Report

이준찬

January 20, 2025

1 Linear Regression Report

1.1 다중선형회귀 수행

sklearn의 linear model 패키지를 사용해 다중선형회귀를 수행하였다. 분석 데이터는 4개의 float 변수, TV, radio, newspaper, sales로 구성된 200개의 데이터를 활용하였으며 sales 를 독립변수로, 나머지 세 변수를 종속변수로 사용하였다.

분석 패키지로는 sckit learn의 linear model 패키지를 사용하였으며 학습 방식으로는 OLS를 사용하였다. 별도의 정규화 방식, 스케일링, 결측 or 이상치 제거는 수행하지 않았다.

1.2 Summary 해석

constant, TV, radio, newspaper 에 대해 추정된 회귀계수는 coef 열에서 확인할 수 있다. 각 변수의 해석은, 각 변수의 값이 1 증가할 때 선형 모델이 추정하는 sales 값의 변화량이다. 표본이 200개 이상으로 대표본이기 때문에, 원본 데이터의 잔차가 정규분포를 따르고 등분산이라는 가정 or CLT에 의한 정규근사에 의해 coef 추정치가 t 분포를 따른다. X와 y가 아무 관련이 없다는 귀무가설 $H_0 : \beta_i = 0$ 을 기각할지 수용할지를 t 통계량을 사용해 결정한다. summary의 t 값과 0.05의 유의확률에 의한 critical value를 비교해 $\beta_i = 0$ 으로 귀무가설이 맞을 때 현재 t값보다 극단적인 값이 나올 확률 $P > |t|$ 를 구한다. 또한 데이터의 분산과 데이터 크기로 구한 표준오차를 바탕으로 회귀계수의 95% 신뢰구간을 구한다. 신뢰구간안에 0이 있는 경우 p value가 유의확률을 벗어나게 되며 귀무가설을 수용한다. summary를 바탕으로, newspaper는 sales에 영향을 미치지 않는다고 볼 수 있다.

이외에 모델의 설명력 R squared는 0.897로 sales의 변화를 모델이 비교적 잘 추정한다고 볼 수 있다.

$$t = \frac{\beta_i - 0}{\mathrm{SE}(\beta_i)}$$

$$s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - k}$$

1.3 Correlation 해석

Correlation Matrix를 통해 각 변수 간의 선형 상관관계를 확인할 수 있다. 변수 간 상관관계가 높으면 회귀모델에서 다중공선성이 발생할 수 있다. 다중공선성의 높을 경우 모델 설명력은 유지되지만 각 변수들의 coefficient의 변동이 커져 해석이 어려워지므로 다중공선성이 클 경우 해당 변수를 제거해주어야 한다.

X 변수 간에는 newspaper와 radio가 0.35의 상관관계를 보였고, 나머지 변수들은 서로 큰 상관관계를 가지지 않았다. X와 y변수 간의 상관관계에서는 TV가 가장 높았고, newspaper가 가장 낮았다. 실제로 summary에서 TV의 회귀계수는 대립가설을 채택했으나, newspaper의 회귀계수는 기각되었다.

상관계수 correlation coefficient

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	570.3
Date:	Mon, 20 Jan 2025	Prob (F-statistic):	1.58e-96
Time:	19:52:24	Log-Likelihood:	-386.18
No. Observations:	200	AIC:	780.4
Df Residuals:	196	BIC:	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011

Omnibus:	60.414	Durbin-Watson:	2.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241
Skew:	-1.327	Prob(JB):	1.44e-33
Kurtosis:	6.332	Cond. No.	454.

Figure 1: OLS summary.

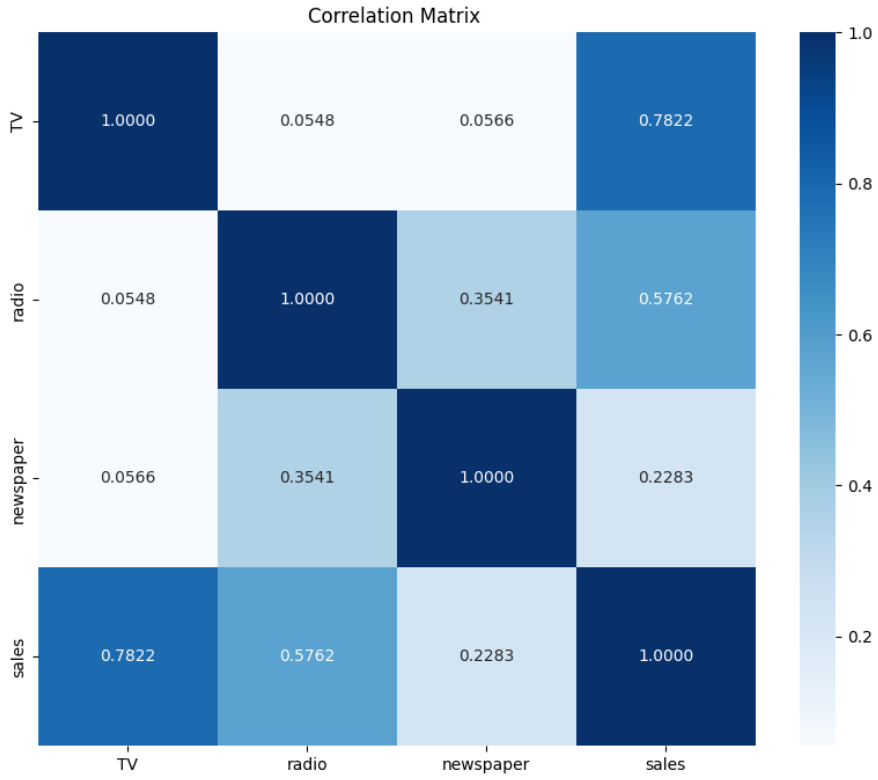

Figure 2: Correlation Matrix.