

데이터마이닝 중간고사

Data Mining Midterm Examination

학 과		학 번		이 름	
일 시			총 점		/ 100

※ 모든 답은 답안 공간에 명확히 기재하시오. O, X 문항은 문항 앞 괄호 안에 O 또는 X로 답하시오. 각 문항의 배점은 문항 끝 (+N)으로 표시하였음.

4. 아래 결과는 R에서 선형회귀를 구축한 후 모델을 프린트해본 결과이다. 세부 문항에 답하시오.

OLS Regression Results						
Dep. Variable:	target	R-squared (uncentered):	0.960			
Model:	OLS	Adj. R-squared (uncentered):	0.959			
Method:	Least Squares	F-statistic:	723.4			
Date:	Wed, 31 Mar 2021	Prob (F-statistic):	4.08e-264			
Time:	10:33:47	Log-Likelihood:	-1214.1			
No. Observations:	404	AIC:	2454.			
Df Residuals:	391	BIC:	2506.			
Df Model:	13					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
CRIM	-0.1016	0.035	-2.911	0.004	-0.170	-0.033
ZN	0.0543	0.016	3.411	0.001	0.023	0.086
INDUS	-0.0306	0.071	-0.429	0.668	-0.171	0.109
CHAS	3.5489	1.057	3.357	0.001	1.470	5.628
NOX	-3.6282	3.599	-1.008	0.314	-10.704	3.447
...						
				(이하 변수 생략)		
Notes:						
[1] R ² is computed without centering (uncentered) since the model does not contain a constant.						
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.						
[3] The condition number is large, 8.24e+03. This might indicate that there are strong multicollinearity or other numerical problems.						

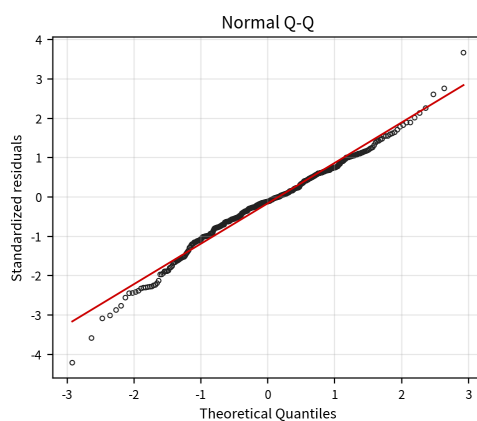
(그림 1) OLS Regression Results

4.1 그림 1을 보고 적합시킨 모델에 대한 선형회귀식을 적어보세요. (변수명은 긴 경우 줄여서 써도 됩니다.)

4.2 그림 1을 보고 유의수준을 0.01로 잡았을 때 각 변수의 유의함을 설명하시오.

4.3 그림 1을 보고 결정계수 값과 수정 결정계수 값을 적고, 수정 결정계수가 존재하는 이유를 설명하시오.

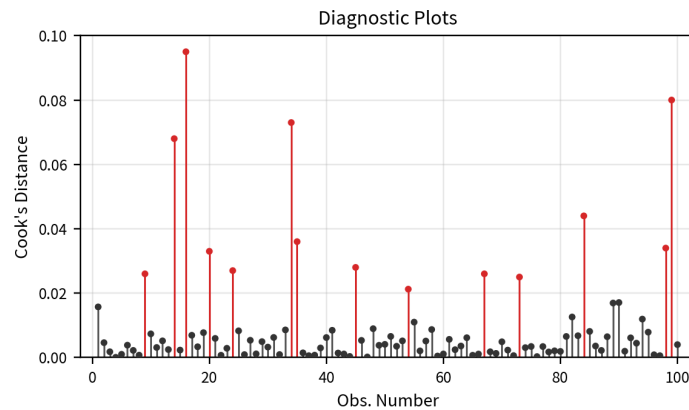
4.4 그림 2는 선형회귀 모델 가정 확인을 위해 Quantile-Quantile Plot을 보여준다. 이 그래프에 대해 설명하고, 이 그래프로부터 알 수 있는 정보를 쓰시오.



(그림 2) Normal Q-Q plot

5. 아래 문항은 선형회귀모델의 가정과 진단에 관한 질문이다. O, X로 답변하세요. (+21)

- 5.1 () 다중공선성(Multicollinearity)은 선형회귀 학습모델 구축 시 독립변수 간 강한 상관관계가 발생하는 문제이다. (+3)
- 5.2 () Gauss-Markov theorem에 따르면, 모델은 선형조건(linearity)을 만족하고, 오차(e)의 기대값과 분산이 각각 0과 σ^2 이며, $\text{var}(e) = \sigma^2 = \text{var}(y)$ 이다. (+3)
- 5.3 () 선형회귀에서 ordinary least squares에 따라 파라미터 β_j 를 추정하면 best linear unbiased estimator가 된다. (+3)
- 5.4 () 선형회귀모델에서 다중공선성 문제 발견을 위해서, 불필요한 중복 설명력 변수를 제거하거나 shrinkage model (Lasso/Ridge/Elastic net) 회귀모형을 구축해볼 수 있다. (+3)
- 5.5 () Variance inflation factor(VIF)는 독립변수 간 선형 상관성을 반영하는 지표로, 종속변수와 직접 비교하지 않고 일정 수준 이상 높을 경우 회귀모델이 불안정하다는 것을 의미한다. (+3)
- 5.6 () 선형회귀모델의 등분산성 문제는 상태지수(condition index)를 기준으로 발견 가능하다. (+3)
- 5.7 () 아래 그림은 선형회귀 모형진단(model diagnostics) 시 Cook's distance를 가시화하고 있는데, 이는 특정 관측치가 회귀계수 추정에 얼마만큼 영향력을 끼치는지 진단하며, 오차항이 서로 독립이고 평균 0의 정규분포를 따르는지 진단하는 데 효과적이다. (+3)



(그림 3) Diagnostic Plots: Cook's Distance

6. 아래 문항은 선형회귀모델과 모델 정규화에 대한 질문이다. O, X로 답변하세요. (+12)

- 6.1 () 선형회귀모델은 $SST = SSE + SSR$ 을 만족하나, outliers가 다수 섞인 경우 $SST \leq SSE + SSR$ 일 수 있다.
※ Total Sum of Squares (SST), Sum of Squares of Error (SSE), Sum of Squares Regression (SSR) (+4)
- 6.2 () 선형모델 $Y = f^*(X) + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$ 에서 ε 는 독립동일분포(independent identically distributed; iid) 조건을 만족해야 하는데, 데이터 분포의 시간변화가 발생할 경우는 보장하지 않는다. (+4)
- 6.3 () Bias-Variance decomposition을 통해 선형 분리한 Bias와 Variance는 각각 예측 오차와 편차를 가리키고 있으며, 이중 Bias는 실제 값 대비 평균 추정 값 간의 차이가 얼마나 큰지를 잘 설명한다. (+4)

7. 선형회귀의 목적식은 **convex function**에 해당하여 각 파라미터에 대해 미분의 0값을 기준으로 최적값을 찾는 과정을 설명하고 있다. 식 1과 식 2를 이용해서 β_0 와 β_1 가 식 3, 4와 같은 수식이 됨을 보이시오. (+10)

$$\frac{\partial C(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)] = 0 \quad \text{Eq. 1}$$

$$\frac{\partial C(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)] X_i = 0 \quad \text{Eq. 2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad \text{Eq. 3}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad \text{Eq. 4}$$

$$f(X) = \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \quad \text{Eq. 5}$$

8. 정규화 선형모델인 Ridge, Lasso, Elastic net regression 중 모델 해석력(interpretation)이 좋은 모델과 일반화 성능(penalizing performance)이 높은 모델을 하나씩 정하고, 그 이유를 **penalty term**을 기준으로 설명하시오. (+5)

9. 아래의 수식은 정규화(Regularization)의 개념을 설명하고 있다. 아래의 첫 번째 term, 두 번째 term이 각각 무엇을 의미하는가? λ 의 의미는 무엇이고 이 값이 클 때와 작을 때는 어떻게 되는지 설명하시오. (+10)

$$L(\beta) = \min_{\beta} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

10. 아래는 Ridge regression에 대한 식을 보여준다. 이때 1) λ 와 t의 관계를 설명하고, 우측의 β_1 , β_2 축면의 그림에서 2) 빨간색 원이 표현하는 의미를 간단히 표현하시오. (+10)

$$\hat{\beta}^{ridge} = \operatorname{argmin}_{\beta} \left(\sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right)$$

$$\hat{\beta}^{ridge} = \operatorname{argmin}_{\beta} \sum_{i=1}^n (y_i - x_i \beta)^2$$

목적식

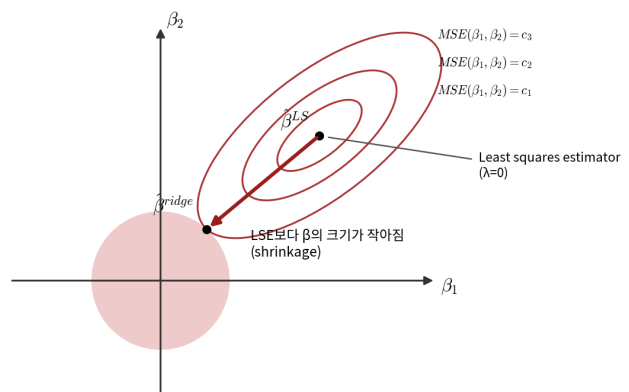
(Objective function)

$$\text{subject to } \sum_{j=1}^p \beta_j^2 \leq t$$

제약식

(Constraints)

Ridge 회귀의 목적식 및 제약식



Ridge solution path

— 끝 —

[복기본 안내] 이 시험지는 시험 중 화면 사진을 바탕으로 복기한 것입니다. 사진에서 잘렸던 문장(5.1~5.5, 4.3, 4.4 등)은 문맥에 맞게 보완했고, 그림 2(Q-Q plot)·그림 3(Cook's distance)·Ridge solution path는 사진을 참고해 다시 그린 것이라 실제 시험지와 세부 모양이 다릅니다. 그림 1은 사진에서 읽힌 변수(CRIM~NOX)만 수록했습니다. 4번 문항의 배점, 1~3번 문항, 6.1의 배점(6번 총점 12점을 균등 배분)은 확인되지 않았습니다.