

Chapter 5.

상관분석과 회귀분석

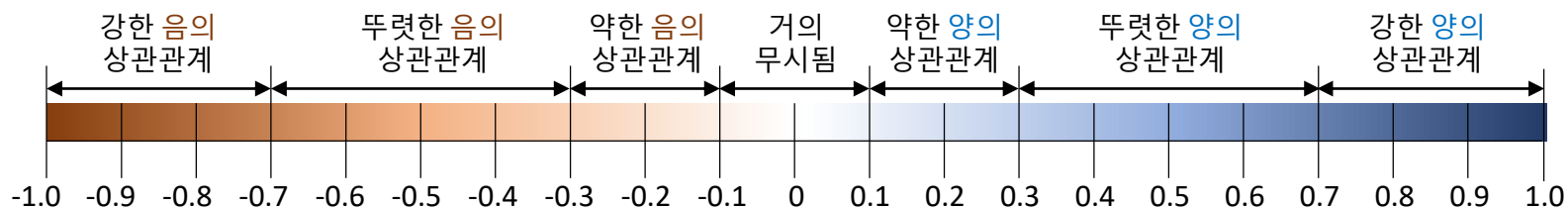
1. 상관분석
2. 회귀분석

1-1. 상관분석

1. 상관분석

■ 상관분석이란?

- 상관분석(Correlation analysis)은 연속형인 두 변수 간에 어떤 선형적인(Linear) 또는 비선형적인(Non-linear) 관계를 갖고 있는지 분석하는 방법
- 두 변수 간의 관계는 **상관계수(Correlation coefficient)**로 나타내고, **-1~1 사이의 값을 갖으며 (-)부호일 경우 반비례 관계인 음의 상관관계, (+)부호일 경우 비례 관계인 양의 상관관계를 나타냄**
- 일반적인 판정기준은 아래와 같으며 상관계수는 두 변수 간의 연관된 정도를 나타내는 것이지 인과관계를 설명하는 것은 아님



1-2. 데이터 불러와서 확인하기

1. 상관분석

▣ 데이터 불러와서 확인

- 분석 목적 : 병아리의 성장에 영향을 미치는 인자 확인
- ch5-1.csv 데이터 셋의 경우 부화한 지 1주일 된 병아리 몸무게(weight), 종란 무게(egg_weight), 하루 평균 이동거리(movement), 하루 평균 사료 섭취량(food) 데이터가 포함되어 있으며 총 5개의 열(변수)과 30개의 행으로 구성되어있음

코딩실습																																					
In [1] :	<i># pandas 패키지 불러오기 및 pd라는 약어로 지칭하기</i> import pandas as pd																																				
In [2] :	w = pd.read_csv('ch5-1.csv') <i># w 변수에 데이터셋 입력</i>																																				
In [3] :	w.head() <i># 위에서 부터 5개 데이터 확인</i>																																				
Out [3] :	<table><tr><th></th><th>chick_nm</th><th>weight</th><th>egg_weight</th><th>movement</th><th>food</th></tr><tr><td>0</td><td>a01</td><td>140</td><td>65</td><td>146</td><td>14</td></tr><tr><td>1</td><td>a02</td><td>128</td><td>62</td><td>153</td><td>12</td></tr><tr><td>2</td><td>a03</td><td>140</td><td>65</td><td>118</td><td>13</td></tr><tr><td>3</td><td>a04</td><td>135</td><td>65</td><td>157</td><td>13</td></tr><tr><td>4</td><td>a05</td><td>145</td><td>69</td><td>157</td><td>13</td></tr></table>		chick_nm	weight	egg_weight	movement	food	0	a01	140	65	146	14	1	a02	128	62	153	12	2	a03	140	65	118	13	3	a04	135	65	157	13	4	a05	145	69	157	13
	chick_nm	weight	egg_weight	movement	food																																
0	a01	140	65	146	14																																
1	a02	128	62	153	12																																
2	a03	140	65	118	13																																
3	a04	135	65	157	13																																
4	a05	145	69	157	13																																

In [4] :	<pre>w.info() # 데이터 구조 및 자료형 확인</pre>
Out [4] :	<pre><class 'pandas.core.frame.DataFrame'> RangeIndex: 30 entries, 0 to 29 Data columns (total 5 columns): # Column Non-Null Count Dtype --- - 0 chick_nm 30 non-null object 1 weight 30 non-null int64 2 egg_weight 30 non-null int64 3 movement 30 non-null int64 4 food 30 non-null int64 dtypes: int64(4), object(1) memory usage: 1.3+ KB</pre>

1-3. 상관분석을 위한 별도 데이터 셋 만들기

1. 상관분석

▣ 별도 데이터 셋 구성

- 첫 번째 열(chick_nm)이 문자(object)라 상관분석이 되지 않기 때문에 첫 번째 열을 제외하고 별도의 데이터 셋 구성

코딩실습

```
In [5]: # w 데이터셋에서 1~4열 데이터만 가져오기  
w_n = w.iloc[:,1:5]  
w_n.head()
```

```
Out [5]:
```

	weight	egg_weight	movement	food
0	140	65	146	14
1	128	62	153	12
2	140	65	118	13
3	135	65	157	13
4	145	69	157	13

1-4. 상관분석 실시

1. 상관분석

■ 상관분석

- 첫 번째 열(chick_nm)이 문자(object)라 상관분석이 되지 않기 때문에 첫 번째 열을 제외하고 별도의 데이터 셋 구성

코딩실습																										
In [6] :	<pre># 상관분석 실시 w_cor = w_n.corr(method = 'pearson') w_cor</pre>																									
Out [6] :	<table><tr><th></th><th>weight</th><th>egg_weight</th><th>movement</th><th>food</th></tr><tr><th>weight</th><td>1.000000</td><td>0.957169</td><td>0.380719</td><td>0.877574</td></tr><tr><th>egg_weight</th><td>0.957169</td><td>1.000000</td><td>0.428246</td><td>0.808147</td></tr><tr><th>movement</th><td>0.380719</td><td>0.428246</td><td>1.000000</td><td>0.319011</td></tr><tr><th>food</th><td>0.877574</td><td>0.808147</td><td>0.319011</td><td>1.000000</td></tr></table>		weight	egg_weight	movement	food	weight	1.000000	0.957169	0.380719	0.877574	egg_weight	0.957169	1.000000	0.428246	0.808147	movement	0.380719	0.428246	1.000000	0.319011	food	0.877574	0.808147	0.319011	1.000000
	weight	egg_weight	movement	food																						
weight	1.000000	0.957169	0.380719	0.877574																						
egg_weight	0.957169	1.000000	0.428246	0.808147																						
movement	0.380719	0.428246	1.000000	0.319011																						
food	0.877574	0.808147	0.319011	1.000000																						

상관계수	대상
피어슨	수치적 데이터 변수의 상관 분석(선형관계)
스피어만	순서적 데이터 변수의 상관 분석(비선형관계)

∴ 병아리 몸무게(weight)와 가장 상관관계가 큰 변수는 종란 무게(egg_weight)로 0.957의 상관계수를 갖고 있으며 다음으로 하루 평균 사료 섭취량(food), 하루 평균 이동거리(movement) 순임

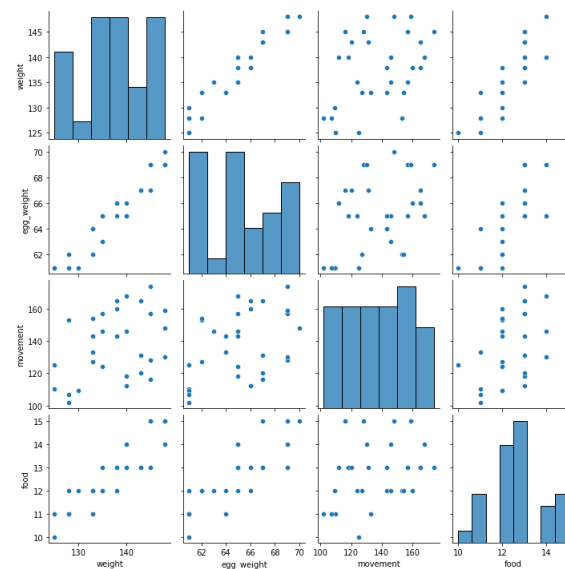
1-5. 상관분석 결과 표현하기

1. 상관분석

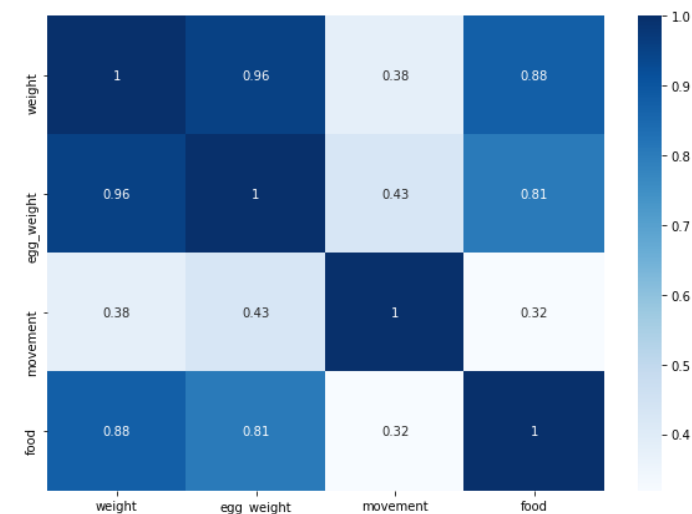
상관분석 결과 시각화

- 상관분석의 결과는 주로 산점도(Scatter plot)도 나타내며 seaborn 패키지 pairplot() 함수 이용

코딩실습	
In [7] :	<pre># 상관관계 시각화를 위한 패키지 불러오기 import matplotlib.pyplot as plt import seaborn as sns</pre>
In [8] :	<pre># 산점도 그리기 sns.pairplot(w_n)</pre>
In [9] :	<pre># 상관행렬도 그리기 plt.figure(figsize = (10,7)) sns.heatmap(w_cor, annot = True, cmap = 'Blues') plt.show()</pre>



<변수 간의 1:1 관계를 표현한 산점도>



<상관행렬도>

Chapter 5.

상관분석과 회귀분석

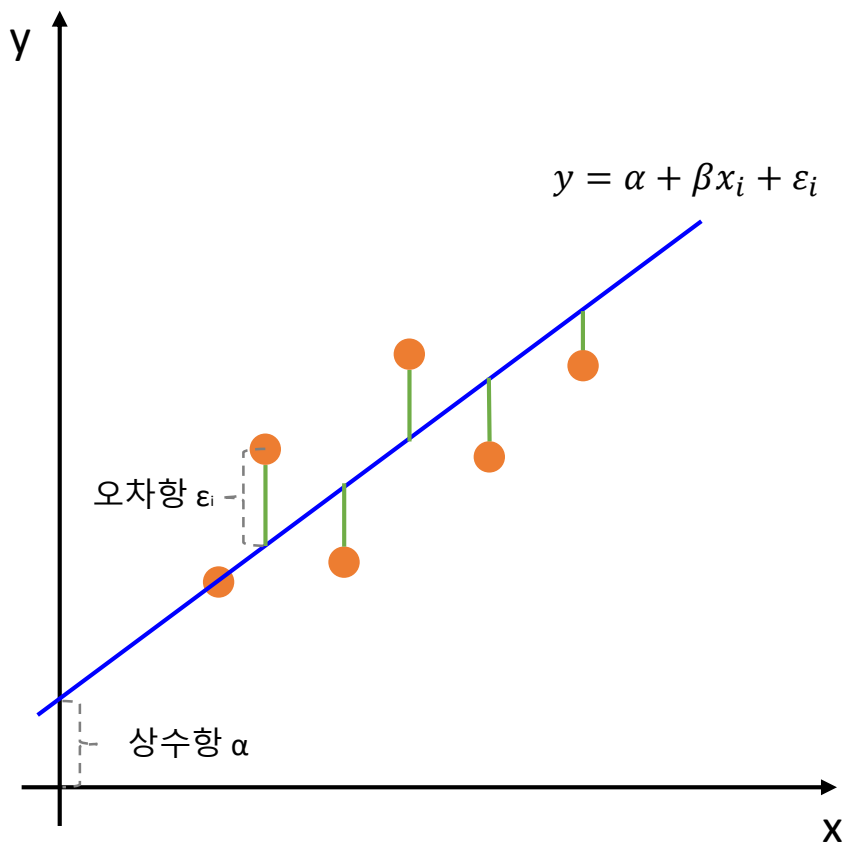
1. 상관분석
2. 회귀분석

2-1. 회귀분석이란?

2. 회귀분석

회귀분석

- 회귀분석(Regression analysis)은 연속형 변수들에 대해 두 변수 사이의 관계를 수식으로 나타내는 분석 방법
- x라는 독립변수, y라는 종속변수가 존재할 때 이 두 변수 사이의 관계를 $y = ax + b$ 와 같은 형태의 수식으로 나타낼 수 있는 방법



<회귀모형 개념>

※ 회귀분석의 5가지 가정

- 선형성 : 독립변수(x)와 종속변수(y)의 관계가 선형관계가 있음
- 독립성 : 잔차(Residual)와 독립변수의 값이 관련없어야 함
- 등분산성 : 독립변수의 모든 값에 대한 오차들의 분산이 일정해야 함
- 비상관성 : 관측치들의 잔차들끼리 상관이 없어야 함
- 정상성 : 잔차항이 정규분포를 이뤄야 함

2-2. 단순 선형 회귀분석

2. 회귀분석

▣ 단순 선형 회귀분석

- 분석 목적 : 병아리의 몸무게 예측 모델 개발
- 병아리 몸무게(weight)와 가장 상관계수가 컸던 종란 무게(egg_weight)로 statsmodels.formula.api 모듈의 ols() 함수 이용
- 회귀분석 결과는 summary() 메소드를 이용해 확인

코딩실습

```
In [10]: # 회귀분석 수행을 위한 모듈 불러오기 및 smf로 지칭하기
import statsmodels.formula.api as smf
# 종란무게 - 병아리 몸무게 단순선형회귀모델 구축
model_lm = smf.ols(formula = 'weight ~ egg_weight', data = w_n)
```

```
In [11]: # 모델 학습
result_lm = model_lm.fit()
```

```
In [12]: # 모델 결과 확인
result_lm.summary()
```

2-2. 단순 선형 회귀분석

2. 회귀분석

■ 단순 선형 회귀분석

■ 회귀분석 결과 해석

OLS Regression Results

Dep. Variable:	weight	R-squared:	0.916
Model:	OLS	Adj. R-squared:	0.913
Method:	Least Squares	F-statistic:	306.0
Date:	Tue, 01 Mar 2022	Prob (F-statistic):	1.32e-16
Time:	18:32:04	Log-Likelihood:	-63.148
No. Observations:	30	AIC:	130.3
Df Residuals:	28	BIC:	133.1
Df Model:	1		

Covariance Type: nonrobust

	coef	std err	t	P> t	[0.025	0.975]
Intercept	-14.5475	8.705	-1.671	0.106	-32.380	3.285
egg_weight	2.3371	0.134	17.493	0.000	2.063	2.611

Omnibus:	15.078	Durbin-Watson:	1.998
Prob(Omnibus):	0.001	Jarque-Bera (JB):	2.750
Skew:	0.032	Prob(JB):	0.253
Kurtosis:	1.518	Cond. No.	1.51e+03

[결과 해석]

① 회귀모델이 통계적으로 유의한지 확인

: F통계량의 p-value(p값)가 0.05 보다 작으면 유의수준 5%(신뢰수준 95%) 하에서 추정된 회귀모델이 통계적으로 유의한 것으로 판단 → 1.32e-16(1.32×10^{-16})으로 통계적으로 유의

② 개별 독립변수가 통계적으로 유의한지 확인

: 개별 독립변수의 p값이 0.05 보다 작으면 유의수준 5% 하에서 통계적으로 유의한 것으로 판단
→ 종란 무게(egg_weight)가 0으로 통계적으로 유의하다고 판단하며 상수(Intercept)의 p값은 의미 없음

③ 결정계수(R-squared)가 높은지 확인

: R^2 는 1에 가까울수록 회귀모델의 성능(설명력)이 뛰어나다고 판단 → 0.9162로 매우 높음

④ 회귀모델은 coef(coefficient)값으로 구할 수 있음

: Intercept는 y절편(상수)을 뜻하며 각 독립변수에 해당되는 coef값은 해당 독립변수의 계수(기울기)를 나타냄

→ $\text{weight} = 2.3371 \times \text{egg_weight} - 14.5475$

2-2. 단순 선형 회귀분석

2. 회귀분석

▣ 단순 선형 회귀분석

- print() 함수 이용 시 보고서 형태로 출력 가능

코딩실습

```
In [13]: # 보고서 형태로 모델 결과 출력  
print(result_lm.summary())
```

OLS Regression Results						
Dep. Variable:	weight	R-squared:	0.916			
Model:	OLS	Adj. R-squared:	0.913			
Method:	Least Squares	F-statistic:	306.0			
Date:	Tue, 01 Mar 2022	Prob (F-statistic):	1.32e-16			
Time:	18:32:04	Log-Likelihood:	-63.148			
No. Observations:	30	AIC:	130.3			
Df Residuals:	28	BIC:	133.1			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	-14.5475	8.705	-1.671	0.106	-32.380	3.285
egg_weight	2.3371	0.134	17.493	0.000	2.063	2.611
Omnibus:	15.078	Durbin-Watson:	1.998			
Prob(Omnibus):	0.001	Jarque-Bera (JB):	2.750			
Skew:	0.032	Prob(JB):	0.253			
Kurtosis:	1.518	Cond. No.	1.51e+03			

Notes:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 1.51e+03. This might indicate that there are strong multicollinearity or other numerical problems.

2-2. 단순 선형 회귀분석

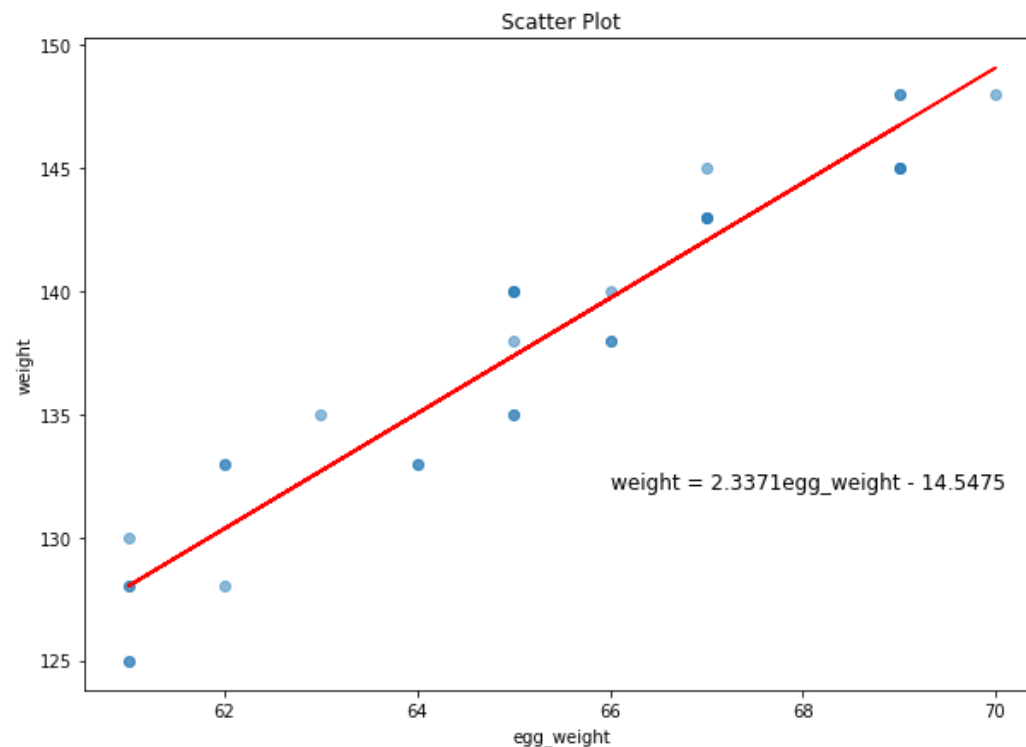
2. 회귀분석

■ 단순 선형 회귀분석 결과 확인 - 산점도

- summary() 메소드를 이용한 결과해석 외에도 산점도나 잔차의 히스토그램을 그려서 모델이 데이터를 잘 대표하는지 확인

코딩실습

```
In [14]: # 종란무게에 따른 병아리 몸무게 산점도
plt.figure(figsize = (10,7))
plt.scatter(w.egg_weight, w.weight, alpha = .5)
plt.plot(w.egg_weight, w.egg_weight*2.3371 - 14.5475, color = 'red')
plt.text(66, 132, 'weight = 2.3371egg_weight - 14.5475', fontsize = 12)
plt.title('Scatter Plot')
plt.xlabel('egg_weight')
plt.ylabel('weight')
plt.show()
```



➔ 빨간색 회귀직선이 데이터의 분포를 잘 대표하고 있음

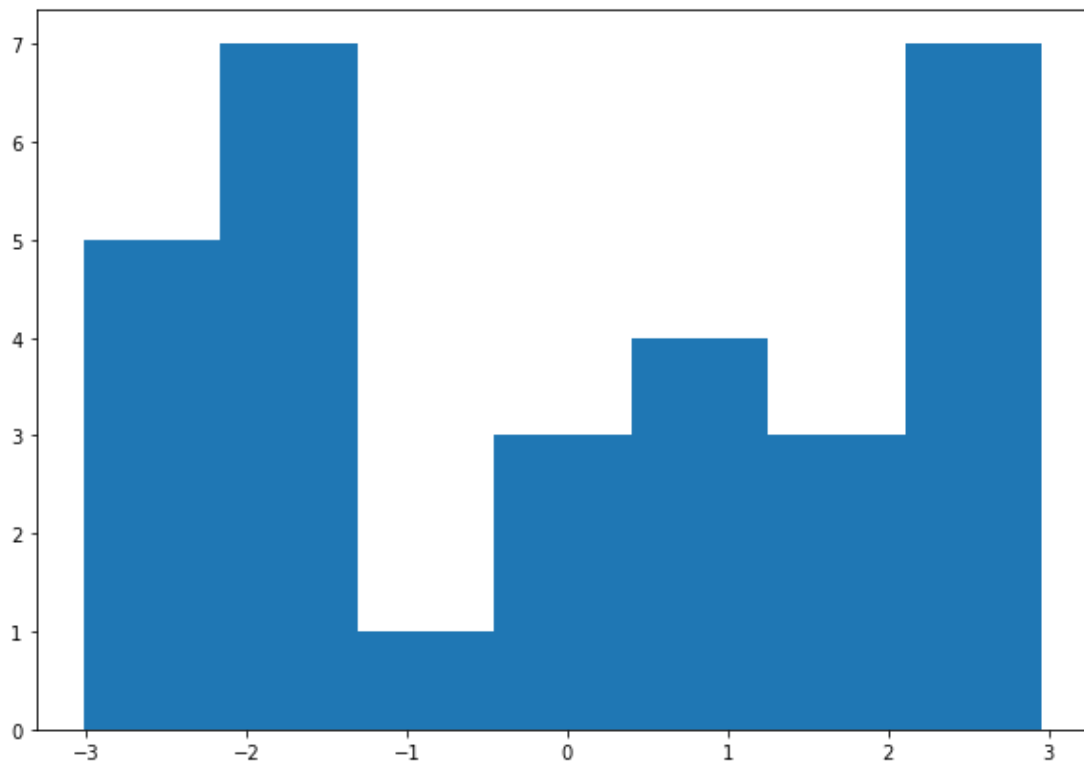
2-2. 단순 선형 회귀분석

2. 회귀분석

▣ 단순 선형 회귀분석 결과 확인 - 잔차 히스토그램

- result_lm의 속성에 잔차(residual)를 포함하고 있어 이를 이용해 잔차 히스토그램을 그릴 수 있음

코딩실습	
In [15] :	<pre># 잔차 5개만 확인 result_lm.resid.head()</pre>
Out [15] :	<pre>0 2.633714 1 -2.354880 2 2.633714 3 -2.366286 4 -1.714829 dtype: float64</pre>
In [16] :	<pre># 잔차 히스토그램 그리기 plt.figure(figsize = (10,7)) plt.hist(result_lm.resid, bins = 7) plt.show()</pre>



➔ 잔차(실제값 - 계산된 값) 결과가 0을 중심으로 종모양의 분포가 나오지 않아 다소 아쉬운 결과임

2-3. 다중 회귀분석

2. 회귀분석

▣ 다중 회귀분석 실시

- ols() 함수에 모든 변수를 추가해서 다중 회귀분석 실시

코딩실습

In [17] :	<i># 병아리 몸무게 예측을 위한 다중회귀분석 실시</i> model_mlm = smf.ols(formula = 'weight ~ egg_weight + food + movement', data = w_n)
In [18] :	result_mlm = model_mlm.fit()
In [19] :	result_mlm.summary()

2-3. 다중 회귀분석

2. 회귀분석

다중 회귀분석 실시

다중 회귀분석 결과 해석

OLS Regression Results

Dep. Variable:	weight	③ R-squared:	0.948			
Model:	OLS	Adj. R-squared:	0.942			
Method:	Least Squares	① F-statistic:	157.7			
Date:	Tue, 01 Mar 2022	Prob (F-statistic):	8.46e-17			
Time:	18:32:05	Log-Likelihood:	-56.008			
No. Observations:	30	AIC:	120.0			
Df Residuals:	26	BIC:	125.6			
Df Model:	3					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	2.9748	8.587	0.346	0.732	-14.676	20.626
egg_weight	1.7763	0.195	9.117	0.000	1.376	2.177
food	1.5847	0.405	3.915	② 0.001	0.753	2.417
movement	-0.0087	0.017	-0.522	0.606	-0.043	0.026
Omnibus:	1.993	Durbin-Watson:	2.030			
Prob(Omnibus):	0.369	Jarque-Bera (JB):	1.746			
Skew:	-0.480	Prob(JB):	0.418			
Kurtosis:	2.311	Cond. No.	4.31e+03			

[결과 해석]

① 회귀모델이 통계적으로 유의한지 확인

: F통계량의 p-value(p값)가 0.05 보다 작으면 유의수준 5%(신뢰수준 95%) 하에서 추정된 회귀모델이 통계적으로 유의한 것으로 판단 → 8.46e-17(8.46×10^{-17})으로 통계적으로 유의

② 개별 독립변수가 통계적으로 유의한지 확인

: 개별 독립변수의 p값이 0.05 보다 작으면 유의수준 5% 하에서 통계적으로 유의한 것으로 판단 → 종란 무게(egg_weight)와 하루 평균 사료 섭취량(food)만 통계적으로 유의

③ 조정된 결정계수(Adjusted R-squared)가 높은지 확인

: R^2 는 1에 가까울수록 회귀모델의 성능(설명력)이 뛰어나다고 판단 → 0.942로 매우 높음

④ 회귀모델은 coef(coefficient)값으로 구할 수 있음

: Intercept는 y절편(상수)을 뜻하며 각 독립변수에 해당되는 coef값은 해당 독립변수의 계수(기울기)를 나타냄

→ $weight = 1.7763 \cdot egg_weight - 0.0087 \cdot movement + 1.5847 \cdot food + 2.9748$

2-3. 다중 회귀분석

2. 회귀분석

▣ 다중 회귀분석 실시 – 유의하지 않은 독립변수 제거

- p값이 0.05보다 높아 유의수준 5%에서 유의하지 않은 독립변수인 하루 평균 이동거리(movement) 변수를 제외하고 다중 회귀분석 실시

코딩실습	
In [20] :	# 병아리 몸무게 예측을 위한 다중회귀분석 실시2 model_mlm2 = smf.ols(formula = 'weight ~ egg_weight + food', data = w_n)
In [21] :	result_mlm2 = model_mlm2.fit()
In [22] :	result_mlm2.summary()

2-3. 다중 회귀분석

2. 회귀분석

다중 회귀분석 실시 - 유의하지 않은 독립변수 제거

다중 회귀분석 결과 해석

OLS Regression Results

Dep. Variable:	weight	③ R-squared:	0.947
Model:	OLS	Adj. R-squared:	0.943
Method:	Least Squares	① F-statistic:	243.0
Date:	Tue, 01 Mar 2022	Prob (F-statistic):	5.44e-18
Time:	18:32:05	Log-Likelihood:	-56.164
No. Observations:	30	AIC:	118.3
Df Residuals:	27	BIC:	122.5
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	3.6638	8.370	0.438	0.665	-13.510	20.837
egg_weight	1.7453	0.183	9.536	0.000	1.370	2.121
food	1.5955	0.399	4.001	0.000	0.777	2.414

Omnibus:	2.302	Durbin-Watson:	2.103
Prob(Omnibus):	0.316	Jarque-Bera (JB):	1.940
Skew:	-0.502	Prob(JB):	0.379
Kurtosis:	2.263	Cond. No.	1.84e+03

[결과 해석]

① 회귀모델이 통계적으로 유의한지 확인

: F통계량의 p-value(p값)가 0.05 보다 작으면 유의수준 5%(신뢰수준 95%) 하에서 추정된 회귀모델이 통계적으로 유의한 것으로 판단 → $2.2e-16(2.2 \times 10^{-16})$ 으로 통계적으로 유의

② 개별 독립변수가 통계적으로 유의한지 확인

: 개별 독립변수의 p값이 0.05 보다 작으면 유의수준 5% 하에서 통계적으로 유의한 것으로 판단
→ 종란 무게(egg_weight)와 하루 평균 사료 섭취량(food) 통계적으로 유의

③ 조정된 결정계수(Adjusted R-squared)가 높은지 확인

: R^2 는 1에 가까울수록 회귀모델의 성능(설명력)이 뛰어나다고 판단
→ 0.9435로 movement 변수가 있을 때 보다 더 높아졌음

④ 회귀모델은 coef(coefficient)값으로 구할 수 있음

: Intercept는 y절편(상수)을 뜻하며 각 독립변수에 해당되는 coef값은 해당 독립변수의 계수(기울기)를 나타냄

→ $weight = 1.7453 \times egg_weight + 1.5955 \times food + 3.6638$

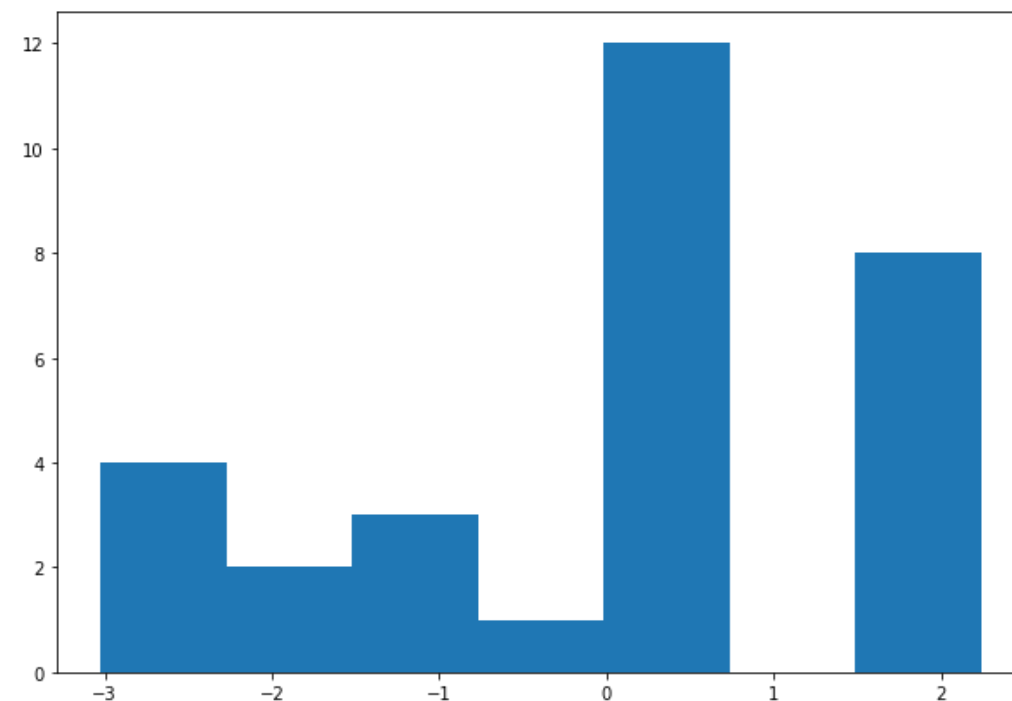
2-4. 다중공선성

2. 회귀분석

▣ 다중공선성

- 다중 회귀분석에서 많은 독립변수들 간의 강한 상관관계가 발생할 경우 다중공선성(Multicollinearity) 문제가 있다고 함
- 다중공선성 문제는 분산팽창요인(VIF, Variance Inflation Factor)을 계산해 구할 수 있으며 10 이상이면 있다고 판단하며 30을 초과하면 심각함
- statsmodels.stats.outliers_influence 모듈에서 variance_inflation_factor() 함수 이용

코딩실습	
In [23] :	<pre># 다중공선성 확인을 위한 함수 불러오기 from statsmodels.stats.outliers_influence import variance_inflation_factor</pre>
In [24] :	<pre># 회귀모델 외생변수이름 속성 model_mlm2.exog_names</pre>
Out [24] :	<pre>['Intercept', 'egg_weight', 'food']</pre>
In [25] :	<pre># 첫 번째 변수(egg_weight) vif 계산 vif1 = variance_inflation_factor(model_mlm2.exog, 1)</pre>
In [26] :	<pre># 두 번째 변수(food) vif 계산 vif2 = variance_inflation_factor(model_mlm2.exog, 2)</pre>
In [27] :	<pre>print(vif1, vif2)</pre>
In [28] :	<pre># 잔차 히스토그램 그리기 plt.figure(figsize = (10,7)) plt.hist(result_mlm2.resid, bins = 7) plt.show()</pre>



➔ 단순 선형 회귀분석 대비 잔차의 히스토그램이
0을 중심으로 한 형태의 종모양에 가까워졌음

∴ 두 변수 모두 10보다 매우 작기 때문에 다중공선성 문제는 없는 것으로 판단됨

▣ 데이터 불러오기 및 확인하기

- 분석 목적 : 성장기간에 따른 병아리의 몸무게 변화 예측 모델 개발
- 독립변수(x)와 종속변수(y)가 비선형 관계일 때 사용하는 분석 방법으로 독립변수에 로그(log)나 거듭제곱 등을 취해서 적합한 모델을 찾아야함
- ch5-2.csv 데이터 셋은 성장 기간(일자)에 따른 병아리의 몸무게(weight) 변화를 기록한 데이터로 총 70일간 관찰하였음

코딩실습																			
In [29] :	w2 = pd.read_csv('ch5-2.csv') # w2 변수에 데이터셋 입력																		
In [30] :	w2.head()																		
Out [30] :	<table><tr><th></th><th>day</th><th>weight</th></tr><tr><td>0</td><td>1</td><td>43</td></tr><tr><td>1</td><td>2</td><td>55</td></tr><tr><td>2</td><td>3</td><td>69</td></tr><tr><td>3</td><td>4</td><td>86</td></tr><tr><td>4</td><td>5</td><td>104</td></tr></table>		day	weight	0	1	43	1	2	55	2	3	69	3	4	86	4	5	104
	day	weight																	
0	1	43																	
1	2	55																	
2	3	69																	
3	4	86																	
4	5	104																	

In [31] :	w2.info()
Out [31] :	<class 'pandas.core.frame.DataFrame'> RangeIndex: 70 entries, 0 to 69 Data columns (total 2 columns): # Column Non-Null Count Dtype --- --- 0 day 70 non-null int64 1 weight 70 non-null int64 dtypes: int64(2) memory usage: 1.2 KB

2-5. 비선형 회귀분석

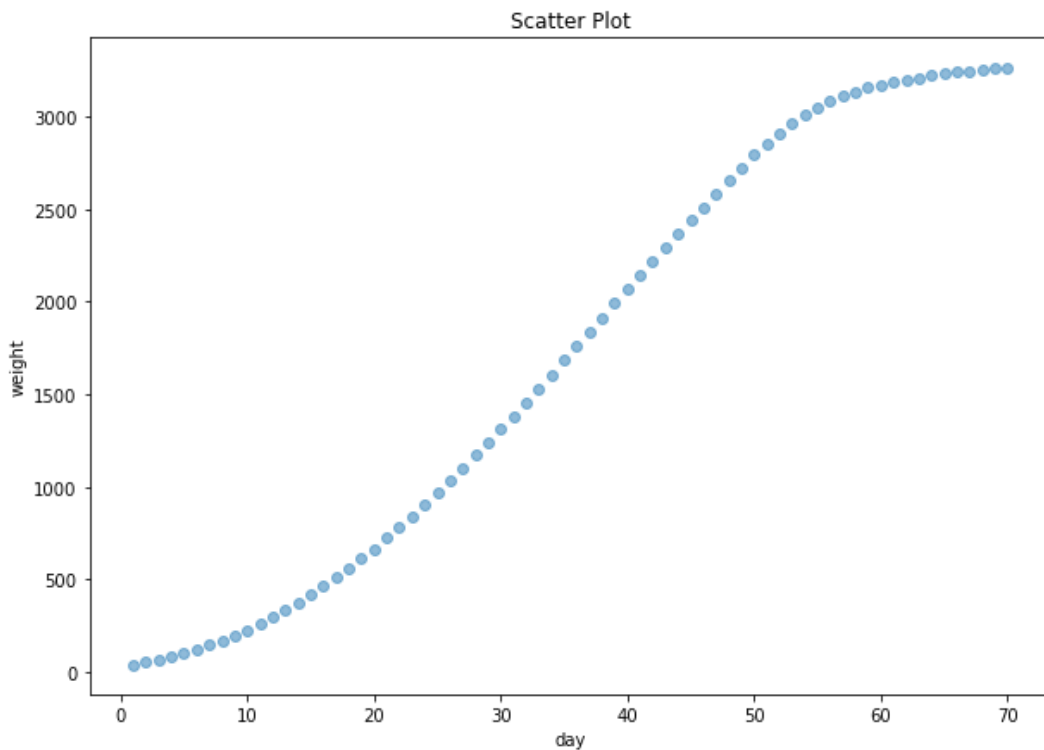
2. 회귀분석

▣ 데이터 불러오기 및 확인하기

- 성장기간에 따른 병아리 몸무게 변화 산점도 그리기

코딩실습

```
In [32]: # 성장기간에 따른 몸무게 변화
plt.figure(figsize = (10,7))
plt.scatter(w2.day, w2.weight, alpha = .5)
plt.title('Scatter Plot')
plt.xlabel('day')
plt.ylabel('weight')
plt.show()
```



<성장기간에 따른 병아리 몸무게 변화 산점도>

2-5. 비선형 회귀분석

2. 회귀분석

■ 선형 회귀분석 실시

- 선형 회귀분석으로 적합(fitting)시켜 보기

코딩실습

In [33] :	<i># 성장기간에 따른 병아리의 몸무게 변화 선형 회귀분석 실시</i> model_lm2 = smf.ols(formula = 'weight ~ day', data = w2)
In [34] :	result_lm2 = model_lm2.fit()
In [35] :	result_lm2.summary()

OLS Regression Results

Dep. Variable:	weight	R-squared:	0.979			
Model:	OLS	Adj. R-squared:	0.979			
Method:	Least Squares	F-statistic:	3189.			
Date:	Tue, 01 Mar 2022	Prob (F-statistic):	7.22e-59			
Time:	18:32:07	Log-Likelihood:	-457.86			
No. Observations:	70	AIC:	919.7			
Df Residuals:	68	BIC:	924.2			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	-295.8671	41.102	-7.198	0.000	-377.885	-213.850
day	56.8216	1.006	56.470	0.000	54.814	58.830
Omnibus:	3.866	Durbin-Watson:	0.025			
Prob(Omnibus):	0.145	Jarque-Bera (JB):	2.079			
Skew:	-0.133	Prob(JB):	0.354			
Kurtosis:	2.199	Cond. No.	82.6			

2-5. 비선형 회귀분석

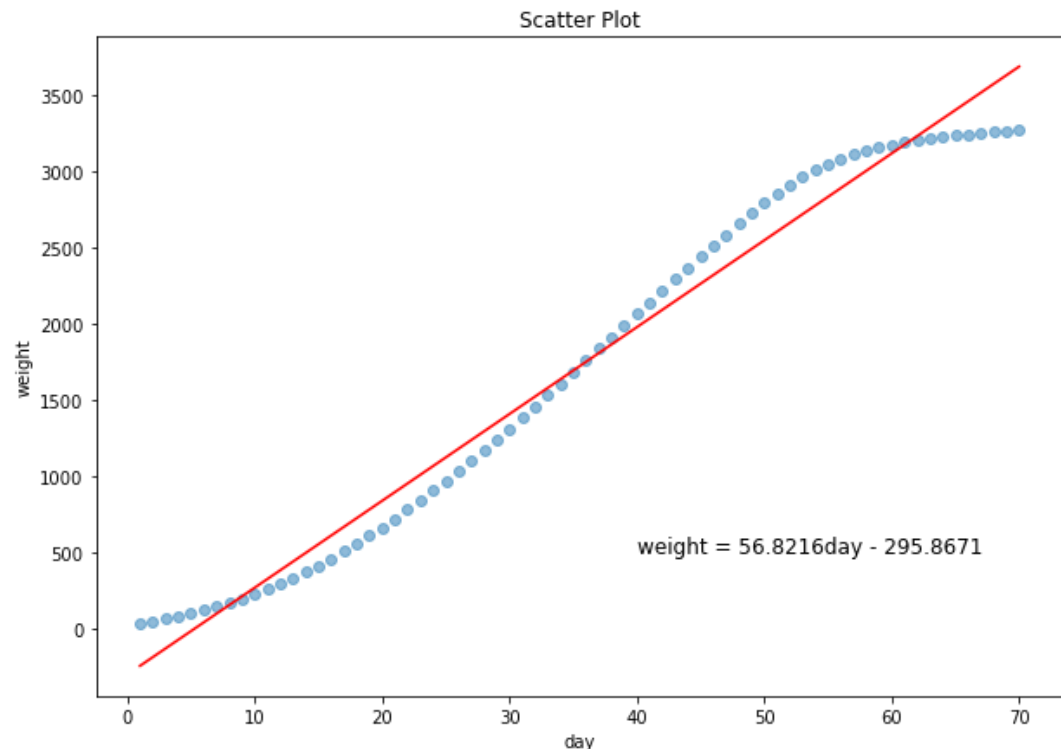
2. 회귀분석

■ 선형 회귀분석 실시

- 회귀모델 적합 결과 산점도에 표시해보기

코딩실습

```
In [36]: # 성장기간에 따른 몸무게 변화
plt.figure(figsize = (10,7))
plt.scatter(w2.day, w2.weight, alpha = .5)
plt.plot(w2.day, w2.day*56.8216 - 295.8671, color = 'red')
plt.text(40, 500, 'weight = 56.8216day - 295.8671', fontsize = 12)
plt.title('Scatter Plot')
plt.xlabel('day')
plt.ylabel('weight')
plt.show()
```



<성장기간에 따른 병아리 몸무게 변화 산점도 및 회귀직선>

∴ 회귀모델 및 개별 독립변수 모두 95% 신뢰수준에서 통계적으로 유의하며 R^2 도 0.9791로 매우 높으나 회귀직선의 형태가 적합하지 않음

2-5. 비선형 회귀분석

2. 회귀분석

■ 비선형 회귀분석 실시

- 비선형 회귀분석으로 적합(fitting)시켜 보기

코딩실습

```
In [37]: # 성장기간에 따른 병아리의 몸무게 변화 비선형 회귀분석 실시
model_nlm = smf.ols(formula = 'weight ~ l(day**3) + l(day**2) + day', data = w2)

In [38]: result_nlm = model_nlm.fit()

In [39]: result_nlm.summary()
```

OLS Regression Results

Dep. Variable:	weight	R-squared:	1.000			
Model:	OLS	Adj. R-squared:	0.999			
Method:	Least Squares	F-statistic:	4.407e+04			
Date:	Tue, 01 Mar 2022	Prob (F-statistic):	7.13e-109			
Time:	18:32:07	Log-Likelihood:	-327.17			
No. Observations:	70	AIC:	662.3			
Df Residuals:	66	BIC:	671.3			
Df Model:	3					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	117.0141	13.476	8.683	0.000	90.108	143.920
l(day ** 3)	-0.0253	0.000	-51.312	0.000	-0.026	-0.024
l(day ** 2)	2.6241	0.053	49.314	0.000	2.518	2.730
day	-15.2978	1.632	-9.373	0.000	-18.557	-12.039
Omnibus:	6.702	Durbin-Watson:	0.082			
Prob(Omnibus):	0.035	Jarque-Bera (JB):	2.680			
Skew:	0.103	Prob(JB):	0.262			
Kurtosis:	2.064	Cond. No.	5.65e+05			

2-5. 비선형 회귀분석

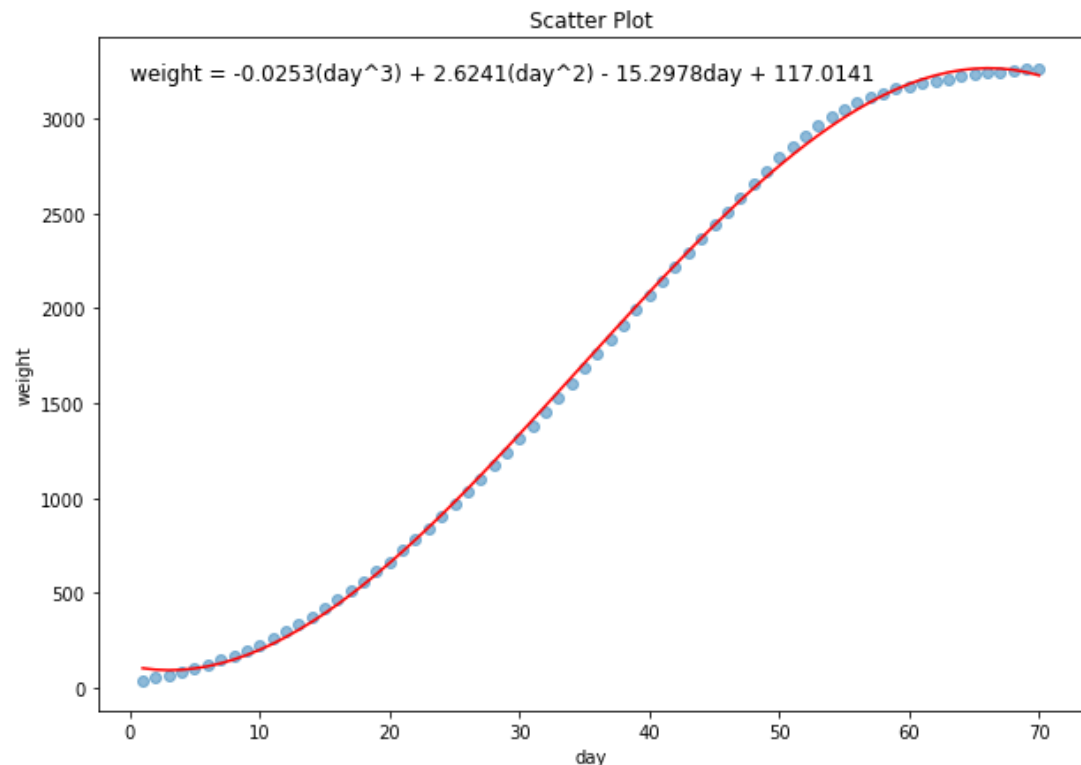
2. 회귀분석

■ 선형 회귀분석 실시

- 회귀모델 적합 결과 산점도에 표시해보기

코딩실습

```
In [40]: # 성장기간에 따른 몸무게 변화
plt.figure(figsize = (10,7))
plt.scatter(w2.day, w2.weight, alpha = .5)
plt.plot(w2.day, (w2.day**3)*(-0.0253) + (w2.day**2)*2.6241 +
w2.day*(-15.2978) + 117.0141, color = 'red')
plt.text(0, 3200, 'weight = -0.0253(day^3) + 2.6241(day^2) -
15.2978day + 117.0141', fontsize = 12)
plt.title('Scatter Plot')
plt.xlabel('day')
plt.ylabel('weight')
plt.show()
```



<성장기간에 따른 병아리 몸무게 변화 산점도 및 회귀곡선>

∴ 선형회귀모델 대비 R^2 도 1으로 매우 높아졌으며 산점도와 회귀곡선이 거의 일치함

$$weight = -0.0253*day^3 + 2.6241*day^2 - 15.2978*day + 117.0141$$