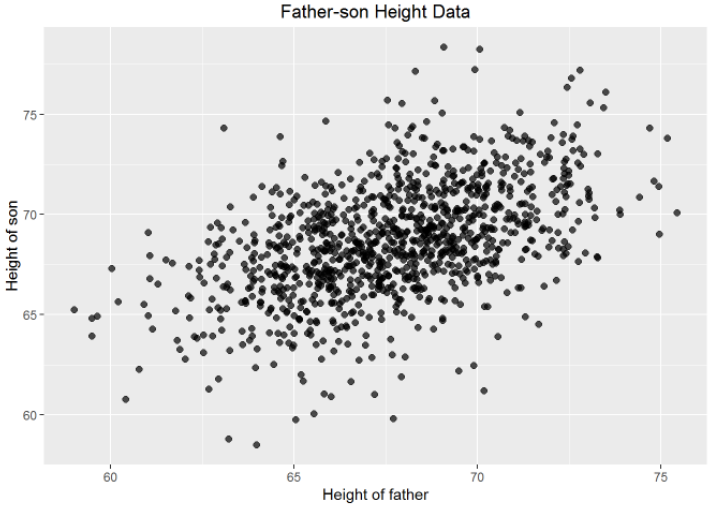
Regression (회귀)

Prof. Hyerim Bae
Department of Industrial Engineering, Pusan National University
hrbae@pusan.ac.kr

우리 아들의 키가 얼마나 클까?

- By Francis Galton



- By Orley Ashenfelter

- 품질 = 12.145
  + 0.00117 * 겨울 강우량
  + 0.06140 * 생장기 평균 기온
  − 0.00386 * 추수기 강우량

Contents

산업현장에서 수집된 데이터를 분석하는데 필요한 기초 소양을 강의합니다.

01 Linear Regression

02 Logistic Regression

Regression

변수의 개념

- 변수의 사전적 정의
  - 각 측정 단위에 대해서 측정하려고 하는 특성

- 질적 변수(Qualitative variable)
  - 변수의 값이 특정 카테고리를 포함하는 변수
  - 예) 사람의 성별, 자동차의 제조사 등

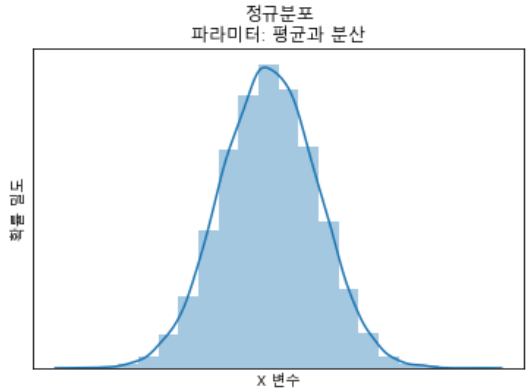
- 양적 변수(Quantitative variable)
  - 변수의 값을 연속형 수치로 표현가능한 변수
  - 예) 사람의 몸무게, 자동차의 가격 등

파라미터의 개념

- 파라미터의 사전적 정의
  - 수학적: 함수의 값을 결정짓는 변수
  - 통계적: 분포의 속성을 결정짓는 변수(모 평균, 모 분산 등)

파라미터

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$



용어정리, 변수 개념, 파라미터 개념

항목	설명
독립변수	다른 변수에게 영향을 주는 변수(x)
종속변수	다른 변수에게 영향을 받는 변수(y)
회귀계수	절편과 기울기를 의미함 β_0(절편)+β_1(x_1의 계수)
회귀방정식	회귀 계수를 이용하여 생성된 방정식
회귀 선	독립변수와 종속변수에 대한 분포를 나타내는 직선중 가장 적합한 직선
적합된 값	독립 변수에 대한 예측된 y값
잔차(Residuals)	예측값과 실제값의 차이
잔차 제곱 합	잔차의 제곱 합

변수 정규화 및 표준화

- 정규화란?
  - 데이터프레임 변수간의 크기가 심하게 차이나는 경우 문제 발생
    - 변수간의 크기를 동등한 정도로 맞추는 과정

- 최소 최대 정규화(Min-max normalization)

$$X' = \frac{X - MIN}{MAX - MIN}$$

- 표준 정규분포 데이터 표준화(Z score normalization)

$$X' = \frac{X - \text{평균}}{\text{표준편차}}$$

설명 모델(Explanatory model) vs. 예측 모델(Predictive model)

- 목표
  - Good EM: 모델이 데이터를 잘 적합시키는 모델
  - Good PM: 새로운 사례를 정확하게 예측하는 모델

- 데이터 셋
  - EM: 전체 데이터셋
    - 모집단에서 가정된 관계에 대한 정보가 최대한 반영된 최적의 적합 모델을 추정하기 위해 전체 데이터 세트를 사용
  - PM: 학습용 및 검증용 데이터셋
    - 데이터는 일반적인 학습 세트와 검증 세트로 나뉘어지며, 학습 세트는 모델을 추정하는데 사용되며, 검증 세트는 새로운 데이터에 대한 모델의 성능을 평가하는데 사용

- 성능 평가 기준
  - EM: 데이터가 모델에 얼마나 잘 적합하는가?
  - PM: 모델이 얼마나 새로운 사례를 잘 예측하는가?

예측 모델링

- 목표: 입력 변수는 있지만, 실제 값은 존재하지 않는 다른 데이터에서 출력값을 예측하기 위함
  - 즉, 새로운 사례에 대한 출력값을 알아내기 위함(DM)

✓ 전형적인 데이터 마이닝의 맥락 중 하나

✓ 모델 목표: 예측 정확도 최적화

✓ 학습용 데이터로 모델 학습

✓ 검증용 데이터로 성능 검증

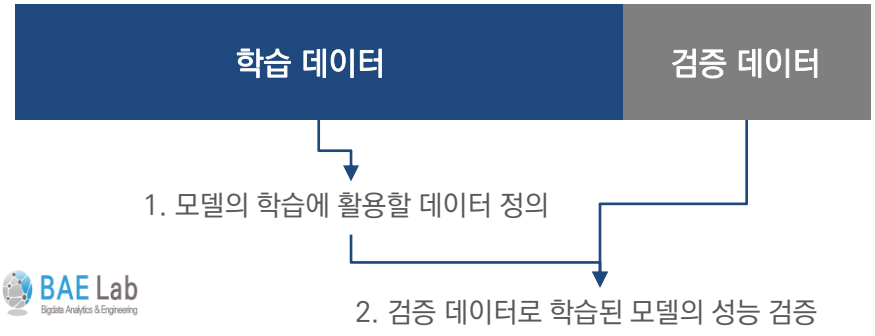
✓ 예측에 활용되는 변수가 미치는 영향을 설명하는 것이 주 목적은 아니지만, 중요한 문제

학습데이터 및 검증데이터 개념 및 분할 방식

- 학습 데이터
  - 모델이 학습하는데 사용할 데이터

- 검증 데이터
  - 학습한 데이터의 성능을 검증하기위한 데이터

- 보편적인 학습데이터 및 검증데이터 분할: 학습 데이터 70%, 검증 데이터 30%
  - 분석가가 상황에 맞게 임의로 분할



회귀 방정식 및 기본 가정

- Regression
  - Simple vs. Multiple
  - Linear vs. Non-linear

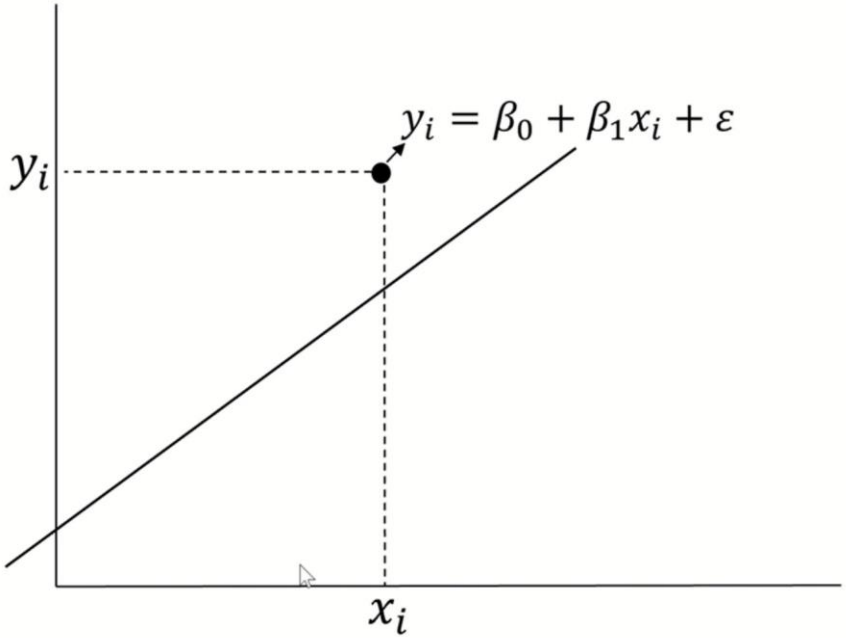
- 선형 회귀 모델(LRM, Linear Regression Model)

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

- 아래 4가지 조건을 만족한다면?
  - 추정한 계수들의 기댓값이 모수와 동일하며(Unbiased), 최소의 오차를 가짐

1. 정규성: 잔차(ε)의 분포가 정규 분포를 따름
2. 선형성: 입력 변수와 출력 변수는 선형 관계를 가짐
3. 독립성: 입력 변수들은 독립적임
4. 등분산성: 잔차(ε)의 분산의 분포가 입력 변수에 상관없이 같음

$$\varepsilon_i \sim (0, \sigma^2)$$



회귀 계수 추정 방법

- Y_i와 $\mathrm{E}(Y_i)$의 차이를 최소화하여 계수 추정

- 최소 제곱법을 활용한 추정

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum {x_i}^2 - (\sum x_i)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \widehat{\beta_1}\bar{x}$$

예제: Toyota Corolla의 가격

- 목표: 데이터를 기반으로 중고 Toyota Corollas의 가격 예측
- 데이터: 1442개 Toyota Corollas의 종속 변수 및 독립변수

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
18600	30	75889	Diesel	90	1	0	2000	3	210	1245
21500	27	19700	Petrol	192	0	0	1800	3	100	1185
12950	23	71138	Diesel	69	0	0	1900	3	185	1105
20950	25	31461	Petrol	192	0	0	1800	3	100	1185

Price in Euros
Age in months as of 8/04
KM (kilometers)
Fuel Type (diesel, petrol, CNG)
HP (horsepower)
Metallic color (1=yes, 0=no)
Automatic transmission (1=yes, 0=no)
CC (cylinder volume)
Doors
Quarterly_Tax (road tax)
Weight (in kg)

학습된 회귀 모델

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3608.418457	1458.620728	0.0137	97276410000
Age_08_04	-123.8319168	3.367589	0	8033339000
KM	-0.017482	0.00175105	0	251574500
Fuel_Type_Diesel	210.9862518	474.9978333	0.6571036	6212673
Fuel_Type_Petrol	2522.066895	463.6594238	0.00000008	4594.9375
HP	20.71352959	4.67398977	0.00001152	330138600
Met_Color	-50.48505402	97.85591125	0.60614568	596053.75
Automatic	178.1519013	212.0528565	0.40124047	19223190
cc	0.01385481	0.09319961	0.88188446	1272449
Doors	20.02487946	51.0899086	0.69526076	39265060
Quarterly_Tax	16.7742424	2.09381151	0	160667200
Weight	15.41666317	1.40446579	0	214696000

Hypthesis

$$H_0 : \quad \beta_1 = \beta_{1,0}$$
$$H_1 : \quad \beta_1 \neq \beta_{1,0}$$

Test Statistic

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$$

$$se(\hat{\beta}_1) = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{\frac{n-2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$

Interval of H0's acceptance

$$-t_{\alpha/2,n-2} < T_0 < t_{\alpha/2,n-2}$$

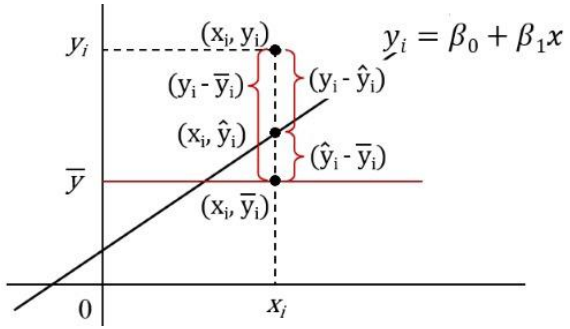
$$pvalue = 2 \times (1 - P(T \leq t_0)$$

```
Residuals:
     Min       1Q   Median       3Q      Max
-8212.5   -839.2    -14.3    831.5   7270.7

Coefficients:
                   Estimate   Std. Error t value             Pr(>|t|)
(Intercept)     -1774.877829  1643.744823  -1.080                0.2807
Age_08_04        -135.430875     4.875906 -27.776 < 0.0000000000000002 ***
KM                 -0.019003     0.002341  -8.116  0.00000000000000283 ***
Fuel_TypeDiesel  1208.339159   534.431400   2.261                0.0241 *
Fuel_TypePetrol  2425.876714   520.587979   4.660  0.00000391697679667 ***
HP                 38.985537     5.587183   6.978  0.00000000000811621 ***
Met_Color          84.792715   126.883452   0.668                0.5042
Automatic         306.684154   289.433138   1.060                0.2898
CC                  0.031966     0.099075   0.323                0.7471
Doors             -44.157742    64.056530  -0.689                0.4909
Quarterly_Tax      16.677343     2.602668   6.408  0.00000000030287017 ***
Weight             12.667487     1.536587   8.244  0.00000000000000109 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1406 on 588 degrees of freedom
Multiple R-squared:  0.8567,    Adjusted R-squared:  0.854
F-statistic: 319.6 on 11 and 588 DF,  p-value: < 0.00000000000000022
```

$$Y_i - \overline{Y} = \left(\hat{Y}_i - \overline{Y}\right) + \left(Y_i - \hat{Y}_i\right)$$
$$= \left(\hat{Y}_i - \overline{Y}\right) + e_i$$



예측 값

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1514553377	1325.527246	-0.000426154

Validation Data scoring - Summary Report

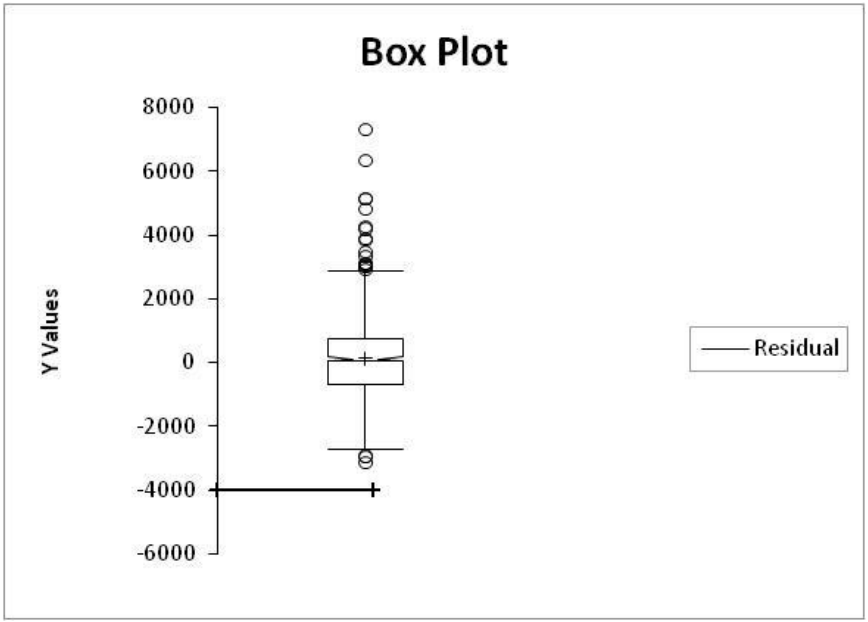
Total sum of squared errors	RMS Error	Average Error
1021587500	1334.079894	116.3728779

회귀 계수로 계산된 예측 값

Predicted Value	Actual Value	Residual
15863.86944	13750	-2113.869439
16285.93045	13950	-2335.930454
16222.95248	16900	677.047525
16178.77221	18600	2421.227789
19276.03039	20950	1673.969611
19263.30349	19600	336.6965066
18630.46904	21500	2869.530964
18312.04498	22500	4187.955022
19126.94064	22000	2873.059357
16808.77828	16950	141.2217206
15885.80362	16950	1064.196384
15873.97887	16250	376.0211263
15601.22471	15750	148.7752903
15476.63164	15950	473.3683568
15544.83584	14950	-594.835836
15562.25552	14750	-812.2555172
15222.12869	16750	1527.871313
17782.33234	19000	1217.667664

잔차: 실제 값과 예측 값의 차이

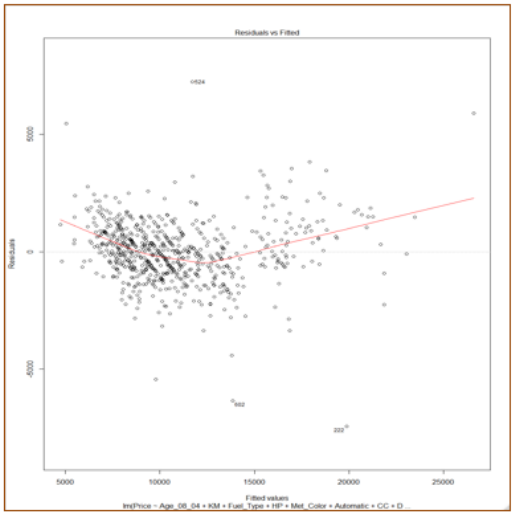
잔차의 분포

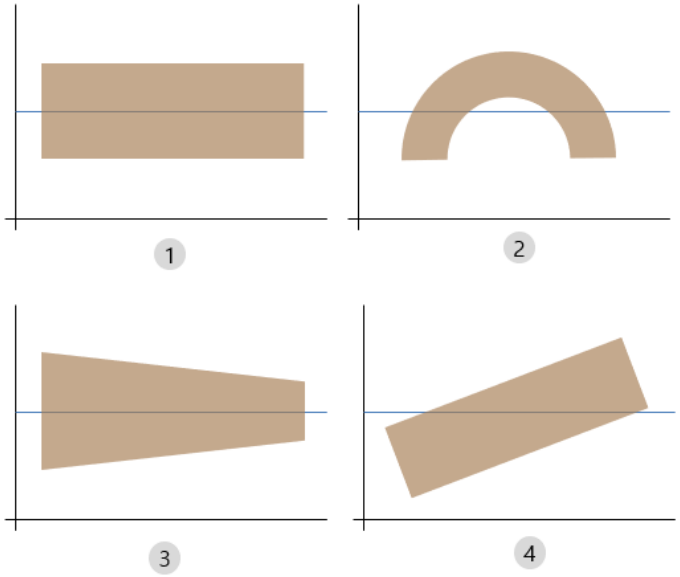


대칭 형태의 분포

일부 이상치 존재

잔차 분석


Residuals vs Fitted
1524
222
5000
0
-5000
Residuals
5000
10000
15000
20000
25000
Fitted values


1
2
3
4

입력 변수의 수를 줄여야 하는 이유

- 많은 입력 변수를 가진다면?
  1. 예측 변수 전부를 수집하는 것이 불가능하거나, 비용이 많이 듬
  2. 적은 수의 예측 변수로 더 정확한 예측을 할 수 있음
  3. 예측 변수가 많을수록 결측치 존재의 위험성이 높아짐
  4. 다중 공선성이 발생하여, 회귀 계수의 추정치가 불안해질 수 있음
  5. 종속변수와 상관없는 예측변수 사용 시, 예측의 분산이 증가할 수 있음
  6. 종속변수와 상관관계가 있는 예측 변수를 누락시킬 시, 예측 오차 혹은 편향도가 증가할 수 있음

참고) 다중공선성

- 두 개 이상의 예측 변수가 종속 변수에 동일한 선형 관계를 공유하는 것

완전 탐색(Exhaustive Search)

- R^2: 모델이 설명할 수 있는 변동성의 비율
  - 평가된 예측 변수들의 모든 부분집합 (single, pairs, triplets, etc.)
  - 계산 집약적임
  - Adjusted R^2를 통해 독립변수가 많아지는 경우에 대한 문제 해결

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

입력 변수의 수에 대한 패널티

입력변수의 수만 증가시켜도 발생하는 R^2의 인위적인 증가를 배제

입력 변수 선택 방법

- 목표: 충분히 좋은 성능을 보이는 가장 간단한 모델을 찾는 것
  - 모델이 더 강건해짐
  - 예측 정확도가 더 높아짐

- 완전 탐색(Exhaustive Search)

- 부분 탐색 알고리즘
  - 전진 선택법(Forward Selection)
  - 후진 제거법(Backward Elimination)
  - 단계 선택법(Stepwise Selection)

1	2	3	4	5	6	7	8
Constant	Age_08_04	*	*	*	*	*	*
Constant	Age_08_04	Weight	*	*	*	*	*
Constant	Age_08_04	KM	Weight	*	*	*	*
Constant	Age_08_04	KM	el_Type_Petrol	Weight	*	*	*
Constant	Age_08_04	KM	el_Type_Petrol	Quarterly_Tax	Weight	*	*
Constant	Age_08_04	KM	el_Type_Petrol	HP	Quarterly_Tax	Weight	*
Constant	Age_08_04	KM	el_Type_Petrol	HP	Automatic	Quarterly_Tax	Weight

모델(예측 변수 6개)

The Regression Model

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3874.492188	1415.003052	0.00640071	97276411904
Age_08_04	-123.4366303	3.33806777	0	8033339392
KM	-0.01749926	0.00173714	0	251574528
Fuel_Type_Petrol	2409.154297	319.5795288	0	5049567
HP	19.70204735	4.22180223	0.00000394	291336576
Quarterly_Tax	16.88731384	2.08484554	0	192390864
Weight	15.91809368	1.26474357	0	281026176

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1514553377	1325.527246	-0.000426154

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1021587500	1334.079894	116.3726779

Model Fit →

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1516825972	1326.521353	-0.000143957

Predictive performance
(compare to 12-predictor model!) →

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1021510219	1334.029433	118.4483556

정리

- 선형 회귀 모델은 설명 뿐 아니라, 예측까지 할 수 있는 유명한 도구임
- 좋은 예측 모델은 높은 정확도를 가짐
- 예측 모델은 학습용 데이터 세트를 사용하여 구성되며, 별도의 검증 데이터 세트에서 평가됨
- 예측 정확성과 강건함을 위해서는 불필요한 입력 변수를 제거하는 것이 중요함
- 변수 선택 방법(전진 선택법, 후진 제거법, 단계 선택법): 좋은 후보 모델을 찾는데 도움이 됨.
  - 이러한 작업을 실행한 후, 평가해야 함