통계학실험(005)
과제 7 (8장)
2022-18608 문현호

1

예제1.(handspan.txt) 다음은 167명의 학생들에 대해 성별(Sex)과 신장(Height) 그리고 손 한 뼘의 길이(HandSpan)를 측정한 자료이다.

Python 코드

```
hand span = pd.read csv("data/handspan.txt", sep="\t")
print(hand span.describe())
```

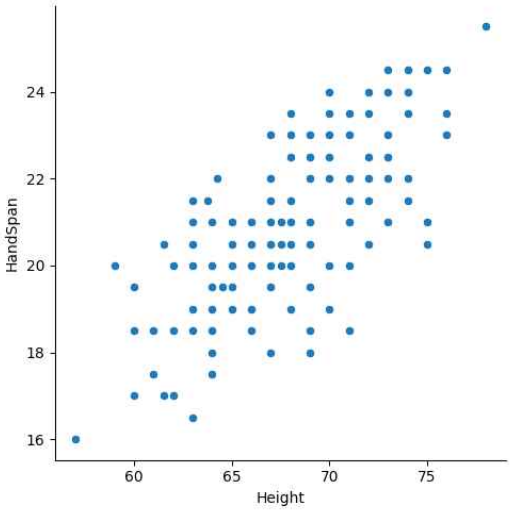
	Height	HandSpan
count	167.000000	167.000000
mean	68.071856	20.862275
std	4.064808	1.926875
min	57.000000	16.000000
25%	65.000000	19.500000
50%	68.000000	21.000000
75%	71.000000	22.000000
max	78.000000	25.500000

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

Python 코드

```
# scatter plot
sns.relplot(x="Height", y="HandSpan", data=hand_span)
plt.show()

# correlation analysis
print("correlation analysis")
print(hand_span.corr())
```

```
correlation analysis
             Height  HandSpan
Height     1.000000  0.739538
HandSpan   0.739538  1.000000
```



- 두 변수의 상관계수가 0.74로, 1에 가까운 편에 해당한다. 또한 그림을 보면, 두 변수 사이에 선형적 연관성, 선형 상관관계가 충분히 존재한다고 할 수 있다.

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

Python 코드

```
# correlation analysis 2
print(pearsonr(hand_span["Height"], hand_span["HandSpan"]))
```

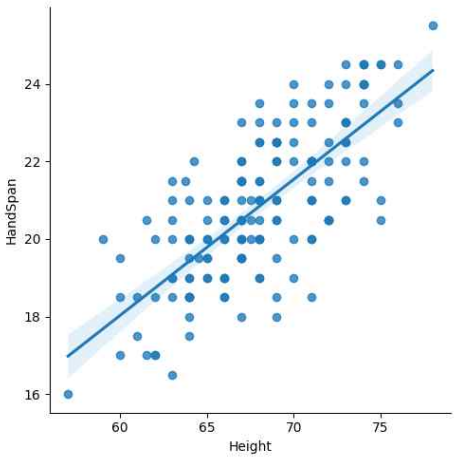
```
PearsonRResult(statistic=0.7395375015506631, pvalue=3.599304059035937e-30)
```

- 상관관계가 없다는 귀무가설의 p value가 0.05 미만이므로, 신장과 손 한 뼘의 길이 사이의 상관관계는 유의수준 5%에서 존재한다.

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

Python 코드

```
# scatter plot with regression line
sns.lmplot(x="Height", y="HandSpan", data=hand_span)
plt.show()

# regression model print
print("regression model print")
model = ols("Height ~ HandSpan", hand_span).fit()
print(model.summary())
```



regression model print

OLS Regression Results

Dep. Variable:	Height	R-squared:	0.547
Model:	OLS	Adj. R-squared:	0.544
Method:	Least Squares	F-statistic:	199.2
Date:	Thu, 08 Dec 2022	Prob (F-statistic):	3.60e-30
Time:	13:26:37	Log-Likelihood:	-404.55
No. Observations:	167	AIC:	813.1
Df Residuals:	165	BIC:	819.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	35.5250	2.316	15.339	0.000	30.952	40.098
HandSpan	1.5601	0.111	14.113	0.000	1.342	1.778

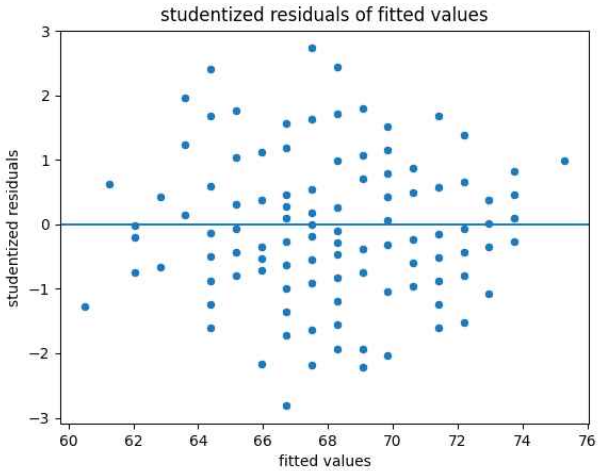
Omnibus:	0.395	Durbin-Watson:	1.988
Prob(Omnibus):	0.821	Jarque-Bera (JB):	0.163
Skew:	0.050	Prob(JB):	0.922
Kurtosis:	3.115	Cond. No.	229.

- 회귀식은 Height = 1.5601*HandSpan + 35.5250으로 나타났다.
- F 검정에 의하면 유의수준은 3.60*10^-30으로, 0.05 미만이므로 회귀 직선이 유의하지 않다는 귀무가설을 기각한다. 따라서 회귀직선은 유의수준 5%에서 유의하다.
- 이때 유의수준은 위에서 구한 상관관계와 같다. 수식이 같기 때문으로 보인다.

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

Python 코드

```
# regression model and residual plot
_, ax = plt.subplots()
sns.scatterplot(x=model.fittedvalues, y=model.resid_pearson)
ax.axhline(y=0)
plt.title("studentized residuals of fitted values")
plt.xlabel("fitted values")
plt.ylabel("studentized residuals")
plt.show()
```



- 잔차도를 보면, 값이 0 기준으로 거의 대칭으로 퍼져있고, 변수 값의 변화에 따라서도 산포가 거의 같으며, 특정한 모양을 가지고 있지 않고, 절댓값 2.5 이상으로의 점이 잘 나타나지 않는다. 따라서 대칭의 선형성, 산포가 같은 등분산성, 모양이 없는 독립성, 값이 크게 벗어나지 않는 정규성을 모두 가지고 있고, 선형 회귀 모형의 적용은 타당한다.

2

예제2.(carstopping.txt) 주어진 자료는 브레이크가 작동되는 순간의 자동차의 주행 속도(Speed)에 따른 자동차 제동 거리(StopDist)를 조사한 자료이다.

Python 코드

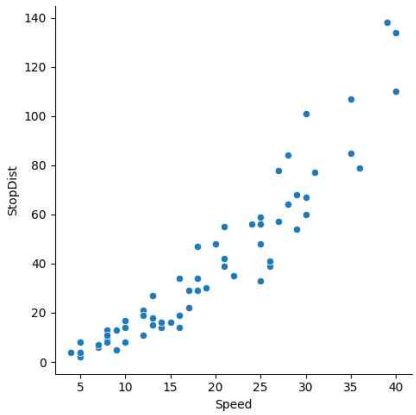
```
car_stopping = pd.read_csv("data/carstopping.txt", sep="\t")
print(car_stopping.describe())
```

	StopDist	Speed
count	63.000000	63.000000
mean	39.222222	18.968254
std	33.125589	9.879872
min	2.000000	4.000000
25%	13.500000	10.000000
50%	30.000000	18.000000
75%	56.500000	26.500000
max	138.000000	40.000000

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

Python 코드

```
# scatter plot
sns.relplot(x="Speed", y="StopDist", data=car_stopping)
plt.show()

# correlation analysis
print("correlation analysis")
print(car_stopping.corr())
print(pearsonr(car_stopping["Speed"], car_stopping["StopDist"]))
```



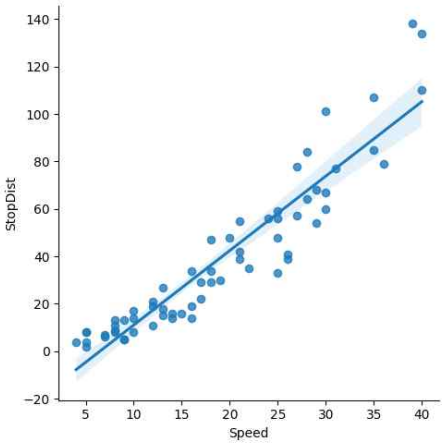

```
correlation analysis
          StopDist     Speed
StopDist  1.000000  0.935504
Speed     0.935504  1.000000
PearsonRResult(statistic=0.9355036785004626, pvalue=2.975121583730146e-29)
```

- 그림으로 보아도 서로 상관관계가 보이고, 계산된 상관계수도 상당히 높게 나타났다. p value도 2.98*10^-29로, 유의수준이 5%보다 매우 작다. 이에 따라 서로 상관관계가 있음을 알 수 있다.

(2) 주어진 자료에 단순 선형회귀모형을 적용한 후 결과를 확인해보자. 유의수준 5%에서 모형은 유의한가?

Python 코드

```
# scatter plot with regression line
sns.lmplot(x="Speed", y="StopDist", data=car_stopping)
plt.show()

# regression model print
print("regression model print")
model = ols("Speed ~ StopDist", car_stopping).fit()
print(model.summary())
```



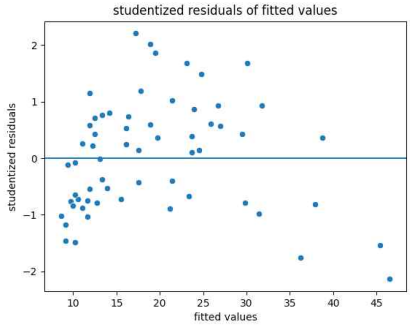

```
regression model print
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Speed   R-squared:                       0.875
Model:                            OLS   Adj. R-squared:                  0.873
Method:                 Least Squares   F-statistic:                     427.7
Date:                Thu, 08 Dec 2022   Prob (F-statistic):           2.98e-29
Time:                        17:09:16   Log-Likelihood:                -167.65
No. Observations:                  63   AIC:                             339.3
Df Residuals:                      61   BIC:                             343.6
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      8.0245      0.690     11.623      0.000       6.644       9.405
StopDist       0.2790      0.013     20.680      0.000       0.252       0.306
==============================================================================
Omnibus:                        1.217   Durbin-Watson:                   1.649
Prob(Omnibus):                  0.544   Jarque-Bera (JB):                1.088
Skew:                           0.135   Prob(JB):                        0.580
Kurtosis:                       2.416   Cond. No.                         79.7
==============================================================================
```

- 단순 선형회귀모형을 적용한 결과, Speed = 0.2790*StopDist + 8.0245의 관계를 가지고, 그 관계의 유의성은 위와 같은 2.98*10^29로 5%보다 매우 작게 나타났다. 따라서 모델은 유의수준 5%에서 유의하다.

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

Python 코드

```
# regression model and residual plot
_, ax = plt.subplots()
sns.scatterplot(x=model.fittedvalues, y=model.resid_pearson)
ax.axhline(y=0)
plt.title("studentized residuals of fitted values")
plt.xlabel("fitted values")
plt.ylabel("studentized residuals")
plt.show()
```

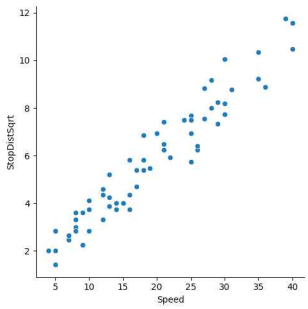


- 잔차도의 값이 위아래로 2.5 이상 퍼지지 않아 정규성은 보이지만, 패턴이 조금 보인다. 가운데에서만 양수가 많은 곡선 형태로, 대칭도 아니기 때문에 등분산성과 선형성, 독립성을 가지지 않아 적용이 별로 타당하지 않다.

(4) 자동차의 주행속도와 자동차의 제동거리 사이의 산점도를 확인해보자. 두 변수 사이에는 곡률(curvature)관계가 존재하며, 또한 x 값이 증가함에 따라 y값의 산포가 증가하는 것을 확인할 수 있다. 따라서 주어진 자료에 대해서는 단순선형회귀 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?

Python 코드

```
# Sqrt value of stop dist
car_stopping["StopDistSqrt"] = car_stopping.apply(lambda row:
math.sqrt(row.StopDist), axis=1)

# scatter plot
sns.relplot(x="Speed", y="StopDistSqrt", data=car_stopping)
plt.show()

# correlation analysis
print("correlation analysis")
print(car_stopping.corr())
print(pearsonr(car_stopping["Speed"], car_stopping["StopDistSqrt"]))
```



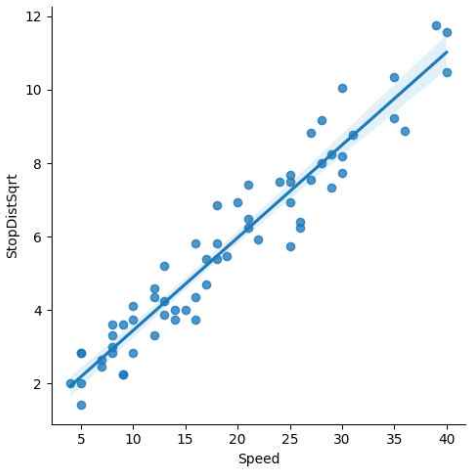

```
correlation analysis
              StopDist     Speed  StopDistSqrt
StopDist      1.000000  0.935504      0.976673
Speed         0.935504  1.000000      0.961474
StopDistSqrt  0.976673  0.961474      1.000000
PearsonRResult(statistic=0.9614739538060763, pvalue=6.505401870243608e-36)
```

- DataFrame에 새로운 column을 lambda 함수로 만들어 추가하였다.
- 산점도는 완전한 직선 형태를 가지고 있다. 또한 Speed와의 상관계수도 0.936에서 0.961로 증가한 것을 확인할 수 있다.

(5) 새로운 변수 sqrt.dist와 Speed에 대해 단순선형회귀모형을 적합 시킨 후 결과를 확인해 보자. 새로운 모형의 결정계수 R^2 값은 얼마인가? (1)번에서 구한 모형의 결정계수 값과 비교해보시오.

Python 코드

```
# scatter_plot with regression line
sns.lmplot(x="Speed", y="StopDistSqrt", data=car_stopping)
plt.show()

# regression model print
print("regression model print")
model = ols("Speed ~ StopDistSqrt", car_stopping).fit()
print(model.summary())
```



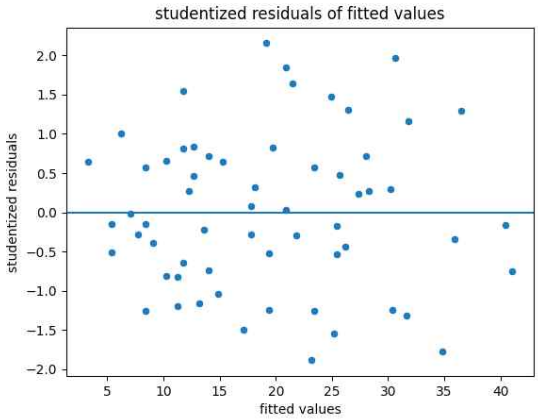

```
regression model print
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Speed   R-squared:                       0.924
Model:                            OLS   Adj. R-squared:                  0.923
Method:                 Least Squares   F-statistic:                     746.2
Date:                Thu, 08 Dec 2022   Prob (F-statistic):           6.51e-36
Time:                        17:09:20   Log-Likelihood:                -151.83
No. Observations:                  63   AIC:                             307.7
Df Residuals:                      61   BIC:                             312.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -1.9276      0.839     -2.297      0.025      -3.606      -0.250
StopDistSqrt   3.6601      0.134     27.317      0.000       3.392       3.928
==============================================================================
Omnibus:                        2.536   Durbin-Watson:                   2.208
Prob(Omnibus):                  0.281   Jarque-Bera (JB):                1.637
Skew:                           0.144   Prob(JB):                        0.441
Kurtosis:                       2.265   Cond. No.                         15.6
==============================================================================
```

- StopDistSqrt와 Speed 사이 결정계수 R^2 값은 0.924였으며, 앞에서 구한 StopDist의 0.875보다 증가한 것을 확인할 수 있다. 더욱 선형인 관계를 가짐을 확인할 수 있다.

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

Python 코드

```
# regression model and residual plot
_, ax = plt.subplots()
sns.scatterplot(x=model.fittedvalues, y=model.resid_pearson)
ax.axhline(y=0)
plt.title("studentized residuals of fitted values")
plt.xlabel("fitted values")
plt.ylabel("studentized residuals")
plt.show()
```



- 값이 0에 대칭을 보이는 선형성을 보이고, 값에 따라 잔차 산포가 거의 같아 등분산성을 보이며, 따로 패턴은 없는 것처럼 나타나기 때문에 독립성을 가진다. 또한 값이 2.5 이상으로 퍼지지 않아 정규성을 가진다. 잔차도로 확인할 수 있는 정보에 따르면, 단순선형회귀모형의 적용은 타당하다고 볼 수 있다.

3

예제3. (hospital.txt) 다음은 미국 내 113개의 병원들을 대상으로 입원 기간 동안 환자들이 받는 감염 위험과 관련된 사항들을 조사하였다. 다음은 주요 변수에 대한 설명이다.

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

Python 코드

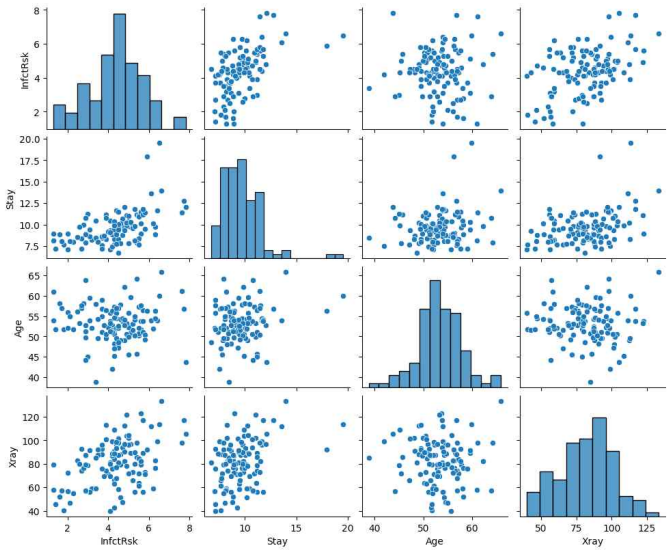
```
hospital = pd.read_csv("data/hospital.txt", sep="\t")
print(hospital.describe())

hospital1 = hospital[["InfctRsk", "Stay", "Age", "Xray"]]
print(hospital1.describe())
```

```
              ID        Stay         Age  ...      Census      Nurses  Facilities
count  113.00000  113.000000  113.000000  ...  113.000000  113.000000  113.000000
mean    57.00000    9.648319   53.231858  ...  191.371681  173.247788   43.159292
std     32.76431    1.911456    4.461607  ...  153.759564  139.265390   15.200861
min      1.00000    6.700000   38.800000  ...   20.000000   14.000000    5.700000
25%     29.00000    8.340000   50.900000  ...   68.000000   66.000000   31.400000
50%     57.00000    9.420000   53.200000  ...  143.000000  132.000000   42.900000
75%     85.00000   10.470000   56.200000  ...  252.000000  218.000000   54.300000
max    113.00000   19.560000   65.900000  ...  791.000000  656.000000   80.000000

[8 rows x 12 columns]
         InfctRsk        Stay         Age        Xray
count  113.000000  113.000000  113.000000  113.000000
mean     4.354867    9.648319   53.231858   81.628319
std      1.340908    1.911456    4.461607   19.363826
min      1.300000    6.700000   38.800000   39.600000
25%      3.700000    8.340000   50.900000   69.500000
50%      4.400000    9.420000   53.200000   82.300000
75%      5.200000   10.470000   56.200000   94.100000
max      7.800000   19.560000   65.900000  133.500000
```

(1) 종속변수와 각각의 설명변수들 사이에는 유의한 상관관계가 존재하는가? 산점도와 상관분석을 통해 이를 확인해보시오.

Python 코드

```
# pair scatter plot
sns.pairplot(data=hospital1)
plt.show()

# correlation analysis
print("correlation analysis")
print(hospital1.corr())
print("InfctRsk, Stay", pearsonr(hospital1["InfctRsk"],
hospital1["Stay"]))
print("InfctRsk, Age", pearsonr(hospital1["InfctRsk"], hospital1["Age"]))
print("InfctRsk, Xray", pearsonr(hospital1["InfctRsk"],
hospital1["Xray"]))
print("Stay, Age", pearsonr(hospital1["Stay"], hospital1["Age"]))
print("Stay, Xray", pearsonr(hospital1["Stay"], hospital1["Xray"]))
print("Age, Xray", pearsonr(hospital1["Age"], hospital1["Xray"]))
```




```
correlation analysis
          InfctRsk      Stay       Age      Xray
InfctRsk  1.000000  0.533444  0.001093  0.453392
Stay      0.533444  1.000000  0.188914  0.382482
Age       0.001093  0.188914  1.000000 -0.018855
Xray      0.453392  0.382482 -0.018855  1.000000
InfctRsk, Stay PearsonRResult(statistic=0.5334438309449094, pvalue=1.1769611863413669e-09)
InfctRsk, Age PearsonRResult(statistic=0.001093166148662385, pvalue=0.9908314602981837)
InfctRsk, Xray PearsonRResult(statistic=0.4533915569793219, pvalue=4.584845452365265e-07)
Stay, Age PearsonRResult(statistic=0.18891397199126475, pvalue=0.04507677688347593)
Stay, Xray PearsonRResult(statistic=0.3824819298637571, pvalue=2.9055594558771173e-05)
Age, Xray PearsonRResult(statistic=-0.018854896501825837, pvalue=0.8428734128226135)
```

- 산점도를 보고, 상관관계에 대한 F 검정을 시행했을 때의 값은 그림과 같다. 이때 InfectRisk와 Age, Age와 Xray만이 서로의 상관관계에 대한 0.05 이상의 p value를 보였다. 이에 따라 서로간의 상관관계는 5% 유의수준에서 없다고 할 수 있다. 이들 둘 사이의 관계는 그림에서 보아도 그저 둥근 모양을 가진다.
- 반대로 InfectRisk와 Stay, InfectRisk와 Xray, Stay와 Age, Stay와 Xray는 서로간의 상관관계에 대한 p value가 0.05 미만이었다. 이에 따라 이들 서로간의 상관관계는 5% 유의수준에서 있다고 할 수 있다. 산점도에서도 확인할 수 있다.

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

Python 코드

```
# regression model print
print("regression model print")
model = ols("InfctRsk ~ Stay + Age + Xray", hospital1).fit()
print(model.summary())
```

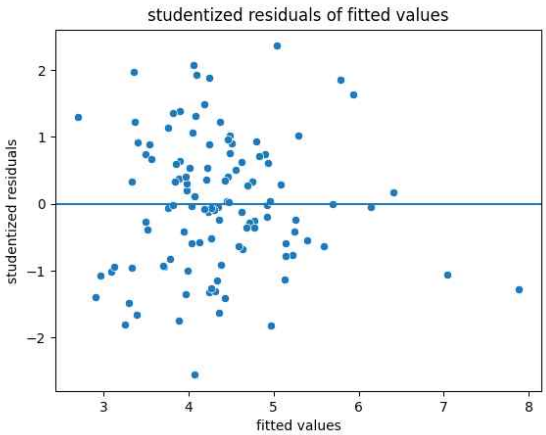
```
                            OLS Regression Results
==============================================================================
Dep. Variable:              InfctRsk   R-squared:                       0.363
Model:                           OLS   Adj. R-squared:                  0.345
Method:                Least Squares   F-statistic:                     20.70
Date:               Thu, 08 Dec 2022   Prob (F-statistic):           1.09e-10
Time:                       17:09:27   Log-Likelihood:                -167.51
No. Observations:                113   AIC:                             343.0
Df Residuals:                    109   BIC:                             353.9
Df Model:                          3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.0012      1.315      0.761      0.448      -1.605       3.607
Stay           0.3082      0.059      5.189      0.000       0.190       0.426
Age           -0.0230      0.024     -0.978      0.330      -0.070       0.024
Xray           0.0197      0.006      3.414      0.001       0.008       0.031
==============================================================================
Omnibus:                        0.750   Durbin-Watson:                   1.881
Prob(Omnibus):                  0.687   Jarque-Bera (JB):                0.823
Skew:                           0.063   Prob(JB):                        0.663
Kurtosis:                       2.601   Cond. No.                     1.28e+03
==============================================================================
```

- 전체 선형 모델의 F 검정 결과 p value는 1.09*10^-10으로, 0.05보다는 충분히 작았다. 이에 따라 이들 사이에 관계가 없다는 귀무가설을 기각하며, 유의수준 5%에서 이 모형은 유의미하다고 할 수 있다.

(3) 다중선형회귀모형의 적용은 타당하다고 볼 수 있는가?

Python 코드

```
# regression model and residual plot
_, ax = plt.subplots()
sns.scatterplot(x=model.fittedvalues, y=model.resid_pearson)
ax.axhline(y=0)
plt.title("studentized residuals of fitted values")
plt.xlabel("fitted values")
plt.ylabel("studentized residuals")
plt.show()
```



- 값이 2.5 이상으로 퍼지지 않아 정규성을 가지고, 중심에 대칭적으로 몰려있어 선형성을 가지며, 전체적으로 잔차의 분포가 유사하여 등분산성을 가지고(조금 의심스러운 부분이 있는 것은 사실이다), 특정 방향으로의 변화나 패턴을 보이지 않으므로 독립성을 가진다. 이들 값으로 확인한 결과로는 다중 선형회귀 모형은 어느 정도 타당하다고 할 수 있다.

add 1

(1) 주어진 데이터를 불러오고, 우리는 주어진 변수들 중에서 “Xray”, “Beds”, “Census”, “Nurses”, “Facilities” 이들 변수만을 사용할 것이다. 이 변수들만 포함된 새로운 데이터프레임 df를 만들어라.

Python 코드

```
# make a new DataFrame
hospital2 = hospital[["Xray", "Beds", "Census", "Nurses", "Facilities"]]
print(hospital2.describe())
```

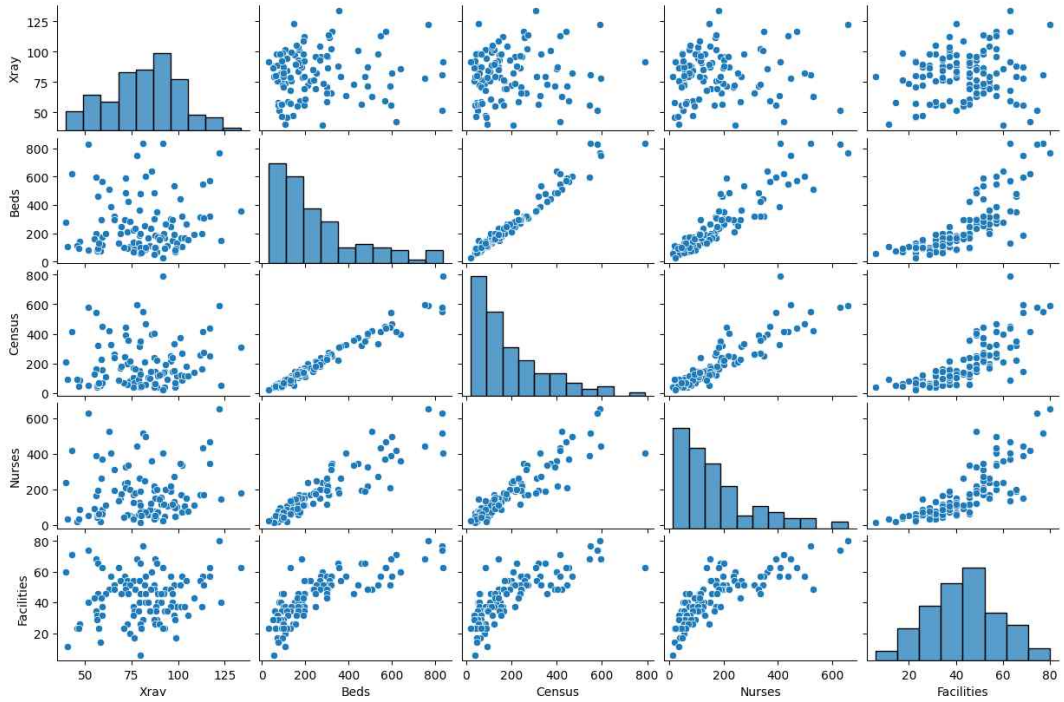
	Xray	Beds	Census	Nurses	Facilities
count	113.000000	113.000000	113.000000	113.000000	113.000000
mean	81.628319	252.168142	191.371681	173.247788	43.159292
std	19.363826	192.842687	153.759564	139.265390	15.200861
min	39.600000	29.000000	20.000000	14.000000	5.700000
25%	69.500000	106.000000	68.000000	66.000000	31.400000
50%	82.300000	186.000000	143.000000	132.000000	42.900000
75%	94.100000	312.000000	252.000000	218.000000	54.300000
max	133.500000	835.000000	791.000000	656.000000	80.000000

- 이와 같은 형태로 만들 수 있다. 그대로 새로운 데이터프레임으로서 이용한다.

(2) df 데이터 프레임의 각 변수에 대한 산점도를 그려라.

Python 코드

```
# pair_scatter_plot
sns.pairplot(data=hospital2)
plt.show()
```



- 산점도의 형태는 위와 같다. 이를 보면 Xray와의 비교를 제외한 나머지 모든 비교 사이에서는 선형으로 보이는 관계가 나타난다. 반대로 Xray와의 비교는 어떤 변수에 대해서도 상당히 넓게 펴진 형태를 나타낸다.

add 2

(1) Census 와 Beds의 상관계수 행렬을 구하고 상관분석을 실시하여라.

Python 코드

```
# correlation analysis
print("correlation analysis")
print(hospital2[["Census", "Beds"]].corr())
print("Census, Beds", pearsonr(hospital2["Census"], hospital2["Beds"]))
```

```
          Census      Beds
Census  1.000000  0.980998
Beds    0.980998  1.000000
Census, Beds PearsonRResult(statistic=0.9809977426417227, pvalue=6.874348044091662e-81)
```

- 상관분석 결과, 상관계수는 서로 0.98이라는 큰 수가 나왔고, p value도 6.87*10^-81이라는, 0.05보다 훨씬 낮은 숫자가 나왔다. 따라서 두 변수 사이의 선형 관계가 없다는 귀무가설을 기각하고, 두 변수 사이에는 유의수준 0.05에서 선형 관계가 있음을 확인할 수 있다.

(2) Beds를 종속변수로, Census를 설명변수로 하여 단순선형회귀모형을 적합하여라. 이 모델이 타당하다고 볼 수 있는가? 이 모델의 결정계수(R^2) 와 F-검정통계량의 값을 구하여라.

Python 코드

```
# regression model print
print("regression model print")
model = ols("Beds ~ Census", hospital2).fit()
print(model.summary())
```

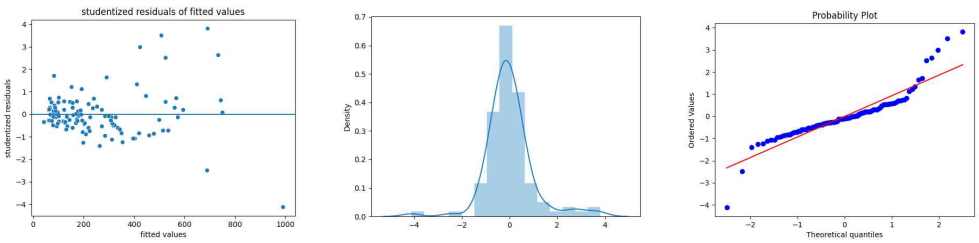
```
regression model print
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Beds   R-squared:                       0.962
Model:                            OLS   Adj. R-squared:                  0.962
Method:                 Least Squares   F-statistic:                     2838.
Date:                Thu, 08 Dec 2022   Prob (F-statistic):           6.87e-81
Time:                        17:09:38   Log-Likelihood:                -569.13
No. Observations:                 113   AIC:                             1142.
Df Residuals:                     111   BIC:                             1148.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     16.7138      5.660      2.953      0.004       5.498      27.930
Census         1.2304      0.023     53.270      0.000       1.185       1.276
==============================================================================
Omnibus:                       29.648   Durbin-Watson:                   1.866
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              142.128
Skew:                           0.687   Prob(JB):                     1.37e-31
Kurtosis:                       8.319   Cond. No.                         392.
==============================================================================
```

- 결정계수 R^2 값은 0.962, F-검정 통계량 값은 2838, p value는 6.87*10^-81로, 위에서와 같이 0.05보다 상당히 낮은 관계가 나왔다. 이에 따라 이 모델과 같은 관계가 없다는 귀무가설을 기각하고, 이 모델은 타당하다고 할 수 있다.

(3) 위의 결과로 얻어진 모델에 대해서 잔차분석을 시행해보여라. 선형성, 등분산성, 독립성, 정규성을 확인하여라.

Python 코드

```
# regression model and residual plot
_, ax = plt.subplots()
sns.scatterplot(x=model.fittedvalues, y=model.resid_pearson)
ax.axhline(y=0)
plt.title("studentized residuals of fitted values")
plt.xlabel("fitted values")
plt.ylabel("studentized residuals")
plt.show()

sns.distplot(model.resid_pearson, bins=15)
plt.show()

probplot(model.resid_pearson, plot=plt)
plt.show()
```



- 잔차도를 보았을 때 0에 대해 대칭적이기 때문에 선형성을 가지고, 3, 4 이상으로 벗어난 값이 조금 있지만 무시 가능하기 때문에 정규성을 가지고, 특별한 패턴이 없기 때문에 독립성을, 값에 따라서 잔차 분포의 변화는 작지만 거의 없기 때문에 등분산성을 가진다. 또한 히스토그램이 0을 중심으로 어느 정도 대칭을 보이고(선형성), 정규분포를 닮았으며(정규성), 정규확률도표 역시 주어진 직선과 상당히 일치한 값을 보여, 각 값들이 정규분포를 따르고 있음을 알 수 있다(정규성). 마지막으로 위에서 구한 값에서 나타나는 Durbin-Watson 값이 1.866으로 2에 가까운 편이다(독립성).
- 잔차분석 결과가 선형성, 등분산성, 정규성, 독립성을 모두 가지므로 이 선형 모델은 타당하다고 할 수 있다.

(4) 더빈왓슨 테스트를 직접 해보고 그것과 model.summary()로 구한값을 비교해라

Python 코드

```
print("Durbin Watson by func : ", durbin_watson(model.resid_pearson))

e = model.resid_pearson

a, b = 0, 0
for i in range(0, len(e)):
a += e[i]*e[i]
b += (e[i]-e[i-1])*(e[i]-e[i-1])
print("Durbin Watson by hand : ", b/a)

a, b = 0, 0
for i in range(1, len(e)):
a += e[i]*e[i]
b += (e[i]-e[i-1])*(e[i]-e[i-1])
print("Durbin Watson by hand : ", b/a)
```

```
Durbin Watson by func :  1.8659352100283118
Durbin Watson by hand :  1.8684671590711128
Durbin Watson by hand :  1.8666235054572025
```

- durbin-watson을 이용한 경우는 정확히 같은 값을 가진다. 반면에 직접 형성한 함수의 값은 조금 다르게 나타났지만, 큰 차이를 보이지는 않았다. 이는 계산의 순서 차이, 서로간의 계산에 의한 차이로 예상할 수 있다.

(5) Census의 회귀계수의 추정값에 대한 t-value를 제곱한 것과 모델의 F-검정통계량의 값을 비교해라.

```
regression model print
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Beds   R-squared:                       0.962
Model:                            OLS   Adj. R-squared:                  0.962
Method:                 Least Squares   F-statistic:                     2838.
Date:                Thu, 08 Dec 2022   Prob (F-statistic):          6.87e-81
Time:                        17:09:38   Log-Likelihood:                -569.13
No. Observations:                 113   AIC:                             1142.
Df Residuals:                     111   BIC:                             1148.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     16.7138      5.660      2.953      0.004       5.498      27.930
Census         1.2304      0.023     53.270      0.000       1.185       1.276
==============================================================================
Omnibus:                       29.648   Durbin-Watson:                   1.866
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              142.128
Skew:                           0.687   Prob(JB):                     1.37e-31
Kurtosis:                       8.319   Cond. No.                         392.
==============================================================================
```

(위의 사진과 같다. 코드는 생략하였다.)

- Census의 t value는 53.270, 그 제곱값은 2837.6929이다. 이 값은 모델의 F-검정통계량인 2838.과 유효숫자 내에서 동일하다. 수식이 동일하기 때문으로, 그리고 모델의 F-검정통계량을 계산하는 방법이 바로 변수의 t value의 제곱과 관계가 있다고 추측할 수 있다.
- 다른 모델에서는 이 수식이 직접 성립하지는 않았다. 다른 여러 t value의 제곱의 단순 합보다는 작고, 이에 따라 어떤 형태의 평균일 수 있으나 산술평균은 아닌 것으로 보인다.

add 3

(1) Facilities를 종속변수로 하고 “Xray”, “Beds”, “Census”, “Nurses”을 설명변수로 해서 다중선형회귀모형을 적용해라. 이 모형이 유의수준 5%에서 타당한지 설명하여라. 4개의 설명변수 중에서 유의미하지 않은 변수가 있는지 확인해라.

Python 코드

```
# regression model print
print("regression model print")
model = ols("Facilities ~ Xray + Beds + Census + Nurses", hospital2).fit()
print(model.summary())
```

```
regression model print
                            OLS Regression Results
==============================================================================
Dep. Variable:             Facilities   R-squared:                       0.656
Model:                            OLS   Adj. R-squared:                  0.643
Method:                 Least Squares   F-statistic:                     51.41
Date:                Thu, 08 Dec 2022   Prob (F-statistic):           3.61e-24
Time:                        17:09:43   Log-Likelihood:                -407.11
No. Observations:                 113   AIC:                             824.2
Df Residuals:                     108   BIC:                             837.9
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     22.8394      3.875      5.894      0.000      15.159      30.520
Xray           0.0521      0.045      1.166      0.246      -0.036       0.141
Beds           0.0499      0.024      2.058      0.042       0.002       0.098
Census        -0.0155      0.029     -0.533      0.595      -0.073       0.042
Nurses         0.0373      0.015      2.408      0.018       0.007       0.068
==============================================================================
Omnibus:                        3.870   Durbin-Watson:                   1.750
Prob(Omnibus):                  0.144   Jarque-Bera (JB):                3.339
Skew:                           0.312   Prob(JB):                        0.188
Kurtosis:                       3.566   Cond. No.                     2.08e+03
==============================================================================
```

- 모형 적용 결과, p value는 3.61&10^-24 <0.05로, 이 모델이 타당하지 않다는 가설이 기각되기 때문에 유의수준 5%에서 이 모델은 타당하다.
- 4개의 변수 중 p value가 0.05 이상인 것은 Xray, 그리고 Census가 있다. 따라서 이 두 변수는 유의미하지 않다는 귀무가설이 채택된다. Beds와 Nurses에 대해서는 귀무가설이 기각되어, 이 두 변수는 유의미하다는 대립가설이 채택된다. [Xray, Census].

(2) 위 결과에서 유의미한 변수만을 가지고 다중선형회귀 모형을 적용해라. 이 모형이 유의수준 5%에서 타당한지 설명하여라. 그리고 각 변수가 유의미한지 확인하여라.

Python 코드

```
# regression model print without Xray, Census (inputs whose p is over 0.05)
print("regression model print")
model = ols("Facilities ~ Beds + Nurses", hospital2).fit()
print(model.summary())
```

```
regression model print
                            OLS Regression Results
==============================================================================
Dep. Variable:             Facilities   R-squared:                       0.651
Model:                            OLS   Adj. R-squared:                  0.644
Method:                 Least Squares   F-statistic:                     102.5
Date:                Thu, 08 Dec 2022   Prob (F-statistic):           7.49e-26
Time:                        17:09:43   Log-Likelihood:                -407.92
No. Observations:                 113   AIC:                             821.8
Df Residuals:                     110   BIC:                             830.0
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     27.1194      1.411     19.216      0.000      24.323      29.916
Beds           0.0376      0.011      3.406      0.001       0.016       0.059
Nurses         0.0378      0.015      2.475      0.015       0.008       0.068
==============================================================================
Omnibus:                        3.089   Durbin-Watson:                   1.778
Prob(Omnibus):                  0.213   Jarque-Bera (JB):                2.656
Skew:                           0.222   Prob(JB):                        0.265
Kurtosis:                       3.605   Cond. No.                         636.
==============================================================================
```

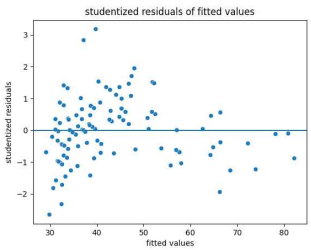
- 이 모델의 F검정 p value는 7.49*10^-26 <0.05로, 유의수준 5%에서 귀무가설이 기각되어 유의수준 5%에서 타당하다고 할 수 있다.
- 각 변수 모두 p value가 0.05 미만으로, 유의미하지 않다는 귀무가설이 유의수준 5%에서 기각되어 두 변수 모두 유의미하다고 할 수 있다.
- 결정계수 R^2 값은 약간 감소하였지만, F 검정통계량 값은 2배 가까이 증가했다.

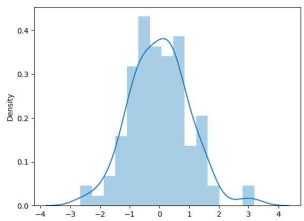
(3) 2의 결과로 얻어진 모델에 대해서 잔차분석을 시행해보여라. 선형성, 등분산성, 독립성, 정규성을 확인하여라.

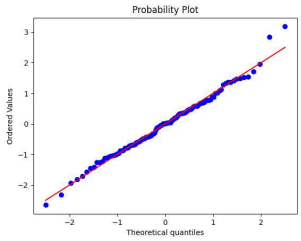
Python 코드

```
# regression model and residual plot
_, ax = plt.subplots()
sns.scatterplot(x=model.fittedvalues, y=model.resid_pearson)
ax.axhline(y=0)
plt.title("studentized residuals of fitted values")
plt.xlabel("fitted values")
plt.ylabel("studentized residuals")
plt.show()

sns.distplot(model.resid_pearson, bins=15)
plt.show()

probplot(model.resid_pearson, plot=plt)
plt.show()

print("Durbin Watson : ", durbin_watson(model.resid_pearson))
```

```
Durbin Watson :  1.7777893663320536
```







- 잔차도를 보면 0에 대한 대칭이 어느 정도 있어 선형성을 가지고, 3 이상으로 벗어난 값이 하나 있지만 무시할 수 있으므로 정규성을 가지며, 전체적인 형태에 특별한 패턴이 없어 독립성을 가진다. 각 값에 따른 잔차의 분포가 완전히 동일하지는 않아 등분산성을 논하기는 어렵지만, 어느 정도 비슷하기 때문에 등분산성이 있다고 할 수 있다. 또한 히스토그램의 형태가 0을 중심으로 대칭을 보여 선형적이고, 정규분포와 그 모양이 유사하며, 정규확률도표 역시 주어진 직선과 거의 일치하므로 정규분포와 비슷하여 정규성을 보인다. 추가로 Durbin-Watson 값은 1.778로, 2에 가까워 독립성을 보인다.
- 잔차 분석 결과, 이 모델은 선형성, 등분산성, 정규성, 그리고 독립성을 모두 어느 정도 이상으로 가지기 때문에, 타당하다고 할 수 있다.