단순선형회귀분석

Park Beomjin[1]
[1] *University of Seoul*

1 예제1: 발길이와 앞팔길이 데이터

- 발길이와 앞팔길이 데이터는 S 대학교 학생들 중 무작위로 남녀 각각 16명을 추출하여 조사한 자료로 발길이와 앞팔길이를 포함하고 있다. 단순선형회귀모형을 통해 두 변수의 관계성을 분석해보자.

- 데이터 읽기

```
libname reg "C:\Users\User\Desktop\19학년도 회귀분석 및 실습\";

PROC IMPORT datafile = "C:\Users\User\Desktop\19학년도 회귀분석 및 실습\
Data\Week2 Data\aflength.csv" dbms = csv replace
out = reg.aflength;
getnames = yes;
LABEL foot = "발길이" forearm = "팔안쪽길이" gender = "성별";
RUN;
```

- 데이터의 요약정보 보기

```
PROC CONTENTS data = reg.aflength;
RUN;
```

데이터셋 이름	WORK.AFLENGTH	관측값	32
멤버 유형	DATA	변수	3
엔진	V9	인덱스	0
생성일	2019.03.11 19:29:46	관측값 길이	24
마지막 수정일	2019.03.11 19:29:46	삭제된 관측값	0
보호		압축여부	아니요
데이터셋 유형		정렬	아니요
레이블			
데이터 표현	WINDOWS_64		
인코딩	euc-kr Korean (EUC)		

변수와 속성 리스트(오름차순)

#	변수	유형	길이	출력형식	입력형식	레이블
1	foot	숫자	8	BEST12.	BEST32.	발길이
2	forearm	숫자	8	BEST12.	BEST32.	팔안쪽길이
3	gender	문자	3	$3.	$3.	성별

- 팔안쪽길이 및 발길이의 요약 통계량 확인하기

2

```
PROC means data = reg.aflength;
var foot forearm;
RUN;
```

변수	레이블	N	평균	표준편차	최솟값	최댓값
foot	발길이	32	241.1875000	17.7535776	200.0000000	277.0000000
forearm	팔안쪽길이	32	240.9062500	16.6275313	201.0000000	272.0000000

- 팔안쪽길이 및 발길이의 히스토그램 그리기

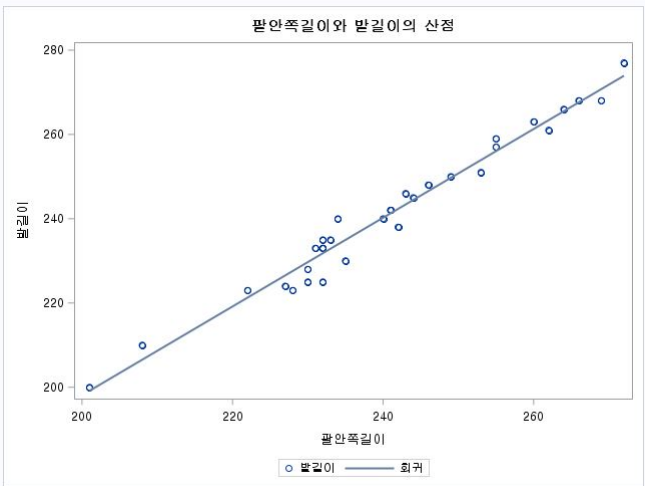
```
PROC SGPLOT data = reg.aflength;
HISTOGRAM foot;
RUN;
```

- 팔안쪽길이 및 발길이의 Boxplot 그리기

```
PROC SGPLOT data = reg.aflength;
HBOX foot;
RUN;
```

- 팔안쪽길이와 발길이의 산점도 확인하기

```
TITLE "팔안쪽길이와 발길이의 산점도";
PROC SGPLOT DATA = reg.aflength;
SCATTER X = forearm Y = foot; /* X, Y 변수의 산점도 작성 */
REG X = forearm Y = foot;  /* 산점도에 추정회귀직선을 추가로 표시*/
RUN;
```



- 팔안쪽길이와 발길이에 대한 상관분석하기

```
PROC CORR data = reg.aflength FISHER(rho0 = 0); /* 상관관계분석 */
VAR foot;          /* FISHER 옵션: 상관관계에 대한 검정과 신뢰구간을 계산한다*/
WITH forearm;
RUN;
```

단순 통계량

변수	N	평균	표준편차	합	최솟값	최댓값	레이블
forearm	32	240.90625	16.62753	7709	201.00000	272.00000	팔안쪽길이
foot	32	241.18750	17.75358	7718	200.00000	277.00000	발길이

피어슨 상관 계수, N = 32
H0: Rho=0 가정하에서 Prob > |r|

	foot
forearm 팔안쪽길이	0.98595 <.0001

피어슨 상관통계량 (피셔의 z 변환)

변수	조합 변수	N	표본 상관계수	피셔의 z	편의 조정	상관계수 추정값	95% 신뢰한계		H0:Rho=Rho0	
									Rho0	p 값
foot	forearm	32	0.98595	2.47559	0.01590	0.98550	0.970202	0.992971	0	<.0001

- 팔안쪽길이와 발길이에 대한 회귀분석하기
  - 반응변수 Y : 발길이
  - 설명변수 X : 팔안쪽길이
  - 모형 : $Y = \beta_0 + \beta_1 X + \epsilon,\ \epsilon \sim N(0, \sigma^2)$

- 최소제곱법을 직접 계산해 회귀계수 구하기

```
proc means data = reg.aflength noprint;
        var foot forearm;
        output out = mean_set mean = /autoname;
run;
/*x, y 평균 구해서 저장하기
데이터 가져와서 foot, forearm 두 변수에 대해 평균을 계산해서 mean_set에 저장
이때, 이름은 autoname으로 자동 지정되게 */

data aflength;
        if _n_ = 1 then set mean_set (drop = _freq_ _type_);
```

```
        set reg.aflength;
        diff_foot = foot - foot_mean;
        diff_forearm = forearm - forearm_mean;
        squ_diff_forearm = diff_forearm**2;
        pro_var = diff_foot*diff_forearm;
run;
/* mean_set과 example.aflength를 결합하여 한꺼번에 계산하려고 사용한 코드
mean_set에서 _freq_와 _type_을 버리고 가져온 후 밑에 식들을 계산*/

proc means data = aflength noprint;
        var pro_var squ_diff_forearm;
        output out = sum_set sum = /autoname;
run;
/*mean계산과 마찬가지로 sum 계산하기 위함*/

data beta_set;
        if _n_ = 1 then set sum_set;
        set mean_set;
        beta_1 = pro_var_sum / squ_diff_forearm_sum;
        beta_0 = foot_mean - beta_1*forearm_mean;
        drop _type_ _freq_ pro_var_sum squ_diff_forearm_sum forearm_mean foot_mea
run;
/*sum_set과 mean_set을 결합하여 beta1과 beta0만 저장할수 있게 만든다*/
```

- PROC REG 를 이용한 회귀분석 하기

```
PROC REG data = reg.aflength plots = diagnostics(stats = none); /*잔차의 산점도 피
MODEL foot = forearm / R; /* 잔차분석을 위한 통계값을 출력하는 옵션 사용*/
RUN;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	9498.22721	9498.22721	1045.11	<.0001
Error	30	272.64779	9.08826		
Corrected Total	31	9770.87500			

Root MSE	3.01467	R-Square	0.9721
Dependent Mean	241.18750	Adj R-Sq	0.9712
Coeff Var	1.24993		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-12.41920	7.86284	-1.58	0.1247
forearm	팔안쪽길이	1	1.05272	0.03256	32.33	<.0001

- 추정된 회귀모형을 통한 반응변수 Y의 평균 추정 값에 대한 신뢰구간 및 새로운 값에 대한 예측구간 구하기

```
PROC REG data = reg.aflength;
MODEL foot = forearm / CLM CLI;
RUN;
```

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	233	231.8117	0.6067	230.5726	233.0508	225.5315	238.0920	1.1883
2	242	241.2862	0.5329	240.1978	242.3746	235.0339	247.5384	0.7138
3	242	241.2862	0.5329	240.1978	242.3746	235.0339	247.5384	0.7138
4	238	242.3389	0.5341	241.2481	243.4297	236.0862	248.5916	-4.3389
5	246	243.3916	0.5373	242.2944	244.4889	237.1378	249.6454	2.6084

2 예제2 : Iris 데이터

- Iris 데이터는 붓꽃의 3가지 종(setosa, versicolor, virginica)에 대해 꽃받침 sepal과 꽃잎 petal의 길이를 정리한 데이터다.
  - 반응변수 Y : Sepal Length
  - 설명변수 X : Sepal Width

- 모형 : $Y = \beta_0 + \beta_1 X + \epsilon,\ \epsilon \sim N(0, \sigma^2)$

- 데이터 읽고 산점도 그리기

```
PROC IMPORT datafile = "C:\Users\User\Desktop\19학년도 회귀분석 및 실습\Data\Week2
out = iris_dat;
getnames = yes;
RUN;

PROC SGPLOT data = iris_dat;
SCATTER X = sepal_width Y = sepal_length;
RUN;
```

- Iris 데이터에서 변수 전체에 대한 상관관계를 알아보자

```
PROC CORR data = iris_dat FISHER(rho0 = 0) plots = matrix; /* 상관관계분석 FISHER
VAR sepal_length -- petal_width;
RUN;
```

- Sepal Length와 Sepal Width간 회귀분석하기 (회귀직선 구하기)

```
PROC REG data = iris_dat;
MODEL sepal_length = sepal_width;
RUN;
```

- 각 붓꽃의 종류에 따른 산점도 나타내기

```
PROC SGPLOT data = iris_dat;
SCATTER X = sepal_width Y = sepal_length;
by Species;
RUN;
```

- 각 붓꽃의 종류에 따른 상관관계 구하기

```
PROC CORR data = iris_dat FISHER(rho0 = 0) plots = matrix;
VAR sepal_length -- petal_width;
by Species;
RUN;
```

- 각 붓꽃의 종류에 따른 회귀분석 하기

```
PROC REG data = iris_dat;
MODEL sepal_length = sepal_width;
by Species;
RUN;
```

- 앞서 붓꽃의 종류를 고려하지 않고 회귀분석 한 결과와 붓꽃의 종류를 나눈 후 각각 회귀분석 한 결과가 어떠한 차이를 보이는가?

3 예제3 : Quadratic 데이터

- Quadratic 데이터는 두 변수 Y, X에 대한 자료이다. 두 변수에 대해 단순선형회귀모형을 적합하고 결과를 확인해보자.
  - 반응변수 Y : Y
  - 설명변수 X : X
  - 모형 : $Y = \beta_0 + \beta_1 X + \epsilon,\ \epsilon \sim N(0, \sigma^2)$
- 데이터 읽고 산점도 그리기

```
PROC IMPORT datafile = "C:\Users\User\Desktop\19학년도 회귀분석 및 실습\Data\Week2
out = quadratic;
getnames = yes;
RUN;

TITLE "Quadratic 데이터 산점도";
PROC SGPLOT DATA = quadratic;
SCATTER X = X Y = Y; /* X, Y 변수의 산점도 작성 */
REG X = X Y = Y;          /* 산점도에 추정회귀직선을 추가로 표시*/
RUN;
```

- X와 Y 변수의 회귀분석을 진행한 후, aflength 데이터의 회귀분석 결과와 차이점이 무엇인지 탐색하기.

```
PROC REG data = quadratic;
MODEL y = x;
RUN;
```