오차로부터 단서 찾기

잔차, 등분산성, VIF

1. Concept Def.

1.1.1 오차(Error)의 정의 : 실제 값과 측정값(계산된 값)의 차이.

===> 이걸 고려해야 되는 이유 또는 원인?

1.1.2 오차의 종류

- 모델오차(Model Error)
- 예측오차(Prediction Error)
- 측정오차(Measurement Error)

1.2.1 잔차(Residual)의 정의

실제값과 예측값 사이의 측정값

1.3 오차와 잔차의 차이점

- 오차는 모집단, 잔차는 표본 데이터에서 사용
- 오차는 이론적, 잔차는 실제적 관측값
- 잔차 —> 오차를 추정

⇒ 오차, 잔차가 작을 수록 궁극의 목적인 집단에 대한 해석이 가능해지고, 불확실성을 낮춤

2. 잔차 분석의 중요성

- 잔차의 측정이 구성 또는 구축한 모델의 품질을 좌우

  => 모집단에 대한 이해, 예측에 대한 불확실성 감소, 궁극적 이해를 높임

- 모델 적합성 평가 :

  잔차의 무작위 분포는 모델의 데이터 구조를 잘 포착

  잔차에 패턴(U선형, 선형경향, 군집)이 있으면 모델의 오류가 있음

- 이상값 발견
- 모델이 개선방향 도출(변수(피처)의 추가, 대안 모델의 구축)

잔차 패턴	의미	해결 방법
U자형 또는 곡선 형태	비선형 관계를 선형 모델이 설명하지 못함	다항 회귀, 비선형 모델 사용
잔차가 특정 구간에서만 크거나 작음	이분산성 존재 (Heteroscedasticity)	가중 회귀, 로그 변환 등
잔차가 시간에 따라 연속적으로 증가/감소	자기상관 존재 (Autocorrelation)	시계열 모델 적용 (ARIMA 등)
잔차가 특정 그룹에 몰림	누락된 변수 또는 그룹 효과	변수 추가, 혼합 모델 고려

3. 잔차 지표

1. 잔차 제곱합 (RSS, Residual Sum of Squares)

RSS는 모델의 예측값과 실제 값의 차이인 **잔차(Residual)**를 제곱하여 모두 더한 값입니다. 오차의 전체 합을 나타내며, 이 값이 작을수록 모델이 데이터를 잘 설명하고 있다고 평가합니다. 잔차를 그냥 더하면 양수와 음수가 상쇄되어 오차의 총합을 정확히 알 수 없기 때문에, 제곱하여 항상 양수 값을 갖도록 합니다. 이 때문에 **특정 오차 값의 크기가 클수록 전체 합에 더 큰 영향을 미칩니다.**

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

활용 : 모델 비교보다는 모델 내부의 오차구조 파악

2. 평균 제곱 오차 (MSE, Mean Squared Error)

MSE는 RSS를 데이터 샘플의 수(

$$n$$

)로 나눈 값입니다. 이는 **오차의 평균적인 크기**를 나타냅니다. RSS와 달리 데이터의 개수가 다른 모델들을 비교할 때 더 유용합니다. 하지만 단위가 실제 값의 단위와 다르기 때문에 직관적인 해석이 어려울 수 있습니다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

활용 : 학습과정에서 손실함수로 사용됨, 모델의 평균적 성능을 평가할때 적합

3. 평균 제곱근 오차 (RMSE, Root Mean Squared Error)

RMSE는 MSE에 제곱근을 취한 값입니다. 가장 큰 특징은 **실제 값과 단위가 동일**하다는 점입니다. 이 때문에 모델의 예측 오차가 평균적으로 얼마나 되는지 직관적으로 이해하기 쉽습니다. 예를 들어, RMSE 값이 10달러라면, 모델의 예측이 평균적으로 실제 가격과 10달러 정도 차이 난다고 해석할 수 있습니다. RMSE는 회귀 모델 평가 지표 중 가장 널리 사용되는 지표 중 하나입니다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

이 세 지표는 모두 값이 작을수록 모델의 성능이 좋다고 판단합니다. 제공해주신 설명은 매우 정확하며, 이 지표들에 대한 핵심적인 내용을 잘 요약해주셨습니다.

활용 : 실제 예측값과의 차이를 직관적으로 보여 줌(예측된 매출 vs. 실제 매출)

화두 : "모델이 남긴 '오차'는 실패의 증거인가, 아니면 더 나은 모델로 가는 길을 알려주는 단서인가?”

- 모델의 오차는 모델이 더 나은 방향 즉 개선을 할수 있도록 하는 단서적 역할 수행
- 한번의 모델 구축이 해답을 찾는 것은 불가하며, 반복적 또는 새로운 접근이 필요
- 오차가 단서?

- 모델의 성능 평가 지표 : 오차의 개선(작아지도록)하는 방향 제시
- 모델 개선의 방향제시

  모델의 취약부분 파악, 과소적합과 과대적합, 이상치 처리문제

- 모델의 기본 가정 검증

  모델에 따라 다르지만 잔차가 정규분포를 따르고, 분산이 일정하다는 가정을 전제함

등분산성(Homoscedasticity) vs. 이분산성 (Heteroscedasticity)

등분산성 정의 : 회귀분석 모델의 잔차(오차) 분산이 모든 독립변수 간에 일정한 현상

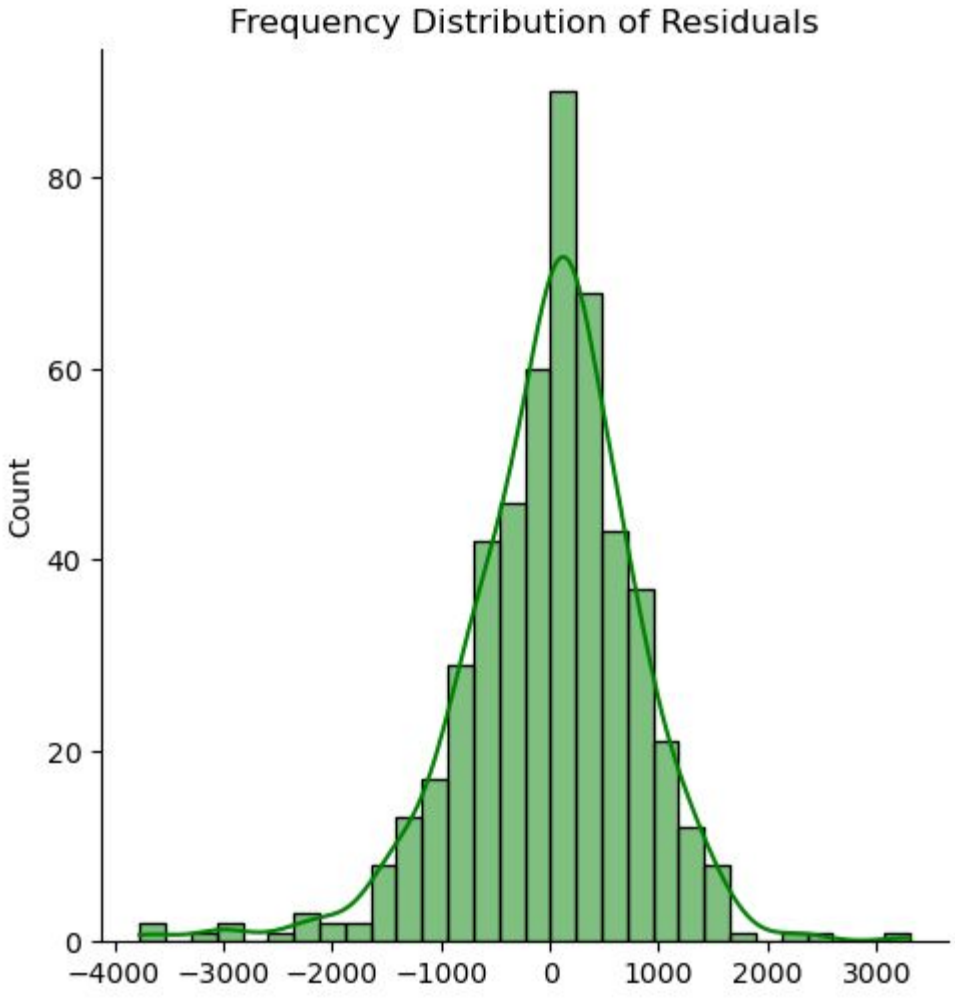
=> 잔차의 분산이 독립변수 값에 무관하게 일정하다

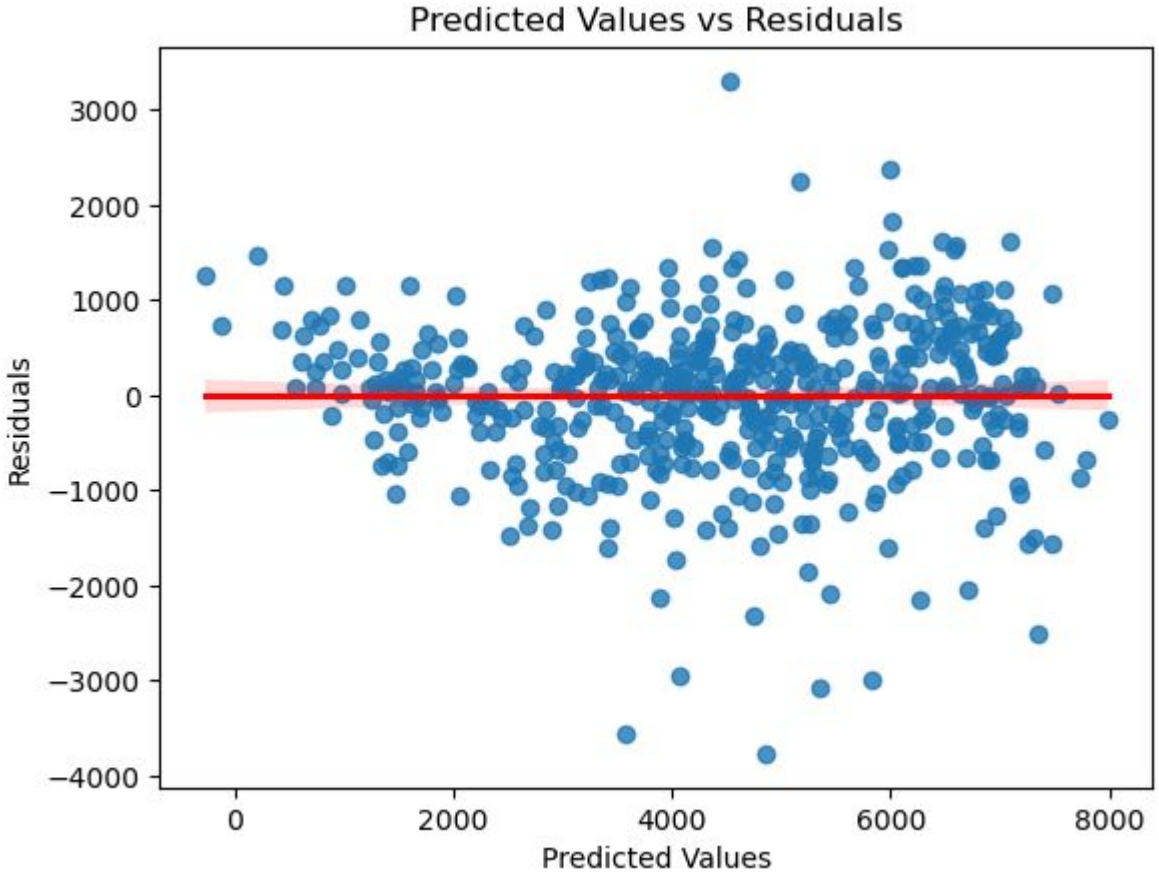
이분산성 정의: 잔차의 분산이 독립변수 값에 따라 달라지는 현상

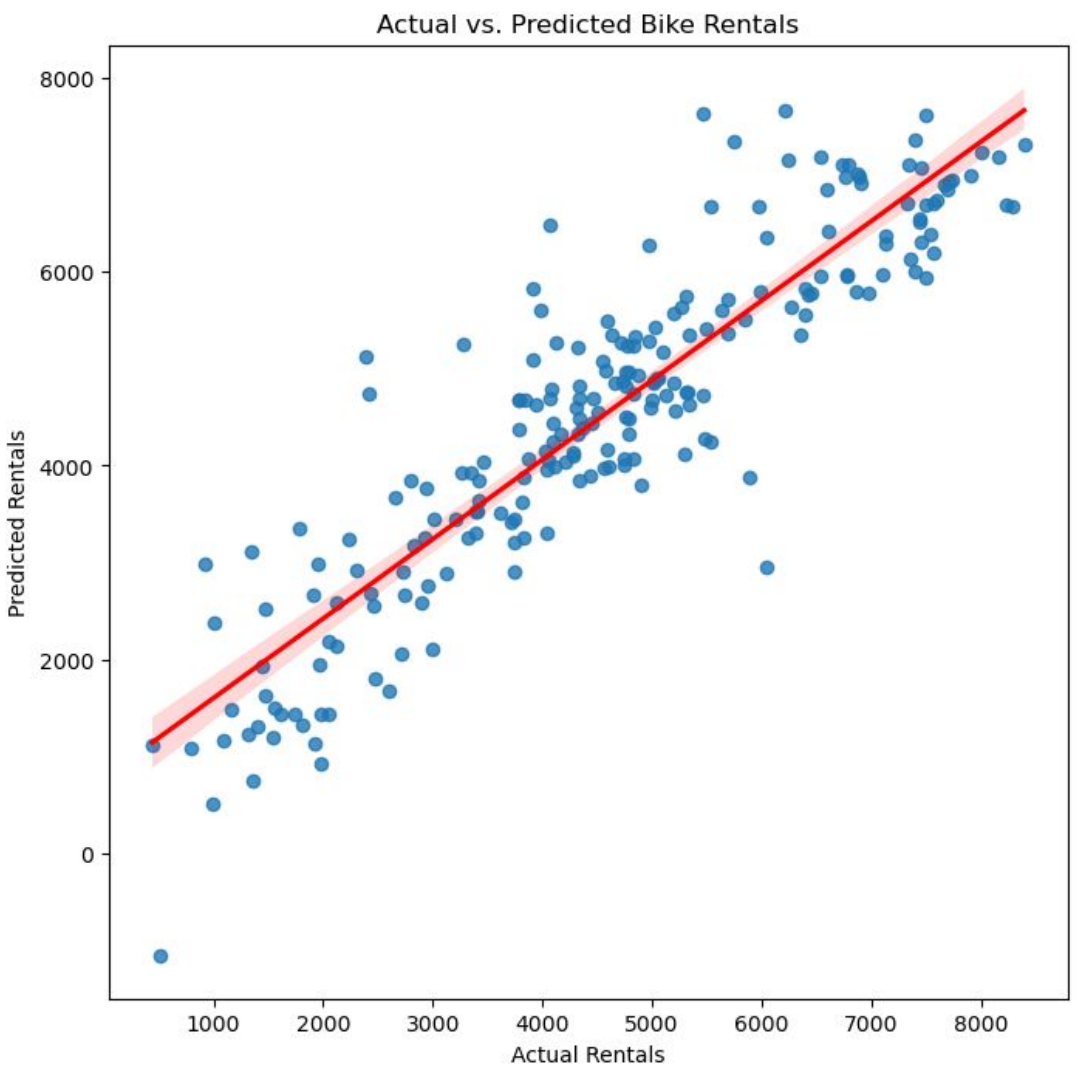
=>모델의 예측력이 일정하지 못한 일관성이 부족함(예측값 ⇧ , 오차값 ⇧)

중요성 : 추정치의 표준오차가 부정확해지며, p값이 왜곡되어 변수의 유의미도에 부정적 영향을 미쳐 모델의 예측의 신뢰도 저하시킴

진단방법 : 잔차의 산점도를 보고 시각적으로 판단


Frequency Distribution of Residuals
Count
0
20
40
60
80
−4000
−3000
−2000
−1000
0
1000
2000
3000


Predicted Values vs Residuals
Residuals
Predicted Values
3000
2000
1000
0
−1000
−2000
−3000
−4000
0
2000
4000
6000
8000


Actual vs. Predicted Bike Rentals
Predicted Rentals
Actual Rentals
8000
6000
4000
2000
0
1000
2000
3000
4000
5000
6000
7000
8000

이분산성 해결방법

- 변수변환 : 로그변화, 제곱근 변환
- 가중 최고 제곱법(WLS) : 오차가 큰데이터 포인트에 더 작은 가중치를 부여하여 재학습 시킴
- 로버스트 표준 오차(Robust Standard Errors): 보정된 표준오차를 사용하여 p값과 신뢰구간을 다시 계산

다중공선성(Multicollinearity) 평가

1. 정의

   회귀분석에서 독립변수들 간에 강한 상관관계가 존재하는 현상(문제점 피처의 개념적 차이가 전혀 무관함에도 발생하는 다중공선성은 무엇일까?)

2. 진단방법

- 상관관계행렬(Correlation Matrix) 확인 : >0.7이상은 다중공선성 의심
- 분산팽창계수(VIF, Variance Inflation, Factor)계산

=>VIF는 변수가 다른 변수에 의해 설명되는 수준을 의미하며, >10이상은 판단

- 문제점
  - 계수 추정치의 불안정성 : 변수가 영향력 판단 불가, 계수(변수의 영향력)을 오류에 영향
  - 통계적 유의성 판단 오류
  - 모델 해석의 어려움

- 해결방법
  - 변수제거
  - 주성분분석(PCA)로 차원 축소 **=>** 새로운 피처적 접근방법
  - 릿지회귀, 라쏘회귀에서 페널티 항을 추가하여 계수 추정치를 보다 구체화 시킴

릿지와 라소를 통한 방법

- 릿지회귀
  - RSS에 L2 페널티항 추가(계수를 모두 제곱하고(베타), 하이퍼패러미터인 람다를 곱함

$$\text{Cost Function} = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- 라쏘회귀
  - RSS에 L1 페널티항 추가

$$\text{Cost Function} = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

특징	릿지 회귀 (L2 페널티)	라쏘 회귀 (L1 페널티)
페널티 형태	계수의 제곱합	계수의 절댓값 합
계수 축소 방식	계수를 0에 가깝게 줄임	일부 계수를 아예 0으로 만듦
장점	모든 변수를 포함하여 모델의 안정성 높임	불필요한 변수를 제거하여 모델을 단순화
주요 사용처	다중공선성이 높을 때	변수 선택이 필요할 때

=> 변수의 계수 값을 0으로 만든 방법

```python
# Checking the VIF
vif = pd.DataFrame()
vif['features'] = X_train_rfe.columns
vif['vif'] = [variance_inflation_factor(X_train_rfe.values,i) for i in range(X_train_rfe.shape[1])]
vif['vif'] = round(vif['vif'],2)
vif = vif.sort_values(by='vif', ascending=False)
vif
```

VIF >10 다중공선성, 값의 기준에 차이에 따라 고려 상황이 달라짐

보수적인 경우 5를 기준으로 하기도 (제시 예에선 5를 적용)

Bike에서는 hum, Temp가 높게 나타남

atemp는 RFE과정에서 삭제됨

=> 변수 제거를 통해 재실행 반복

	features	vif
3	hum	29.32
2	temp	16.81
4	windspeed	4.72
5	Spring	4.37
7	Winter	3.78
6	Summer	2.80
13	Misty	2.31
0	yr	2.09
8	Jan	1.67
10	Nov	1.60
9	July	1.59
11	Sep	1.40
14	Rain	1.26
12	Sun	1.19
1	holiday	1.06

	features	vif
2	temp	5.14
3	windspeed	4.60
5	Summer	2.24
4	Spring	2.11
0	yr	2.07
6	Winter	1.81
7	July	1.59
10	Misty	1.56
8	Sep	1.34
9	Sun	1.18
11	Rain	1.08
1	holiday	1.05

```
temp        4368.639370
yr          2020.888840
const       1384.115819
Winter       747.048000
Sep          698.367431
Summer       327.723759
Sun         -420.590597
July        -430.368213
Spring      -651.340030
Misty       -699.638403
holiday     -938.761398
Rain       -2644.415685
dtype: float64
```

```
X_train_sm_full = sm.add_constant(X)

lr = sm.OLS(y_train,X_train_sm_full)

lr_model = lr.fit()

lr_model.summary()
```

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.829
Model:	OLS	Adj. R-squared:	0.826
Method:	Least Squares	F-statistic:	220.2
Date:	Tue, 02 Sep 2025	Prob (F-statistic):	2.83e-183
Time:	16:13:07	Log-Likelihood:	-4135.9

- `Omnibus` / `Prob(Omnibus)` : `Prob(Omnibus)` 값이 `0.000` 이므로, 잔차가 **정규 분포를 따르지 않는다**는 것을 강하게 시사합니다. 이는 선형 회귀의 핵심 가정이 위반되었음을 의미하는 중요한 문제입니다.
- `Durbin-Watson` : 값이 ** `2.056` **으로 2에 매우 가깝습니다. 이는 잔차에 **유의미한 자기상관(autocorrelation)이 없다**는 것을 나타내며, 좋은 신호입니다.
- `Jarque-Bera (JB)` / `Prob(JB)` : `Prob(JB)` 값이 ** `5.61e-43` **으로 0에 가까운 매우 작은 수치입니다. 이 검정 역시 잔차가 **정규 분포를 따르지 않는다**는 것을 강력하게 뒷받침합니다.
- `Cond. No.` : 값이 ** `16.8` **로, 다중공선성(multicollinearity)의 일반적인 임계값인 30보다 훨씬 낮습니다. 이는 모델에 **심각한 다중공선성 문제가 없다**는 것을 의미하며, 계수 추정치가 안정적일 가능성이 높습니다.

#RFE를 통한 변수 제거(vs.Stepwise)

- 최적의 피처(Feature)조합을 찾는 방법
- 반복적 학습과정에서 중요하지 않은 특성을 하나씩 제거해 나가는 방법

=> 주어진 예에서는 이 방법으로 변수를 제거하는 방법으로 매출량과의 관계에 미치는 변수를 도출함

Decision Matrix for Feature Selection

- **High p-value, High VIF**: Eliminate the column
- **Low p-value, Low VIF**: Retain the column
- **High p-value, Low VIF**: Tentatively remove and reassess VIF impact
- **Low p-value, High VIF**: Consider removal if VIF remains elevated

Threshold Definitions

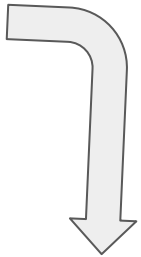
- High p-value: ≥ 0.05
- High Variance Inflation Factor (VIF): ≥ 5

```
# Coefficients and intercept

lr_model.params.sort_values(ascending=False)
```

✓ 0.0s

```
temp       4368.639370
yr         2020.888840
const      1384.115819
Winter      747.048000
Sep         698.367431
Summer      327.723759
Sun        -420.590597
July       -430.368213
Spring     -651.340030
Misty      -699.638403
holiday    -938.761398
Rain      -2644.415685
dtype: float64
```



```
#Equation of our best fitted line is:
# cnt = 1384.115819 + 4368.639370 temp + 2020.888840 yr
#       + 747.048000 Winter + 698.367431 Sep
#       + 327.723759 Summer - 420.590597 Sun
#       - 430.368213 July - 651.340030 Spring
#       - 699.638403 Misty - 938.761398 holiday
#       - 2644.415685 Rain
```

질문 : 변수간 개념 중복이 없는 경우 다중공선성은?

- 데이터 수집과 생성 과정에서 발생한 우연성

- 숨겨진 잠재적 공통 요인(latent factor) 존재: 공통된 제3요인의 영향

- 데이터의 전처리 또는 파생 변수를 만드는 과정에서의 중복

=> VIF 값으로 결정, 주성분 분석을 수행(통합, 제거), 도메인 지식을 활용 재검토와 재정의(Context를 이해)