

회귀분석 report

인공지능학과 정민지

Linear Regression 모델

해당 과제는 advertisement.csv에 담긴 data를 읽어낸 후 다중 선형 회귀 모델을 사용해 TV, radio, newspaper이 sales에 미치는 영향을 분석하는 것을 목표로 한다.

Python의 `statsmodels` library를 사용해 multiple linear regression model을 구현했고 `intercept`(절편)을 만들기 위해 상수항을 추가했으며, `sales`를 반응 변수로, `TV`, `radio`, `newspaper` 광고비는 예측 변수로 설정했다. 회귀 모델 생성은 `statsmodels` 모듈의 `OLS` (Ordinary Least Squares) 함수를 이용해 생성했고 `fit` 함수를 통해 적합 과정을 거쳤다.

또한 R-squared, Adjusted R-squared과 각 예측 변수의 계수(coef), p-value 등의 지표를 그래프를 통해 통계로 나타내어 모델의 유의성을 평가했다.

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:	19:32:58		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]
Intercept	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.		

회귀분석 모델 결과

왼쪽 사진을 참고하면 회귀 모델의 성능 지표를 두 가지로 판단할 수 있다.

1. R-squared: 0.897
2. Adjusted R-squared: 0.896

이는 매출 변동성의 약 89.7%가 TV, radio, newspaper 광고비로 설명될 수 있다는 걸 보여준다.

회귀분석 결과: 주요변수들의 지표 분석

	coef	std err	t	P> t	[0.025	0.975]
Intercept	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

주요변수들 지표

1. Intercept(절편) 값은 모든 광고비가 0일 때 기본 매출 수준을 나타내고 위의 결과에서는 약 2.939 (단위)의 매출이 예상된다.
2. TV 광고비 계수는 0.046으로, 광고비가 1단위 증가할 때 매출이 0.046단위만큼 증가한다는 것을 보여준다. t-statistic 값이 높고(32.81) p-value 값은 낮으므로(<0.0001) TV 광고가 통계적으로 유의미한 예측 변수임을 알 수 있다.
3. radio 광고비 계수는 0.189로, 광고비가 1단위 증가할 때 매출은 0.189단위만큼 증가한다. p-value가 작으므로 (<0.0001) radio 또한 통계적으로 유의미한 변수임을 확인할 수 있다.
4. newspaper 광고비 계수는 -0.001로 매출에 유의미한 영향을 갖지 못한다. p-value가 높은 편으로 (0.8599) 위의 두 변수와 달리 newspaper 광고는 통계적으로 유의미하지 않다는 걸 알 수 있다.

Correlation 분석

correlation matrix는 각 변수 간 상관 관계를 나타낸다.

	TV	radio	newspaper	sales
TV	1.000000	0.054809	0.056648	0.782224
radio	0.054809	1.000000	0.354104	0.576223
newspaper	0.056648	0.354104	1.000000	0.228299
sales	0.782224	0.576223	0.228299	1.000000

변수 간 correlation matrix

1. TV - sales

TV 광고와 매출 간의 상관계수는 0.782로, 꽤나 강한 (+) 양의 상관관계에 있다. TV 광고비가 증가할수록 매출도 증가하는 경향이 크다는 뜻이다.

2. radio - sales

상관계수는 0.576이고, 중간 정도의 (+) 양의 상관관계이다. radio를 통한 광고 또한 매출에 긍정적인 영향을 갖지만, TV 광고보다는 그 정도가 작다.

3. newspaper - sales

상관계수는 0.228이고 가장 작은 상관관계를 가진다.(+ 상관관계이긴 하다.) newspaper 을 통한 광고는 매출에 거의 영향을 갖지 못한다는 것을 알 수 있다.

⇒ 1~3번의 결과는 앞서 진행한 회귀분석 결과와 일치한다.

4. 세 변수 간 상관계수

TV - radio(0.0548)와 TV - newspaper(0.0566)로 서로 거의 독립적이다.

radio - newspaper(0.3541)만 유일하게 약한(+)양의 상관관계를 갖는다.

Conclusion

regression 결과, TV와 radio 광고는 sales 증가에 효과적인 채널로 확인되었으며, newspaper 광고는 유의미한 영향을 갖지 않는 것으로 나타났다. 이 분석 결과를 바탕으로 광고비를 최적화해 매출 성과를 올릴 수 있다. 예를 들어, TV와 radio 광고는 sales 증대에 기여할 수 있는 채널로, 유지하고 키우는 전략을 취하고, sales에 미미한 영향을 보이는 newspaper 광고는 예산을 재검토하고 더 효과적인 다른 채널을 찾는 전략을 취할 수 있다.

또한 매출을 예측할 때 TV 광고가 가장 믿을 만한 지표이며,(가장 높은 상관계수) radio, newspaper순으로 믿을 만하다.(newspaper은 관련이 거의 없다고 봐야한다.) 세 독립변수들 간의 상관관계는 매우 낮으므로 Multicollinearity, 다중공선성 문제가 발생할 확률이 낮는데, 해당 regression 모델이 꽤 높은 신뢰성을 가지는 것으로 해석할 수 있다.