제 8장. 상관분석과 회귀분석

8.1 상관분석

상관분석: 두 변수 사이의 **선형적 연관성**에 관한 추론

X와 Y의 모상관계수:

$$\rho = Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

(X,Y)에 대한 관측값을 $(x_1,y_1),\dots,(x_n,y_n)$이라 할 때, 표본상관계수는 다음과 같다.

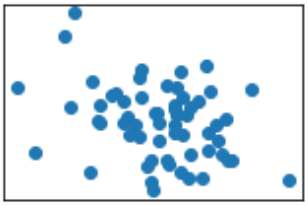
$$r = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2 \sum_{i=1}^n (y_i-\bar{y})^2}}$$

표본상관계수의 성질

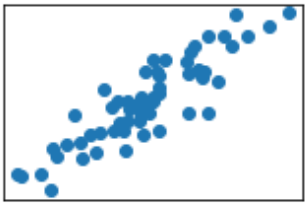
- $-1 \le r \le 1$
- $|r|$의 값이 1에 가까울수록 강한 선형 상관관계
- $|r|$의 값이 0에 가까울수록 약한 선형 상관관계

여러 가지 경우의 산점도와 표본 상관계수

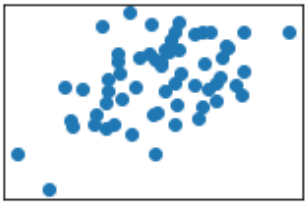
(1) r = 0



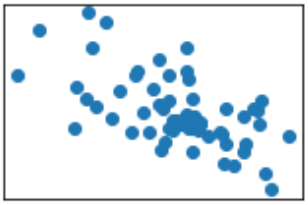
(2) r = 0.9



(3) r = 0.5



(4) r = -0.5



상관관계의 유무에 관한 검정

귀무가설 H_0: $\rho = 0$

검정통계량 $T = \frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}}$

유의수준 α에서 기각역

$$H_1 : \rho > 0 \Rightarrow T \geq t_\alpha(n-2)$$
$$H_1 : \rho < 0 \Rightarrow T \leq t_\alpha(n-2)$$
$$H_1 : \rho \neq 0 \Rightarrow |T| \geq t_{\alpha/2}(n-2)$$

[참고] 제곱합과 곱의 합의 기호와 간편계산법:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum x_i^2 - (\sum x_i)^2/n$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2/n$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - (\sum x_i)(\sum y_i)/n$$

(예) 표본상관계수

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

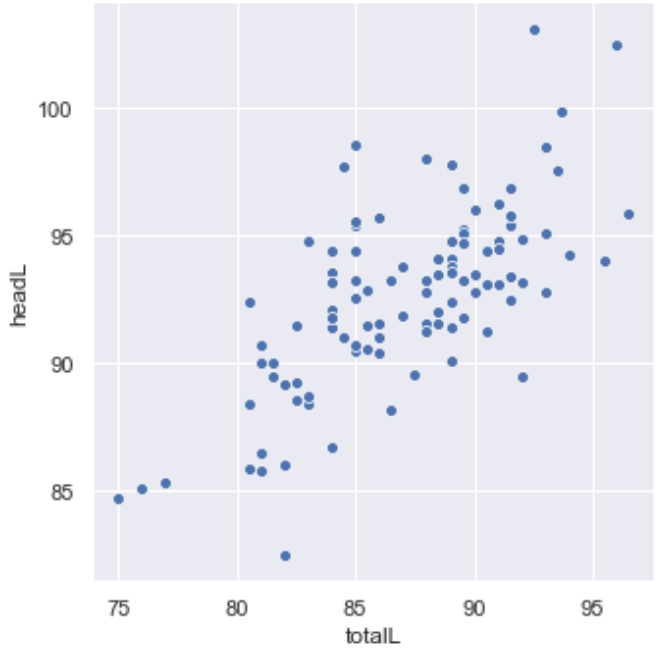
예 1) (possum.txt) 주어진 자료는 주머니쥐에 관한 조사 자료이다. 주머니쥐의 머리 둘레길이(headL)와 전체 몸통길이(totalL)의 관련성에 대해 알아보려고 한다.

먼저 두 변수에 대한 산점도를 그려보면 다음과 같다. 산점도는 seaborn 패키지의 relplot 함수를 사용하여 그릴 수 있다.

```
%cd D:/
```

```
D:₩
```

```
import pandas
from scipy import stats
import matplotlib.pyplot as plt
import seaborn

seaborn.set(rc={'figure.figsize':(3,3)})
possum = pandas.read_table("possum.txt", sep = '₩t')

seaborn.relplot(x = 'totalL', y = 'headL', data = possum)
plt.show()
```



상관관계의 유무에 대한 검정은 scipy.stats의 pearsonr 함수를 사용해서 시행할 수 있다.

```
pearsonr(x, y)
```

모수 값:

- x, y: 검정을 원하는 자료

실행 결과:

- 상관계수, 양측 유의확률

```
t, p = stats.pearsonr(possum['headL'], possum['totalL'])
print("Pearson's correlation coefficient: %s, p-value: %s" % (t, p))
```

```
Pearson's correlation coefficient: 0.6910936973935055, p-value: 4.680578654379455
e-16
```

산점도 상으로 두 변수 간에는 어느 정도 선형적 연관성이 있음을 확인할 수 있고 실제로 상관계수를 구해보면 0.691로 나타난다. 상관관계 유무의 검정 결과, 유의확률은 매우 작은 것으로 확인되었다. 따라서 유의수준 5%에서 두 변수 사이의 상관관계가 존재한다는 매우 뚜렷한 증거가 있다.

8.2 단순선형회귀모형

회귀분석(regression analysis) : 반응변수와 설명변수 사이의 함수관계를 규명

$y = f(x)$ 의 함수관계가 있을 때,

- x : 설명변수(explanatory variable) 또는 독립변수(independent variable)
- y : 반응변수(response variable) 또는 종속변수(dependent variable)

단순회귀모형(simple regression) : 설명변수가 1개인 회귀모형

중회귀모형(multiple regression) : 설명변수가 2개 이상인 회귀모형

선형회귀모형(또는 직선회귀모형) : 설명변수와 반응변수 사이에 직선관계가 있는 모형

단순선형회귀모형

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \dots, n$$

다만,

α, β: 회귀모수 (α: 상수항, β: 기울기)

$x_1, \dots, x_n$: 설명변수 (독립변수)

$y_1, \dots, y_n$: 반응변수 (종속변수)

$\epsilon_1, \dots, \epsilon_n$: 오차항. 서로 독립인 $N(0, \sigma^2)$ 확률변수

- $y = \alpha + \beta x$: 모회귀직선(population regression line)
- $\hat{y} = \hat{\alpha} + \hat{\beta} x$: 적합된 회귀직선 (fitted regression line)
- $\hat{\epsilon} = y - \hat{y}$: 잔차(residual)

최소제곱법 (least squares method):

잔차제곱합인 $\sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n (y_i - \hat{\alpha} + \hat{\beta} x_i)$ 를 최소로 하는 $\hat{\alpha}, \hat{\beta}$을 각각 α, β의 최소제곱추정량(least squares estimator)이라 하며, 이러한 추정법을 최소제곱법이라 한다.

최소제곱추정량:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

최소제곱회귀직선: $\hat{y} = \hat{\alpha} + \hat{\beta} x$

[참고]

1. $\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$에서 $\bar{y} = \hat{\alpha} + \hat{\beta} \bar{x}$가 성립하므로 최소제곱회귀직선은 언제나 $(\bar{x}, \bar{y})$를 지난다.
2. 적합된 회귀직선 $\hat{y} = \hat{\alpha} + \hat{\beta} x$에서 x가 1단위 증가하면 y는 평균 $\hat{\beta}$단위 증가한다.

제곱합의 분해

$$\begin{aligned} & (y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \\ & \Rightarrow \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\ & \Leftrightarrow SST = SSE + SSR \end{aligned}$$

- SST (총제곱합) = SSE (잔차제곱합) + SSR (회귀제곱합)
- 결정계수(coefficient of determination):

$$R^2 = \frac{SSR}{SST}$$

- 평균제곱: $MSE = \frac{SSE}{n-2}$, $MSR = \frac{SSR}{1}$
- 오차항의 분산 σ^2의 추정량: $\hat{\sigma^2} = MSE$

회귀직선의 유의성검정 :

직선관계 또는 회귀모형의 유의성을 나타내는 가설:

$$H_0 : \beta = 0, \quad H_1 : \beta \neq 0$$

H_0하에서 $F = \frac{MSR}{MSE} \sim F(1, n-2)$

분산분석표

요인	제곱합	자유도	평균제곱	F 값	유의확률
회귀	SSR	1	MSR	f=MSR/MSE	$P(F \geq f)$
잔차	SSE	$n-2$	MSE		
계	SST	$n-1$			

회귀모수에 대한 추론 :

- β에 대한 추론

1.$SE(\hat{\beta}) = \sqrt{MSE/S_{xx}}$

2.β의 $100(1-\alpha)\%$신뢰구간: $\hat{\beta} \pm t_{\alpha/2}(n-2)SE(\hat{\beta})$

3.$H_0 : \beta = \beta_0$의 검정:

검정통계량

$$T = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})}$$

- 주어진 x에서 평균반응값 $E(y) = \alpha + \beta x$에 대한 추정

1.예측값: $\hat{y} = \hat{\alpha} + \hat{\beta}x$

2.$SE(\hat{y}) = \sqrt{MSE\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)}$

3.신뢰구간: $(\hat{\alpha} + \hat{\beta}x) \pm t_{\alpha/2}(n-2)SE(\hat{y})$

예 2) (possum.txt) 머리 둘레길이(headL)과 전체 몸통길이(totalL) 간에 유의한 상관관계가 존재한다는 것을 확인했으므로 두 변수를 이용하여 단순선형 회귀모형을 적합해보자.

회귀식의 적합은 statsmodels.formula.api의 ols 함수를 사용할 수 있다. ols 함수 안에 선형회귀식의 독립변수(x: totalL)와 종속변수(y: headL)를 다음과 같이 입력한다. 상수항의 경우에는 별도로 쓰지 않아도 기본적으로 포함하게 된다. 적합 결과를 모두 확인하기 위해서는 summary함수를 사용한다.

```
from statsmodels.formula.api import ols

model = ols("headL ~ totalL", possum).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  headL   R-squared:                       0.478
Model:                            OLS   Adj. R-squared:                  0.472
Method:                 Least Squares   F-statistic:                     93.26
Date:                Wed, 15 Aug 2018   Prob (F-statistic):           4.68e-16
Time:                        14:14:54   Log-Likelihood:                -245.75
No. Observations:                 104   AIC:                             495.5
Df Residuals:                     102   BIC:                             500.8
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     42.7098      5.173      8.257      0.000      32.450      52.970
totalL         0.5729      0.059      9.657      0.000       0.455       0.691
==============================================================================
Omnibus:                        5.577   Durbin-Watson:                   1.881
Prob(Omnibus):                  0.062   Jarque-Bera (JB):                5.117
Skew:                           0.422   Prob(JB):                       0.0774
Kurtosis:                       3.684   Cond. No.                     1.77e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly s
pecified.
[2] The condition number is large, 1.77e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

summary함수의 첫 번째 출력 결과는 회귀모형의 유의성 검정 결과를 보여준다. 검정 통계량 (F - statistic)의 값은 93.26이고 유의확률 (Prob (F-statistic))은 매우 작은 것으로 나타났다. 따라서 유의수준 5%에서 회귀직선은 유의하다고 말할 수 있다. 또한 결정계수(Multiple R-squared)는 0.4776으로써 전체 자료의 산포 중 약 47.76%가 회귀직선으로 설명이 가능한 것을 알 수 있다.

두 번째 결과는 회귀 분석에 사용된 fomula를 보여준다. coefficient는 추정된 회귀계수를 나타낸다. ols 함수의 시행 결과, 적합된 회귀 모형은 $\hat{y} = 42.71 + 0.573x$ 임을 알 수 있다. 회귀계수의 통계적 유의성을 검정한 결과, 설명변수 totalL의 t-검정 통계량의 값은 9.657로써 유의확률은 매우 작다. 따라서 유의수준 5%에서 이 회귀직선은 유의하다고 말할 수 있다.

적합된 회귀 모형의 분산분석표를 출력하기 위해서는 anova_lm 함수를 사용한다. 분산분석표를 확인해보면 summary에서 출력되는 검정 통계량과 동일한 것을 볼 수 있다.

```
import statsmodels.api as sm

table = sm.stats.anova_lm(model, typ=2)
print(table)
```

	sum_sq	df	F	PR(>F)
totalL	628.148138	1.0	93.256604	4.680579e-16
Residual	687.040996	102.0	NaN	NaN

8.3 잔차분석

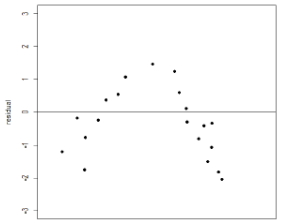
단순선형회귀모형 $y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, n$에서 전제된 주요 가정

1. 선형관계: $E(y) = \alpha + \beta_1 x$
2. 정규성: $\epsilon_i \sim N(0, \sigma^2)$
3. 등분산성: $\epsilon_1, \ldots, \epsilon_n$의 분산은 모두 σ^2
4. 독립성: $\epsilon_1, \ldots, \epsilon_n$은 서로 독립

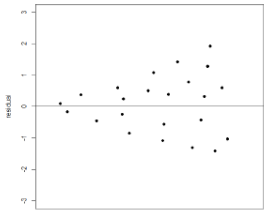
잔차분석(residual analysis): 잔차를 이용하여 모형의 가정에 대한 검토하는 것

- 오차항인 ϵ_i들이 서로 독립이고 같은 분포 $N(0, \sigma^2)$를 따른다면 잔차인 $\hat{\epsilon}_i = y_i - \hat{y}_i$는 0을 중심으로 랜덤하게 분포되어 있어야 한다.
- 이러한 잔차 분석은 통계 패키지에서 제공되는 잔차도(residual plot)을 이용해서 수행할 수 있다.

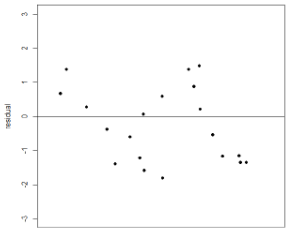
(예) 단순 선형 회귀모형의 가정에 어긋나는 경우



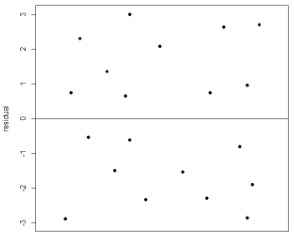
(a) 선형성에 벗어나는 경우



(b) 등분산성에 벗어나는 경우



(c) 독립성에 벗어나는 경우

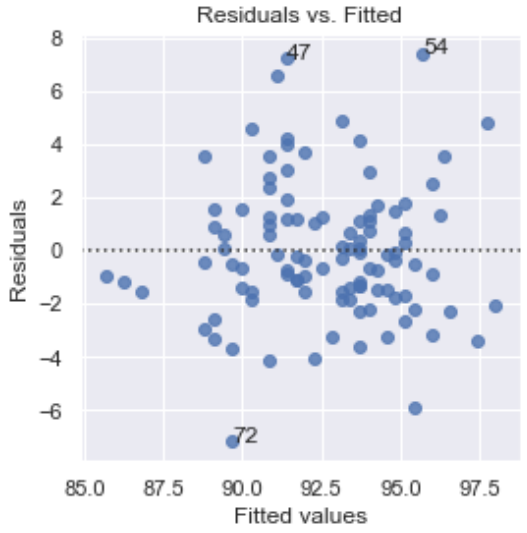


(d) 정규성에 벗어나는 경우

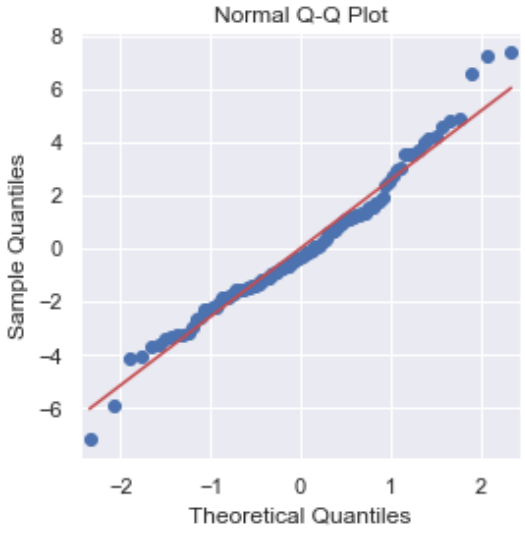
예 3) (예 2) 에서 적합한 직선에 대한 잔차 분석을 시행해보자.

예 2) 에서 적합한 회귀분석의 결과는 model에 저장되어 있다. 적합된 결과에 대한 잔차도(residual plot)와 잔차의 정규 분위수 그래프는 다음의 스크립트를 이용해서 출력할 수 있다.

```
import seaborn
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.graphics.gofplots import ProbPlot


# 1. 잔차도(residual plot): Residuals vs. Fitted

f = model.fittedvalues                # 예측값 (fitted values)
r = model.resid                       # 잔차 (residuals)
ar = np.abs(r)                        # 잔차의 절대값 (absolute residuals)

ar_sort = ar.sort_values(ascending = False)
top_3_ar = ar_sort[0:3]               # top 3 absolute residuals

seaborn.set(rc={'figure.figsize':(4,4)})
seaborn.residplot(f, r, data = possum)

plt.title("Residuals vs. Fitted")
plt.ylabel("Residuals")
plt.xlabel("Fitted values")

for i in top_3_ar.index:
    plt.annotate(i, xy = (f[i], r[i]))  # 잔차의 절대값이 큰 관측값 3개를 그래프에 표시
plt.show()
```



```
# 2. 정규 분위수 (Normal Q-Q plot)

QQ = ProbPlot(r)
plot = QQ.qqplot(line = 's', color='C0', lw=1)
plt.title("Normal Q-Q Plot")

plt.show()
```



잔차도를 확인한 결과, 잔차들은 특정한 패턴을 보이거나 등분산성 가정을 위배한 것으로 보이지는 않는다. 하지만 몇몇 잔차의 경우 범위를 벗어난 큰 값을 갖는 것으로 확인된다. 정규 분위수 그래프를 확인해 보면 양 끝쪽으로 직선에서 벗어난 점들이 관찰되는데 대부분의 잔차는 직선 주위에 몰려있는 것을 확인할 수 있다. 따라서 단순선형회귀모형의 적용은 타당함을 알 수 있다.

8.4 중회귀분석

중회귀모형

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{ki} + \epsilon_i, \quad i = 1, \ldots, n$$

다만,

$\beta_0, \beta_1, \ldots, \beta_k$: 회귀모수

$x_{1i}, \ldots, x_{ki}, i = 1, \ldots, n$: 설명변수 (독립변수)

$y_1, \ldots, y_n$: 반응변수 (종속변수)

$\epsilon_1, \ldots, \epsilon_n$: 오차항. 서로 독립인 $N(0, \sigma^2)$ 확률변수

위의 모회귀직선(population regression line)을 추정하고자 한다. 중회귀모형의 회귀계수도 단순회귀모형과 마찬가지로 최소제곱추정법을 사용하여 구한 최소 제곱추정량을 사용한다.

중회귀모형의 유의성을 검정하기 위한 가설 :

$$H_0 : \beta_1 = \ldots = \beta_k = 0, \quad H_1 : not \quad H_1$$

요인	제곱합	자유도	평균제곱	F 값	유의확률
회귀	SSR	k	MSR	f =MSR/MSE	$P(F \geq f)$
잔차	SSE	$n-k-1$	MSE		
계	SST	$n-1$			

결정계수(R-squared)와 수정결정계수(adjusted R-squared)

- 결정계수(R-squared): $R^2 = \frac{SSR}{SST}$
- 수정결정계수(adjusted R-squared): $R^2_{adj} = 1 - \frac{n-1}{n-k-1}\frac{SSE}{SST}$

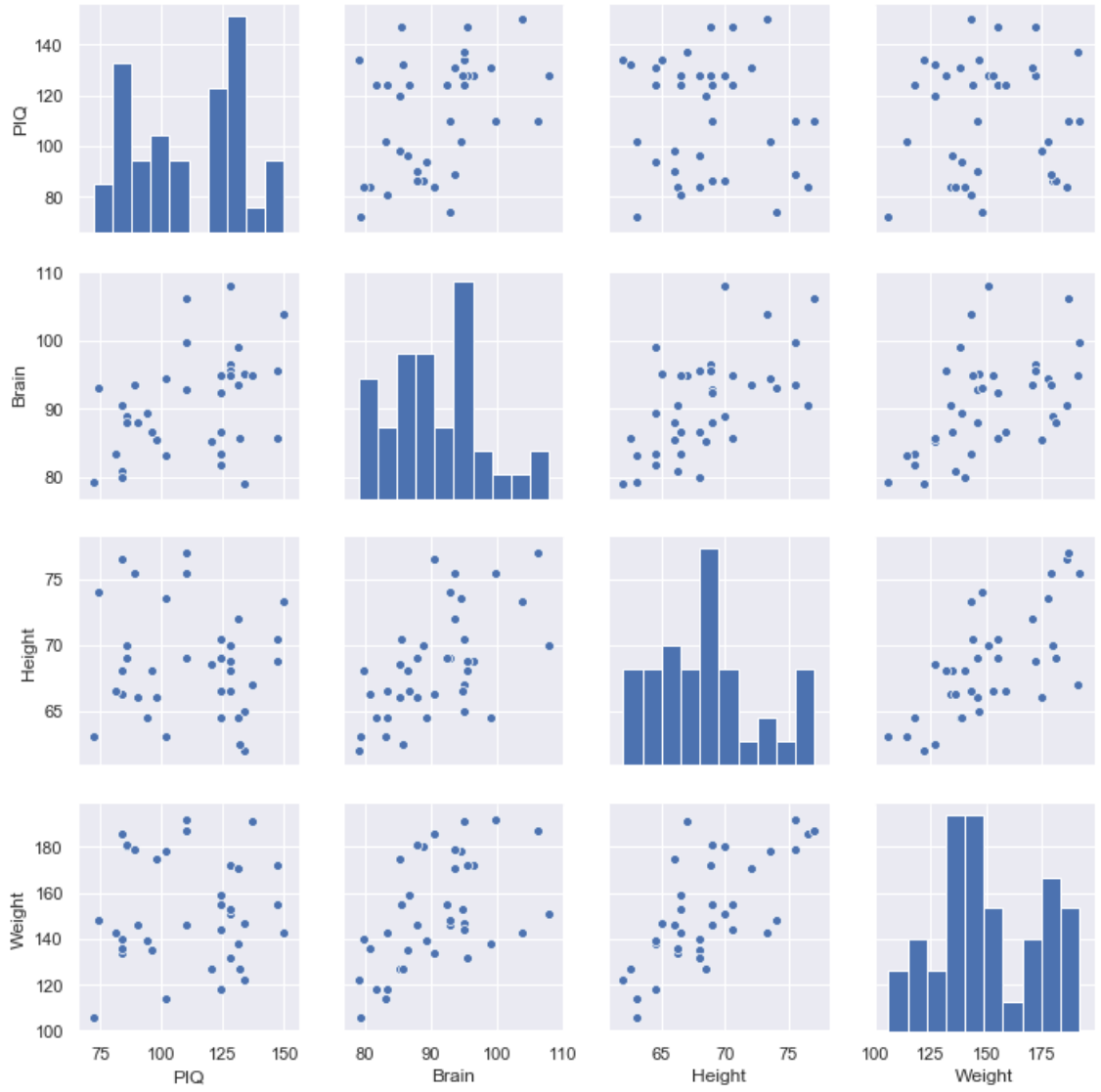
예 4) (iqsize.txt) 다음은 사람의 신체적 특성과 지적 능력 간에 관련성이 있는지를 알아보기 위해 38명의 학생들에 대해 조사한 자료이다. 총 4개의 변수가 포함되어 있고 각 변수에 대한 설명은 다음과 같다.

- PIQ: 종속변수. IQ score.
- Brain: 설명변수1. 두뇌 크기.
- Height: 설명변수2. 신장 (inch).
- Weight: 설명변수3. 체중 (pound).

[풀이] 분석의 첫 단계는 종속변수와 나머지 설명변수들 간의 상관 관계를 살펴보는 것이다. 이를 위해 산점도 행렬을 그려본 결과가 다음과 같다. 변수들 간의 상관계수는 pandas의 corr함수를 이용하여 구할 수 있다.

```
import pandas
from scipy import stats
import matplotlib.pyplot as plt
import seaborn

iqsize = pandas.read_table("iqsize.txt", sep = ' ')
```

```
seaborn.pairplot(data = iqsize)
plt.show()
```




```
iqsize.corr(method='pearson')
```

	PIQ	Brain	Height	Weight
PIQ	1.000000	0.377815	-0.093156	0.002512
Brain	0.377815	1.000000	0.588367	0.513487
Height	-0.093156	0.588367	1.000000	0.699614
Weight	0.002512	0.513487	0.699614	1.000000

```
t, p = stats.pearsonr(iqsize['PIQ'], iqsize['Brain'])
print("Pearson's correlation coefficient: %s, p-value: %s" % (t, p))
```

```
Pearson's correlation coefficient: 0.3778154625064401, p-value: 0.019354318790415
208
```

PIQ 과 Brain 변수 사이에는 선형 관계가 존재하는 것으로 보이지만 Height와 Weight 변수와는 눈에 띄는 강한 선형 관계가 보이지는 않는다. 상관계수를 확인해 보면 PIQ와 Brain의 상관계수는 0.3778로 나타났고 상관계수 검정의 유의확률은 0.01935로 나타난 것을 알 수 있다.

중회귀분석을 통해서 종속변수인 PIQ가 여러 개의 설명변수들과 동시적으로 어떠한 관련성이 있는 지를 살펴 볼 수 있다. 중회귀분석도 마찬가지로 statsmodels.formula.api의 ols 함수를 사용하고 여러 개의 독립변수를 '+'기호로 연결하여 차례대로 입력한다. 중회귀분석의 결과를 model에 저장하고 summary를 이용하여 결과를 출력하면 다음과 같다.

```
from statsmodels.formula.api import ols

model = ols("PIQ ~ Brain + Height + Weight", iqsize).fit()
print(model.summary())
```

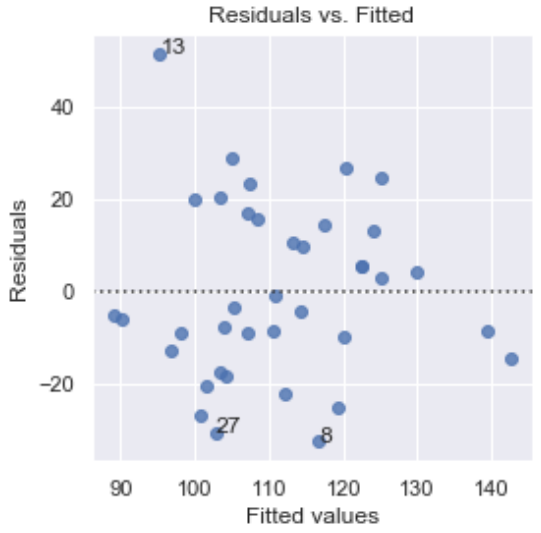
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    PIQ   R-squared:                       0.295
Model:                            OLS   Adj. R-squared:                  0.233
Method:                 Least Squares   F-statistic:                     4.741
Date:                Wed, 15 Aug 2018   Prob (F-statistic):            0.00722
Time:                        02:21:16   Log-Likelihood:                -165.25
No. Observations:                  38   AIC:                             338.5
Df Residuals:                      34   BIC:                             345.1
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    111.3536     62.971      1.768      0.086     -16.619     239.326
Brain          2.0604      0.563      3.657      0.001       0.915       3.205
Height        -2.7319      1.229     -2.222      0.033      -5.230      -0.233
Weight         0.0006      0.197      0.003      0.998      -0.400       0.401
==============================================================================
Omnibus:                        1.379   Durbin-Watson:                   1.827
Prob(Omnibus):                  0.502   Jarque-Bera (JB):                1.088
Skew:                           0.409   Prob(JB):                        0.580
Kurtosis:                       2.859   Cond. No.                     3.73e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly s
pecified.
[2] The condition number is large, 3.73e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

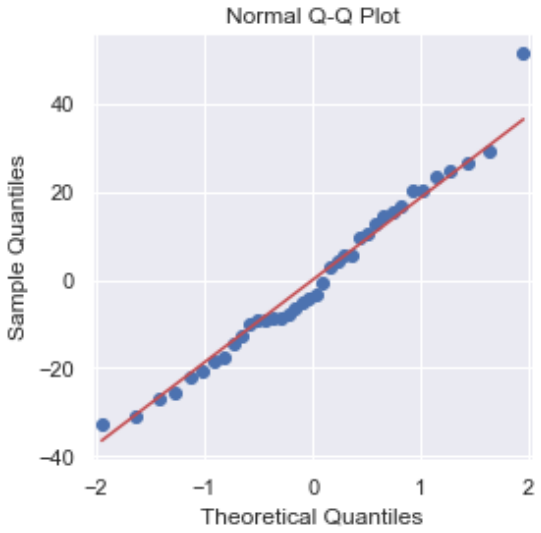
먼저 추정된 회귀식은 다음과 같다.

$$\hat{y} = 111.35 + 2.060Brain - 2.732Height + 0.0006Weight$$

모형의 유의성 검정 결과, 검정 통계량은 (F - statistic) 4.741이고 유의확률 (Prob (F-statistic))은 0.007로 나타났다. 따라서 유의수준 5%에서 모회귀함수는 유의하다고 할 수 있다. 결정계수 (R-squared)의 값은 0.295 (수정된 결정계수 (Adj. R-squared)는 0.233)로 나타났고 따라서 자료 전체의 산포 중 약 29%가 모회귀함수에 의해서 설명됨을 알 수 있다. 각 계수별 유의성의 검정 결과, Brain과 Height 계수의 유의확률이 유의수준 5%보다 작게 나타났다. 따라서 Brain과 Height는 PIQ를 설명함에 있어서 유의한 변수라고 볼 수 있다.

```
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.graphics.gofplots import ProbPlot


# 1. 잔차도(residual plot): Residuals vs. Fitted

f = model.fittedvalues              # 예측값 (fitted values)
r = model.resid                     # 잔차 (residuals)
ar = np.abs(r)                      # 잔차의 절대값 (absolute residuals)

ar_sort = ar.sort_values(ascending = False)
top_3_ar = ar_sort[0:3]             # top 3 absolute residuals

seaborn.residplot(f, r, data = possum)

plt.title("Residuals vs. Fitted")
plt.ylabel("Residuals")
plt.xlabel("Fitted values")

for i in top_3_ar.index:
    plt.annotate(i, xy = (f[i], r[i]))  # 잔차의 절대값이 큰 관측값 3개를 그래프에 표시
plt.show()

# 2. 정규 분위수 (Normal Q-Q plot)

QQ = ProbPlot(r)
plot = QQ.qqplot(line = 's', color='C0', lw=1)
plt.title("Normal Q-Q Plot")

plt.show()
```





적합된 모형의 잔차도를 확인해본 결과 잔차의 값이 매우 큰 관측치가 몇 개 존재하기는 하지만 특별한 패턴이 관측되지는 않았다. 그리고 정규 분위수 그래프에서도 정규분포를 벗어난다는 뚜렷한 증거는 발견되지 않았다. 따라서 주어진 자료에 대한 중선형회귀모형의 적용은 타당함을 알 수 있고, 적용된 모형을 통한 추론은 의미가 있다고 할 수 있다.

8.5 자료를 이용한 예제

예제 1. (handspan.txt) 다음은 167명의 학생들에 대해 성별(Sex)과 신장(Height) 그리고 손 한뼘의 길이(HandSpan)를 측정한 자료이다.

(1) 신장과 손 한뼘의 길이는 서로 상관관계가 존재하는가? 표본 상관계수를 구하고 두 변수 의 산점도를 그려보자. 두 변수 사이에 선형적 연관성이 존재하는가?

(2) 신장과 손 한뼘의 길이사이에 상관관계가 존재하는지 유의수준 5%에서 검정하여라.

(3) 신장(y)과 손 한뼘의 길이(x)에 대해 단순선형회귀모형을 적용해보자. 추정된 회귀식을 구 하고 유의수준 5%에서 회귀 직선의 유의성을 검정하시오.

(4) 단순 선형 회귀모형의 적용은 타당한가? 잔차도를 이용하여 답하시오.

예제 2. (carstopping.txt) 주어진 자료는 브레이크가 작동되는 순간의 자동차의 주행 속도 (Speed)에 따른 자동차 제동 거리(StopDist)를 조사한 자료이다.

(1) 자동차의 주행 속도에 따른 자동차의 제동거리 간에는 서로 상관관계가 존재하는가? 상관 분석을 통해 이를 확인해보자.

(2) 주어진 자료에 단순 선형회귀모형을 적용한 후 결과를 확인해 보자. 유의수준 5%에서 모 형은 유의한가?

(3) 적합된 회귀 모형의 잔차도를 확인해 보자. 단순선형회귀모형의 적용이 타당하다고 볼 수 있는가?

(4) 자동차의 주행속도와 자동차의 제동거리 사이의 산점도를 확인해보자. 두 변수 사이에는 곡률(curvature)관계가 존재하며, 또한 x 값이 증가함에 따라 y값의 산포가 증가하는 것을 확 인할 수 있다. 따라서 주어진 자료에 대해서는 단순 선형회귀 모형의 적용이 적절하지 않다. 이러한 문제를 해결하기 위한 방법 중 하나는 반응변수에 적절한 함수 변환(transformation) 을 취하는 것이다. 즉, 반응변수에 제곱근을 취한 새로운 변수(sqrt.dist)를 만든 후, 새로운 변수 sqrt.dist와 주행속도(Speed)의 산점도를 다시 한번 그려보자. 새로운 산점도는 어떠한 형태를 보이고 있는가?

(5) 새로운 변수 sqrt.dist와 Speed에 대해 단순선형회귀모형을 적합 시킨 후 결과를 확인해 보자. 새로운 모형의 결정계수 □값은 얼마인가? (1)번에서 구한 모형의 결정계수 값과 비교 해보시오.

(6) 새로운 모형의 잔차도를 확인해보자. 단순선형회귀모형의 적용이 타당하다고 볼 수 있는 가?

예제 3. (hospital.txt) 다음은 미국 내 113개의 병원들을 대상으로 입원 기간 동안 환자들이 받는 감염 위험과 관련된 사항들을 조사하였다. 다음은 주요 변수에 대한 설명이다.

- InfctRsk: 종속변수. 감염 위험 정도
- Stay: 설명변수1. 환자들의 평균 입원 기간
- Age: 설명변수2. 환자들의 평균 나이
- Xray: 설명변수3. 해당 병원의 X-ray 검진 횟수

(1) 종속변수와 각 설명변수들 간에는 유의한 상관관계가 존재하는가? 산점도와 상관분석을 통해 이를 확인해보시오.

(2) 주어진 자료에 다중선형회귀모형을 적용해보자. 유의수준 5%에서 모형은 유의하다고 할 수 있는가? 각 변수들은 유의한가?

(3) 다중선형회귀모형의 적용은 타당하다고 볼 수 있는가? 잔차도를 통해 확인해보자.