

Chapter. 01

[비주얼코딩] 코딩없이 하는 데이터 분석 SAS

106. 회귀분석

FAST CAMPUS
ONLINE

직장인을 위한 데이터 분석

강사. 최윤진

Chapter. 06

코딩 없이 하는 데이터 분석 SAS 06. 회귀분석

I 데이터

<http://bit.ly/vc-sas-a>
d

변수	변수 설명
TV	TV 광고비
radio	라디오 광고비
newspaper	신문 광고비
sales	매출액

I 회귀분석

```
[ ] import statsmodels.formula.api as sm
model1 = sm.ols(formula="sales~TV+radio+newspaper", data=df).fit()
#sales~TV+radio+newspaper
print(model1.summary())
```



OLS Regression Results

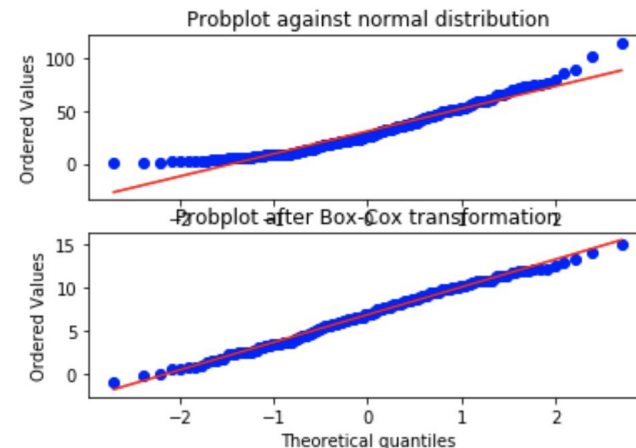
=====						
Dep. Variable:	sales	R-squared:	0.897			
Model:	OLS	Adj. R-squared:	0.896			
Method:	Least Squares	F-statistic:	570.3			
Date:	Tue, 10 Mar 2020	Prob (F-statistic):	1.58e-96			
Time:	06:24:46	Log-Likelihood:	-386.18			
No. Observations:	200	AIC:	780.4			
Df Residuals:	196	BIC:	793.6			
Df 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.			
=====						

```
[ ] import matplotlib.pyplot as plt

fig = plt.figure()
ax1 = fig.add_subplot(211)
x = df['newspaper']
prob = stats.probplot(x, dist=stats.norm, plot=ax1)
ax1.set_xlabel('')
ax1.set_title('Probplot against normal distribution')
#We now use boxcox to transform the data so it's closest to normal:
ax2 = fig.add_subplot(212)
df['newspaper'], _ = stats.boxcox(x)
prob = stats.probplot(df['newspaper'], dist=stats.norm, plot=ax2)
ax2.set_title('Probplot after Box-Cox transformation')

plt.show()
```



I 회귀분석

SAS® Studio

▶ 서버 파일 및 폴더

▼ 작업 및 유틸리티

내 작업

작업

▶ 데이터

▶ 그래프

▶ 지도

▶ 통계량

▶ 선형 모델

일원분산분석

비모수 일원분산분석

다원 ANOVA

공분산분석

선형회귀

이진 로지스틱 회귀

예측회귀모델

일반화선형모델

혼합모형

부분최소제곱회귀

*프로세스 플로우 1

실행 | 코드 생성

플로우 | 결과 | 속성

색상

모두 선택

ADVERTISING

선형회귀

선형회귀 작업은 선형 모델을 적합하여 하나

I 회귀분석

*프로세스 플로우 1 x

프로세스 플로우 1 > 선형회귀

설정 코드/결과 분할

데이터 모델 옵션

데이터

AD.AVERTISING

필터: (없음)

역할

*종속변수: (1개 항목)

sales

분류변수:

칼럼

연속변수:

TV

radio

newspaper

선택 출력 정보

출력 데이터셋

☒ 관측값 방향 통계량 데이터셋 생성

*데이터셋 이름:

work.Reg_stats 찾아보기

예측값

☐ 예측값

☐ 개별 예측값에 대한 신뢰구간

☐ 평균 예측값에 대한 신뢰구간

잔차

☐ 잔차

☒ 스튜던트화 잔차

☐ 현재 관측값이 제거된 스튜던트화 잔차

☐ Press 통계량

영향 통계량

I 회귀분석

*프로세스 플로우 1 x

프로세스 플로우 1 > 선형회귀

설정 코드/결과 분할

데이터 모델 옵션

데이터

AD.ADVERTISING

필터: (없음)

역할

*종속변수: (1개 항목)

sales

분류변수:

칼럼

연속변수:

TV

radio

newspaper

선택 출력 정보

출력 데이터셋

☒ 관측값 방향 통계량 데이터셋 생성

*데이터셋 이름:

work.Reg_stats [찾아보기](#)

예측값

☐ 예측값

☐ 개별 예측값에 대한 신뢰구간

☐ 평균 예측값에 대한 신뢰구간

잔차

☐ 잔차

☒ 스튜던트화 잔차

☐ 현재 관측값이 제거된 스튜던트화 잔차

☐ Press 통계량

영향 통계량

*프로세스 플로우 1 x

프로세스 플로우 1 > 선형회귀

설정 코드/결과 분할

데이터 모델 옵션 선택

모델 효과

모델 효과

편집

절편

TV

radio

newspaper

I 회귀분석

```
[ ] import statsmodels.formula.api as sm
model1 = sm.ols(formula="sales~TV+radio+newspaper", data=df).fit()
#sales~TV+radio+newspaper
print(model1.summary())
```

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:	Tue, 10 Mar 2020	Prob (F-statistic):	1.58e-96
Time:	06:24:46	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.

Model: MODEL1
Dependent Variable: sales

Number of Observations Read	200
Number of Observations Used	200

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4860.32349	1620.10783	570.27	<.0001
Error	196	556.82526	2.84095		
Corrected Total	199	5417.14875			

Root MSE	1.68551	R-Square	0.8972
Dependent Mean	14.02250	Adj R-Sq	0.8956
Coeff Var	12.02004		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	2.93889	0.31191	9.42	<.0001
TV	1	0.04576	0.00139	32.81	<.0001
radio	1	0.18853	0.00861	21.89	<.0001
newspaper	1	-0.00104	0.00587	-0.18	0.8599

I 변수선택

*프로세스 플로우 1

프로세스 플로우 1 > 선형회귀

설정

코드/결과

분할

선택

출력

정보

모델 선택

선택 방법:

단계별 선택

다음으로 효과 추가/제거:

Schwarz Bayesian 정보기준

다음으로 효과 추가/제거 중지:

기본 기준

최적 모델 선택변수:

기본 기준

선택 통계량

선택 도표

상세 정보

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	2.921100	0.294490	9.92	<.0001
TV	1	0.045755	0.001390	32.91	<.0001
radio	1	0.187994	0.008040	23.38	<.0001