)의 단위증가는 고개 만족도(Y)가 평균 5 포인트 상승하는 관계를 보여준다"
- 모집단에 있는 근본적인 관계를 알아보기 위해 데이터에 가장 적합한 모델 적용
- 전체 데이터셋 사용
- 초점
  - ✓ 데이터가 모델에 얼마나 잘 적합하는지(즉, 모델이 데이터의 실제 값과 얼마나 비슷한 값을 주는지)
  - ✓ 평균 관계가 얼마나 강한지에 초점
  - ✓ 회귀계수(β)

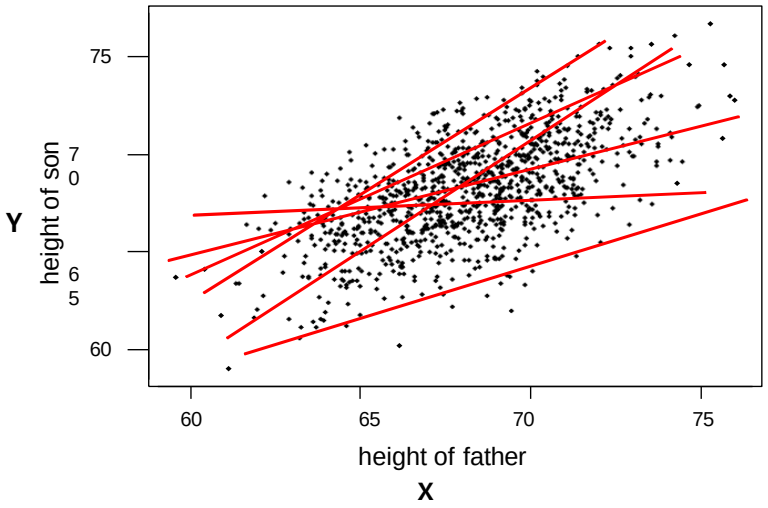
Predictive Model (Data Mining)

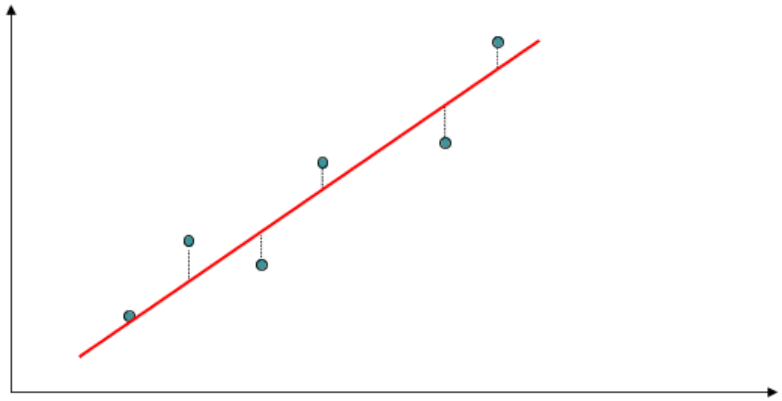
- 주어진 입력 값을 활용하여 새로운 레코드의 결과 값 예측
- 회귀계수나 평균 레코드를 다루지 않고 생성되는 모델을 활용하여 새로운 레코드 예측
- 새로운 개별 레코드에 대해 예측력이 가장 뛰어난 회귀모델 탐색
- 전체 데이터를 학습용 집합과 평가용 집합으로 분리. 학습용 집합은 모형을 추정하기 위해 사용되고 검증용 집합(validation set), 즉 예비용 집합(holdout set) 은 새로운 알려지지 않은 데이터에 대한 모형의 성과를 평가하기 위해 사용
- 초점
  - ✓ 모델이 새로운 사례를 얼마나 잘 예측하는지
  - ✓ 예측($\hat{Y}$)

6.3 Estimating the Regression Equation and Prediction

Estimating Method(모델 추정 기법)

- 최소제곱법(Least Square Estimates): 어떤 계의 해방정식을 근사적으로 구하는 방법으로, 근사적으로 구하려는 해와 실제 해의 오차의 제곱의 합(Sum of Squares)이 최소가 되는 해를 구하는 방법





Minimizes $SS = \sum {e_i}^2$

6.3 Estimating the Regression Equation and Prediction

Estimating Method(모델 추정 기법)

- 일반적으로 최소자승법(Ordinary Least Squares: OLS)을 이용하여 데이터로부터 계수를 추정
- 실제값(Y)과 모델에서의 예측을 기반으로 한 값($\hat{Y}$) 사이의 잔차제곱합(RSS: Residual Sum of Squares)을 최소화시키는 추정 값 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$를 찾음
- Minimize $\sum(Y_i - \hat{Y}_i)^2 \rightarrow$ Estimates $\hat{\beta}_i$
- 입력값: $x_1, x_2, \cdots, x_p$
- 예측값: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2, \cdots, \hat{\beta}_p x_p$

- 가장 좋은 예측인 이유
  - ✓ 예측값이 불편값(unbiased: 예측값의 평균값이 실제값과 동일)이고, 다른 불편 추정값들과 비교했을 때 평균제곱오차가 가장 작음
- 선형 회귀의 가정
  - ✓ 가정 1: 잡음 ε(또는 종속변수 Y)는 정규분포를 따른다.
  - ✓ 가정 2: 종속변수가 예측변수와 선형회귀 계수의 선형 조합으로 표현 가능하다. 파라미터에 대한 선형성만 가정한다. (linearity)
  - ✓ 가정 3: 사례들은 서로 독립적이다. 예측변수의 오차가 서로 무관함을 가정
  - ✓ 가정 4: 예측변수 집단이 주어진 상황에서 Y값의 변동성은 예측변수 값과 관계없이 일정하다. (등분산성: homoskedasticity).

6.3 Estimating the Regression Equation and Prediction

Example: Predicting the Price of Used Toyota Corolla Car

- 도요타 자동차 신규 구매자들에게 보상판매(trade-in)의 일환으로 중고차를 대리점이 구입해주는 선택권 제공
- 딜러는 대리점에서 중고차를 판매할 가격을 예측해야 함
- 데이터 개수: 1,000개 (학습: 60% / 검증: 40%)
- 범주형 변수 → 더미 변수
  - ✓ 연료 유형(Fuel Type) – 휘발유(Petrol), 경유(Diesel), 천연가스(CNG)
  - ✓ 휘발유(Fuel_Type_Petrol: 0/1), 경유(Fuel_Type_Diesel: 0/1)
  - ✓ 불필요한 변수(예. 천연가스)를 포함하면 이 변수가 다른 두 변수의 완벽한 선형 조합이 되므로 회귀모형이 작동하지 않을 수 있음

Price	Age	KM	Fuel_Type	HP	Metallic	Automatic	cc	Doors	Quarterly_Tax	Weight
13500	23	46986	Diesel	90	1	0	2000	3	210	1165
13750	23	72937	Diesel	90	1	0	2000	3	210	1165
13950	24	41711	Diesel	90	1	0	2000	3	210	1165
14950	26	48000	Diesel	90	0	0	2000	3	210	1165
13750	30	38500	Diesel	90	0	0	2000	3	210	1170
12950	32	61000	Diesel	90	0	0	2000	3	210	1170
16900	27	94612	Diesel	90	1	0	2000	3	210	1245

6.3 Estimating the Regression Equation and Prediction

[실습] Table 6.3, 6.4

Example: Predicting the Price of Used Toyota Corolla Car

모델 계수 추정 결과

intercept: 추정된 상수항

```
intercept   -1319.354380041219
```

	Predictor	coefficient
0	Age_08_04	-140.748761
1	KM	-0.017840
2	HP	36.103419
3	Met_Color	84.281830
4	Automatic	416.781954
5	CC	0.017737
6	Doors	-50.657863
7	Quarterly_Tax	13.625325
8	Weight	13.038711
9	Fuel_Type_Diesel	1066.464681
10	Fuel_Type_Petrol	2310.249543

Regression statistics

```
                    Mean Error (ME) : -0.0000
     Root Mean Squared Error (RMSE) : 1400.5823
          Mean Absolute Error (MAE) : 1046.9072
        Mean Percentage Error (MPE) : -1.0223
Mean Absolute Percentage Error (MAPE) : 9.2994
```

검증 셋의 20개 데이터에 대한 예측값

잔차: $Y_i - \hat{Y}_i$

	Predicted	Actual	Residual
507	10607.333940	11500	892.666060
818	9272.705792	8950	-322.705792
452	10617.947808	11450	832.052192
368	13600.396275	11450	-2150.396275
242	12396.694660	11950	-446.694660
929	9496.498212	9995	498.501788
262	12480.063217	13500	1019.936783
810	8834.146068	7950	-884.146068
318	12183.361282	9900	-2283.361282
49	19206.965683	21950	2743.034317
446	10987.498309	11950	962.501691
142	18501.527375	19950	1448.472625
968	9914.690947	9950	35.309053
345	13827.299932	14950	1122.700068
971	7966.732543	10495	2528.267457
133	17185.242041	15950	-1235.242041
104	19952.658062	19450	-502.658062
6	16570.609280	16900	329.390720
600	13739.409113	11250	-2489.409113
496	11267.513740	11750	482.486260

Regression statistics

```
                    Mean Error (ME) : 103.6803
     Root Mean Squared Error (RMSE) : 1312.8523
          Mean Absolute Error (MAE) : 1017.5972
        Mean Percentage Error (MPE) : -0.2633
Mean Absolute Percentage Error (MAPE) : 9.0111
```

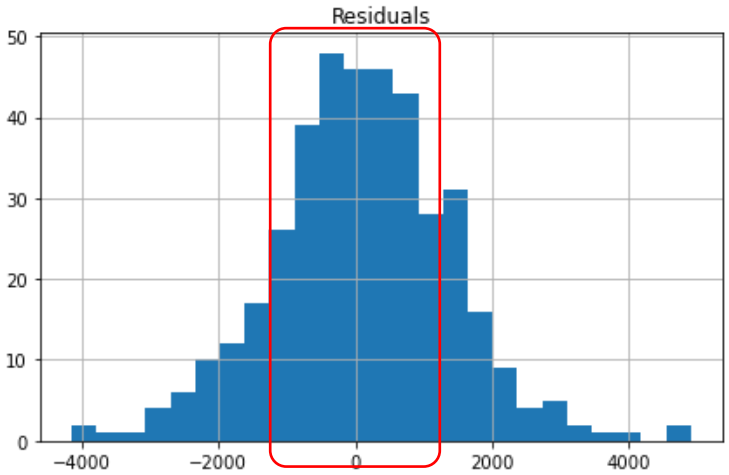
6.3 Estimating the Regression Equation and Prediction

[실습] Figure 6.1

Example: Predicting the Price of Used Toyota Corolla Car

잔차에 대한 히스토그램(검증용 데이터 셋)

- 선형회귀의 가정 1(잡음 ε(또는 종속변수 Y)는 정규분포를 따른다.)을 거의 만족함을 확인
- 대부분의 오차가 −$2,000에서 $2,000 사이에 있음. 특히 [−1406, 1406] → 74.25%
- 오차가 자동차 가격에 비하면 상대적으로 작은 금액이지만 이익 면에서는 고려해야 함



6.4 Variable Selection in Linear Regression

Reducing the Number of Predictors

- 예측변수들 전부를 수집하는 것이 실행 가능하지 않거나 비용이 너무 비쌀 수 있음
- 적은 수의 예측변수를 사용하면 더 정확한 측정을 수행할 수 있음(예를 들어, 설문조사)
- 예측변수가 많을수록 데이터에 결측값이 존재할 위험성이 높아짐
- 간결성(Parsimony)은 좋은 모델이 갖는 중요한 특징. 변수의 개수가 적은 모형에서 예측변수의 영향력을 더 잘 이해할 수 있음
- 많은 변수를 사용하는 모델에서는 다중공선성(Multicollinearity)으로 인해서 회귀계수의 추정치들이 불안정할 수 있음.
  - ✓ 다중공선성: 2개 이상의 예측변수가 종속변수에 동일한 선형관계를 공유
  - ✓ Rule of Thumb: 레코드 개수 ≥ 5(p+2) (p: 예측변수의 개수)

- 다중공선성(Multicollinearity)
  - ✓ 독립변수들 사이에 상관관계를 갖고 있는 현상
  - ✓ 독립변수들끼리 높은 선형관계가 존재하면 추정된 회귀 계수의 큰 오차 유발
  - ✓ 다른 예측 변수들을 고정한 가운데 특정 예측 변수가 한 단위 증가할 때 종속 변수의 기대치의 변화를 측정하는 것이 회귀 계수의 일반적인 해석이지만, 다중공선성이 존재한다면 다른 예측 변수들을 고정한 채, 특정 예측 변수만을 한 단위 증가시키는 것은 불가능
  - ✓ VIF(Variance Inflation Factor, 분산팽창지수)를 통해 확인 가능

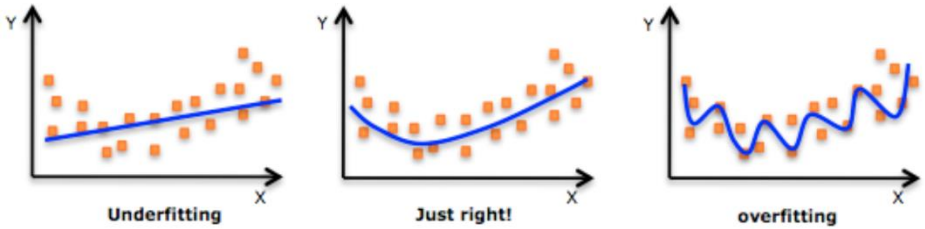
6.4 Variable Selection in Linear Regression

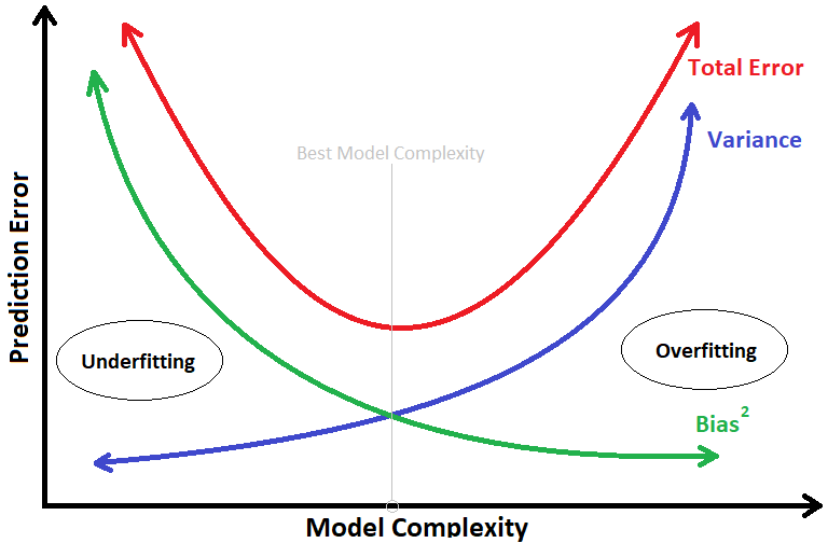
Reducing the Number of Predictors

Bias-Variance Trade-off

1. 종속변수와 실제 상관관계가 있는 예측변수를 누락시킬 경우(예측변수를 적게 사용하면), 예측 오차의 평균 (또는 bias) 증가
2. 종속변수와 상관관계가 없는 예측변수를 사용하는 경우(예측변수를 많이 사용하면), 예측 값 자체의 분산(variance) 증가

- 일반적으로 약간의 편향(bias)을 허용하여 예측의 분산(variance)을 줄이는 방법 사용
- 예측변수가 많은 경우, 잡음의 표준편차에 비해 상대적으로 작은 회귀계수를 갖는 변수들이 모델에 존재하고, 이들은 다른 변수들과도 어느 정도 이상의 상관관계를 보일 수 있는 가능성이 높음 → 이와 같은 변수들을 제거하면 예측 분산을 줄일 수 있기 때문에 예측 성능이 향상됨





6.4 Variable Selection in Linear Regression

How to reduce the number of predictors

- Domain Knowledge 활용: 예측변수들이 무엇을 측정하고 있는지, 왜 이 변수들이 종속변수의 반응 예측에 적절한지 등
- 계산력과 통계적 유의성 활용:
  - ✓ 전역탐색(Exhaustive search): 예측변수의 개수에 penalty 부여
  - ✓ 대표적인 부분집합 선택(Popular subset selection): 예측변수의 개수가 많을 때 사용

Exhaustive Search (전역 탐색)

- 모든 예측 변수들의 부분집합을 평가
- 적당한 p값(예측변수의 수)을 찾기 위한 부분집합의 수가 매우 많기 때문에 가장 가능성이 높은 부분집합을 조사하고 이를 기준으로 예측변수를 선택하는 방법들이 필요
- 과소적합 모델과 과적합 모델의 평가, 비교 기준: 학습용 데이터를 적합시키는 정도.

1. 수정 결정계수(R^2_{adj}, adjusted R^2)
2. Akaike Information Criterion(AIC), Schwartz's Bayesian Information Criterion(BIC)

6.4 Variable Selection in Linear Regression

How to reduce the number of predictors

Exhaustive Search

□ 결정계수(R^2)

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

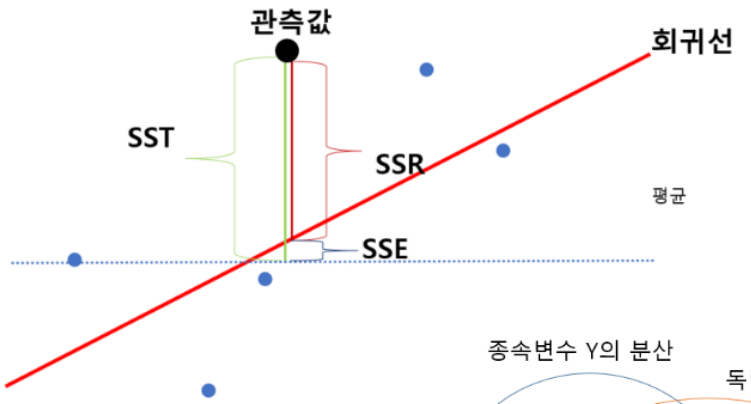
$$SST(Total\ Sum\ of\ Squares) = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

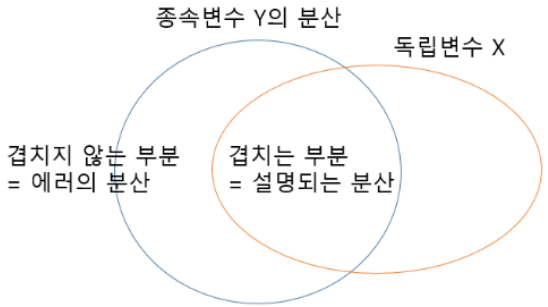
$$SSE(Explained\ Sum\ of\ Square) = \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2$$

$$SSR(Resudual\ Sum\ of\ Square) = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

- 결정계수 or 설명력(R^2): 모델에서 독립변수가 종속변수를 얼마나 잘 설명해주는지를 가리키는 지표
- 독립변수의 개수가 증가하면 일방적으로 증가

- SST: 총 제곱합(Y의 분산)
  - ✓ (관측값 – 관측값의 평균)의 제곱의 총합
- SSE: 회귀제곱합(설명되는 분산)
  - ✓ (추정값 – 관측값의 평균)의 제곱의 총합
- SSR: 잔차제곱합(에러의 분산)
  - ✓ (관측값 – 추정값)의 제곱의 총합





6.4 Variable Selection in Linear Regression

How to reduce the number of predictors

Exhaustive Search

☐ 수정 결정계수(R^2_{adj}, adjusted R^2)

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

예측변수의 개수에 대한 벌점

- 결정계수(R^2): 모델에서 설명할 수 있는 변동성의 비율. (단일 예측변수를 갖는 모델에서는 상관계수의 제곱)
- 결정계수(R^2)와 수정 결정계수(R^2_{adj}, adjusted R^2) 모두: 높은 값 → 보다 나은 적합성
- 결정계수(R^2): 예측변수의 개수를 고려하지 않음. 정보량의 증가가 아니라 단순히 예측변수의 개수만을 증가시켜도 값이 증가함
- 수정 결정계수(R^2_{adj}, adjusted R^2): 예측변수의 개수에 대한 벌점(penalty) 반영.
- 수정 결정계수를 사용한 부분집합 선택 = 학습 데이터에 대한 RMSE를 최소화하는 부분집합 선택

6.4 Variable Selection in Linear Regression

How to reduce the number of predictors

Exhaustive Search

☐ Akaike Information Criterion(AIC), Schwartz's Bayesian Information Criterion(BIC)

$$AIC = n \ln \frac{SSE}{n} + n(1 + \ln(2\pi)) + 2(p + 1)$$

$$BIC = n \ln \frac{SSE}{n} + n(1 + \ln(2\pi)) + \ln(n)\,(p + 1)$$

SSE(Sum of Squared Error) : 오차제곱합

- 모델의 적합도(the goodness of fit) 측정
- 모델에 있는 변수의 개수에 대한 벌점(penalty) 포함
- 정보 이론에 기초한 예측 에러에 대한 추정값
- 작은 값 → 보다 나은 적합성

6.4 Variable Selection in Linear Regression

[실습] Table 6.5

How to reduce the number of predictors

Exhaustive Search

- 부분집합의 크기가 고정되어 있다면, R^2, R^2_{adj}, AIC, BIC 모두 동일한 부분집합 선택
- 서로 다른 개수의 예측변수를 사용하는 모델들을 평가하는 경우 차이가 있음

- ex) Toyota Corolla price data에 대한 전역탐색 결과(11개 변수)
- 변수를 증가시키면서 최적의 모델 제시
- R^2_{adj} 값: 8개의 예측변수가 사용될 때까지 증가한 후 감소
- AIC: 8개의 예측변수를 사용한 모델이 최적
- 중요한 변수: Age, HP, Weight, KM (True or False: 예측변수에 포함 여부)

n	r2adj	AIC	Age_08_04	Automatic	CC	Doors	Fuel_Type_Diesel	Fuel_Type_Petrol	HP	KM	Met_Color	Quarterly_Tax	Weight
1	0.767901	10689.712094	True	False	False	False	False	False	False	False	False	False	False
2	0.801160	10597.910645	True	False	False	False	False	False	True	False	False	False	False
3	0.829659	10506.084235	True	False	False	False	False	False	True	False	False	False	True
4	0.846357	10445.174820	True	False	False	False	False	False	True	True	False	False	True
5	0.849044	10435.578836	True	False	False	False	False	False	True	True	False	True	True
6	0.853172	10419.932278	True	False	False	False	False	True	True	True	False	True	True
7	0.853860	10418.104025	True	False	False	False	True	True	True	True	False	True	True
8	0.854297	10417.290103	True	True	False	False	True	True	True	True	False	True	True
9	0.854172	10418.789079	True	True	False	True	True	True	True	True	False	True	True
10	0.854036	10420.330800	True	True	False	True	True	True	True	True	False	True	True
11	0.853796	10422.298278	True	True	True	True	True	True	True	True	True	True	True

6.4 Variable Selection in Linear Regression

How to reduce the number of predictors

Popular Subset Selection Algorithms

- 모든 가능한 회귀모델로 이루어진 공간에 대해 부분적이며 반복적으로 예측변수군을 탐색
- 비록 다양한 크기의 예측변수 부분집합에 대하여 최적에 가까운 선택을 찾아내는 방법이 존재하지만 최종적인 결과물은 하나의 최적 예측변수 부분집합
- 계산적으로는 간단하지만, 좋은 성과를 나타내는 예측변수들의 조합을 누락시킬 가능성이 있음. 어떠한 방법도 일정 기준에 대하여(예를 들어 수정 결정계수(R^2_{adj})와 같은 기준) 최적의 예측변수 부분집합을 선택한다고 보장하지 않음
- 예측변수가 많을 때 적합한 방법, 예측변수의 개수가 보통 수준인 경우 전역탐색이 더 좋음

6.4 Variable Selection in Linear Regression

How to reduce the number of predictors

Popular Subset Selection Algorithms

☐ 전진선택(Forward Selection)

- 예측변수가 없는 상태에서 예측변수를 하나씩 추가하는 방법
- 추가되는 예측변수: 모델의 R^2 증가에 가장 크게 기여하는 변수
- 추가되는 예측변수의 기여도가 통계적으로 유의하지 않을 때 중단
- 단점: 함께 사용될 때는 효과적이지만 각각 단일변수로 사용될 때는 낮은 성능을 보이는 예측변수들을 누락시킬 수 있음

☐ 후진제거(Backward Elimination)

- 처음에는 모든 예측변수들을 사용하여 시작, 단계별로 가장 유용하지 않은(통계적 유의성을 따라) 예측변수들을 제거
- 모든 남아있는 예측변수들이 유의미한 기여도를 가질 때 중단
- 단점: 모든 예측변수들을 포함하는 초기 모델 계산에 시간이 많이 소요되고 불안정

☐ 단계적 선태(Stepwise Selection)

- 전진선택법과 후진제거법을 복합적으로 사용하는 방법. 첫 번째 단계에서는 전진선택법을 수행
- 전진선택법에 의하여 변수를 추가하면서 새롭게 추가된 변수에 기인하여 기존 변수가 그 중요도가 약화되어 제거될 수 있는 지를 매 단계별로 검토하여 해당변수를 제거
- 추가 또는 제거되는 변수가 더 이상 없을 때 중단

6.4 Variable Selection in Linear Regression

[실습] Table 6.6, 6.7

How to reduce the number of predictors

Popular Subset Selection Algorithms

Ex) Toyota Corolla 학습용 데이터 셋

☐ 전진선택(Forward Selection) / 단계적 선태(Stepwise Selection) → 동일한 결과 (8개 예측변수 선택)

```
Variables: Age_08_04, KM, HP, Met_Color, Automatic, CC, Doors, Quarterly_Tax, Weight, Fuel_Type_Diesel, Fuel_Type_Petrol
Start: score=11565.07, constant
Step: score=10689.71, add Age_08_04
Step: score=10597.91, add HP
Step: score=10506.08, add Weight
Step: score=10445.17, add KM
Step: score=10435.58, add Quarterly_Tax
Step: score=10419.93, add Fuel_Type_Petrol
Step: score=10418.10, add Fuel_Type_Diesel
Step: score=10417.29, add Automatic
Step: score=10417.29, add None
['Age_08_04', 'HP', 'Weight', 'KM', 'Quarterly_Tax', 'Fuel_Type_Petrol', 'Fuel_Type_Diesel', 'Automatic']
```

☐ 후진제거(Backward Elimination) → 8개 예측변수 선택

```
Variables: Age_08_04, KM, HP, Met_Color, Automatic, CC, Doors, Quarterly_Tax, Weight, Fuel_Type_Diesel, Fuel_Type_Petrol
Start: score=10422.30
Step: score=10420.33, remove CC
Step: score=10418.79, remove Met_Color
Step: score=10417.29, remove Doors
Step: score=10417.29, remove None
['Age_08_04', 'KM', 'HP', 'Automatic', 'Quarterly_Tax', 'Weight', 'Fuel_Type_Diesel', 'Fuel_Type_Petrol']
```

score: AIC Score 사용

6.4 Variable Selection in Linear Regression

[실습] Table 6.4, 6.6

How to reduce the number of predictors

Popular Subset Selection Algorithms

Ex) Toyota Corolla 검증용 데이터 셋

☐ 후진제거(Backward Elimination) (8개 예측변수)

Regression statistics

Mean Error (ME)	103.3045
Root Mean Squared Error (RMSE)	1314.4844
Mean Absolute Error (MAE)	1016.8875
Mean Percentage Error (MPE)	-0.2700
Mean Absolute Percentage Error (MAPE)	8.9984

☐ 초기 모델 (10개 예측변수)

Regression statistics

Mean Error (ME)	103.6803
Root Mean Squared Error (RMSE)	1312.8523
Mean Absolute Error (MAE)	1017.5972
Mean Percentage Error (MPE)	-0.2633
Mean Absolute Percentage Error (MAPE)	9.0111

- 후진제거 방법(8개 예측변수)의 성능과 초기 모델(10개 예측변수)의 성능에 거의 차이가 없음
- 간결성(Parsimony)이라는 측면에서 후진제거 방법이 선호됨

6.4 Variable Selection in Linear Regression

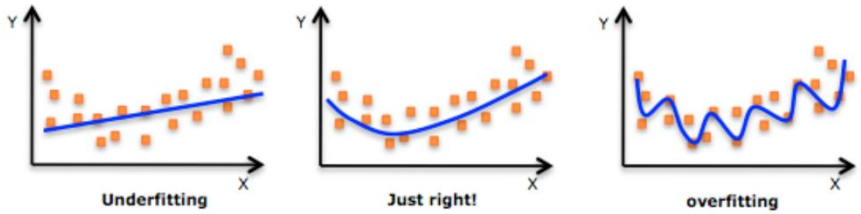
Regularization (Shrinkage Models)

- 예측변수들의 부분집합을 선택하는 것 = 모델의 일부 계수(가중치)를 0으로 만드는 것과 동일
- Regularization(규제) or Shrinkage(수축): 모델의 계수(가중치)들을 0 쪽으로 줄어들게(shrink) 하는 방법
  - ✓ 수정 결정계수(R^2_{adj}): 예측변수의 개수에 페널티를 부여함
  - ✓ Shrinkage: 계수(가중치)들의 절대값의 합에 페널티를 부여함(일반적으로 예측변수들은 우선 동일한 스케일로 정규화 됨)
- 계수들의 규모를 제한하는 이유: 높은 상관관계를 보이는 예측변수들이 높은 오차를 갖는 계수들을 보이는 경향이 있음 → 학습용 데이터에서의 작은 변화가 상관관계를 갖는 변수들을 강화하는 방향으로 급격한 이동을 초래할 수 있음 → 계수들의 전체적인 크기의 제한은 이러한 변동성을 감소시킴
- 일반적인 선형 회귀: Minimize $\sum(Y_i - \widehat{Y_i})^2$ (RSS: 잔차제곱합) → Estimates $\widehat{\beta_i}$
- Shrinkage: Minimize RSS + α Penalty → Estimates $\widehat{\beta_i}$
- 릿지 회귀(Ridge regression) / 라쏘 회귀(Lasso regression)

6.4 Variable Selection in Linear Regression

Regularization (Shrinkage Models)

- bias-variance trade-off 활용
- α가 증가할 때 ridge regression 적합의 유연성은 감소하게 되어 분산은 감소하지만 편향은 증가
- α=0을 갖는 ridge regression에 대응하는 최소제곱 계수 추정치에서 분산은 크지만 편향은 없음. 그러나 α가 증가할 때, ridge 계수 추정치의 축소는 편향이 약간 증가한 것에 비해 예측 분산에서 상당한 감소를 유도
- MSE는 α가 0에서 10까지 증가하면 상당히 감소. 이점을 넘어서면 α 증가에 따른 분산 감소는 느려지고 계수들의 축소는 현저하게 과소추정되어 결과적으로 편향에서 큰 증가를 가져옴



✓ 편향(bias): 예측 오차의 평균
✓ 분산(variance): 예측 값 자체의 분산

☐ 릿지 회귀(Ridge regression)

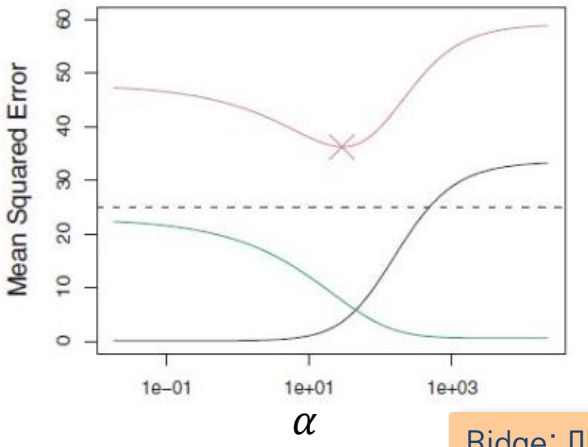
Minimize: RSS + $\alpha \sum_{j=1}^{p} \beta_j^2$ (L2 penalty)

- 페널티: 계수의 제곱합(L2 penalty)에 기초

☐ 라쏘 회귀(Lasso regression)

Minimize: RSS + $\alpha \sum_{j=1}^{p} |\beta_j|$ (L1 penalty)

- 페널티: 계수의 절대값합(L1 penalty)에 기초



MSE(분산 + 편향의 제곱의 함수)

편향(Bias)

예측의 분산(Variance)

Ridge: 마루 / Lasso: 올가미

6.4 Variable Selection in Linear Regression

Regularization (Shrinkage Models)

Ridge: 마루 / Lasso: 올가미

☐ 릿지 회귀(Ridge regression)

Minimize: RSS + $\alpha \sum_{j=1}^{p} \beta_j^2$ (L2 penalty)

- 변수가 많고 계수의 크기가 거의 동일한 크기일 때 성능이 좋음 → 높은 분산을 가지는 상황에서 가장 잘 작동하기 때문
- 최소 제곱 추정치가 아주 높은 분산을 가질 때, 편향의 작은 비용의 증가에서 분산에서의 축소를 이끌어, 결과적으로 좀 더 정확한 예측을 생성

☐ 라쏘 회귀(Lasso regression)

Minimize: RSS + $\alpha \sum_{j=1}^{p} |\beta_j|$ (L1 penalty)

- 적은 수의 설명변수가 상당히 큰 계수를 가질 때 잘 작동
- 일부의 설명변수만 포함하므로 단순하고 해석력 높은 모델을 만든다. → 반응변수와 관련있는 설명변수는 신호이고 변수는 잡음이 된다.
- 변수 선택을 수행하므로 모형을 해석하기가 쉽다.

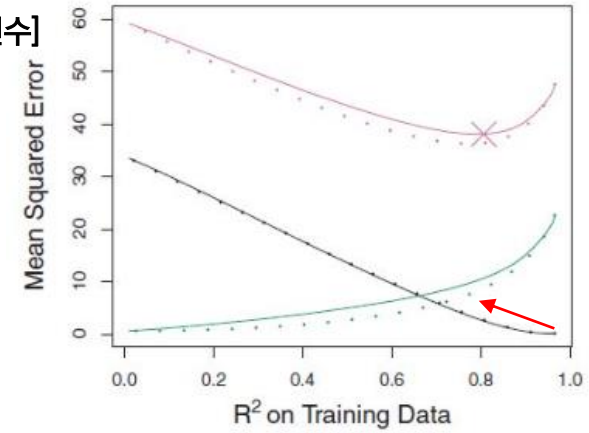
Ridge: ·····
Lasso: ——

MSE(분산 + 편향의 제곱의 함수)

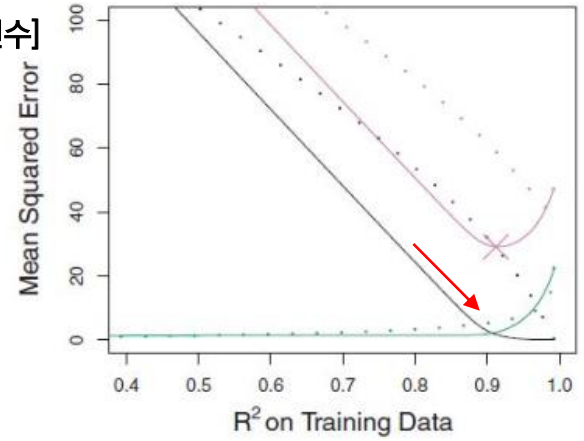
편향(Bias)

예측의 분산(Variance)

[45개 변수]



[2개 변수]



6.4 Variable Selection in Linear Regression

[실습] Table XX

Regularization (Shrinkage Models)

☐ Lasso CV / Ridge CV

- Cross-validation 방법을 사용하여 페널티 패러미터 α 의 최적값을 자동으로 선택함

☐ Bayesian Ridge

- 순차적인 접근방법(iterative approach)을 사용하여 전체 학습용 집합에서 페널티 패러미터 α 를 만들어 냄

✓ 11개 변수 중 6개 사용, 초기 모델과 비교하여 성능 차이 없음

[Normal Linear Regression]

Regression statistics

Mean Error (ME)	103.6803
Root Mean Squared Error (RMSE)	1312.8523
Mean Absolute Error (MAE)	1017.5972
Mean Percentage Error (MPE)	-0.2633
Mean Absolute Percentage Error (MAPE)	9.0111

[Lasso] Regression statistics

Mean Error (ME)	120.6311
Root Mean Squared Error (RMSE)	1332.2752
Mean Absolute Error (MAE)	1021.5286
Mean Percentage Error (MPE)	-0.2364
Mean Absolute Percentage Error (MAPE)	9.0115

[Lasso CV] Regression statistics

Mean Error (ME)	145.1571
Root Mean Squared Error (RMSE)	1397.9428
Mean Absolute Error (MAE)	1052.4649
Mean Percentage Error (MPE)	-0.2966
Mean Absolute Percentage Error (MAPE)	9.2918

Lasso-CV chosen regularization: 3.5138446691310588

```
[-1.40370575e+02 -1.76669006e-02  3.38674037e+01  0.00000000e+00
  6.94393427e+01  0.00000000e+00  0.00000000e+00  2.70913468e+00
  1.24342596e+01 -0.00000000e+00  0.00000000e+00]
```

[Ridge] Regression statistics

Mean Error (ME)	154.3286
Root Mean Squared Error (RMSE)	1879.7426
Mean Absolute Error (MAE)	1353.2735
Mean Percentage Error (MPE)	-2.3897
Mean Absolute Percentage Error (MAPE)	11.1309

[Bayesian Ridge] Regression statistics

Mean Error (ME)	105.5382
Root Mean Squared Error (RMSE)	1313.0217
Mean Absolute Error (MAE)	1017.2356
Mean Percentage Error (MPE)	-0.2703
Mean Absolute Percentage Error (MAPE)	9.0012

Bayesian ridge chosen regularization: 0.004622833439968832

✓ 매우 작은 regulation parameter → 이 데이터 셋은 규제의 효과가 거의 없음

6.4 Variable Selection in Linear Regression

[실습] Table XX

Regularization (Shrinkage Models)

Coefficients by regularization methods

	features	linear regression	lassoCV	bayesianRidge
0	Age_08_04	-140.748761	-140.370575	-139.754059
1	KM	-0.017840	-0.017667	-0.018131
2	HP	36.103419	33.867404	35.856074
3	Met_Color	84.281830	0.000000	85.088966
4	Automatic	416.781954	69.439343	408.599781
5	CC	0.017737	0.000000	0.020405
6	Doors	-50.657863	0.000000	-47.917629
7	Quarterly_Tax	13.625325	2.709135	13.269979
8	Weight	13.038711	12.434260	13.114412
9	Fuel_Type_Diesel	1066.464681	-0.000000	955.581484
10	Fuel_Type_Petrol	2310.249543	0.000000	2162.115763

✓ Lasso CV: 11개 변수 중 6개 사용

✓ 매우 작은 regulation parameter
→ 이 데이터 셋은 규제의 효과가 거의 없음

6.4 Variable Selection in Linear Regression

Which is the best? → Case by case

- 예측변수들의 부분집합을 선택하는 방법론 → “좋은 모델”일지 모르는 후보 모델을 제시
- “Best” 모델이 정말 최적의 모델을 담보하지는 않음
- “Best” 모델은 아직도 불충분한 예측 성능을 가지고 있을 수 있음
- 후보 모델들을 계속 테스트하고 예측 성능을 평가해야 함