Handling Categorical Variables

House;

‘waterfonrt’ -> binary

‘view’, ‘condition’, ‘grade’ -> ordinal ( Category )

‘zipcode’ -> nominal

Regression with categorical variables

Dummy variable;

= indicator variable / binary variable

Convert to 0, 1;

Selected variable: 1, else: 0

One-way ANOVA

* Split into groups by factor ‘X’
* determine whether there are any statistically significant differences between the means of two or more independent (unrelated) groups

Assumption;

Normality – That each sample is taken from a normally distributed population

Sample independence – that each sample has been drawn independently of the other samples

Variance Equality – That the variance of data in the different groups should be the same

Interpretation

모형 적합도

P > |t| < 0.01

Statistically significant to predict target

Prob (F-Statistic) < 0.01

이 회귀모형은 통계적으로 유의미하다

AIC, BIC

작을수록 좋다

Residuals; Last section

Prob(JB) -> 0;

Residual does not follow normal distribution

Durbin-Watson

< 2; There is positive correlation

= 2; No correlation

:> 2; Negative correlation

Omnibus

Checking whether follow normal distribution using skewness and kurtosis

Condition number

연립방정식을 풀 때 b 추정치의 변동이 큰가? = 다중 공선성 체크

X 변수가 많으면 자연스레 커진다 -> 신경 ㄴ