R 데이터(텍스트) 마이닝 및 시각화 분석

Seongmin Mun

26th January 2022

아주대학교
AJOU UNIVERSITY

Outline

회귀분석(regression analysis)

회귀분석?

- 자연현상 혹은 사회현상이 변수들의 인과관계에 의해 발생할때, 이를 수학적으로 설명하기 위해 사용되는 통계적 방법들중의 하나가 회귀분석(Regression analysis)이다.
- 통계학에서, 회귀 분석(回歸 分析, 영어: regression analysis)은 관찰된 연속형 변수들에 대해 두 변수 사이의 모형을 구한뒤 적합도를 측정해 내는 분석 방법이다.

단순 회귀분석(simple regression analysis)

실습문제

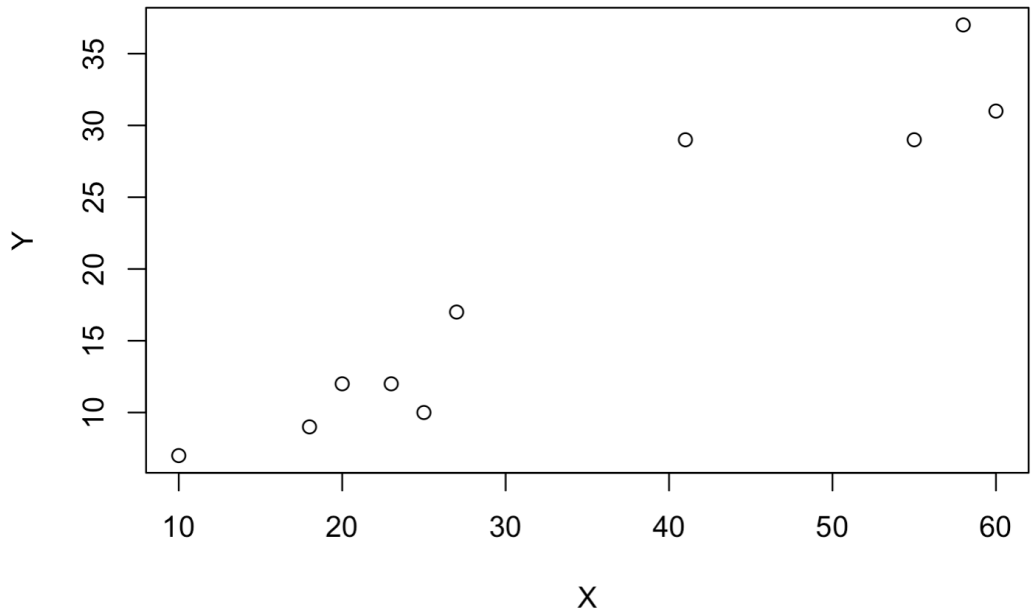
다음은 오염된 물고기의 수은 섭취량과 혈액내 수은량에 대한 자료이다. 해당 자료를 활용하여 수은섭취량(X)이 혈중수은량(Y)에 영향을 미치는지를 분석하시오.

```
> X <- c(18,20,23,41,60,55,27,58,10,25)
> Y <- c(9,12,12,29,31,29,17,37,7,10)
```

가설설정

- H0(귀무가설) : 수은섭취량(X)은 혈중수은량(Y)에 영향을 주지 않는다.
- H1(대립가설): 수은섭취량(X)은 혈중수은량(Y)에 영향을 준다.

산점도 확인




```
> #1.산점도 확인
> #par(mfrow=c(1,1))
> plot(X,Y)
```

회귀분산분석표

```
> lm.simple <- lm(Y~X)
> #회귀분산분석표: F값을 통해 회귀식의 유의성을 설명한다.
> anova(lm.simple)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
X          1 1005.83 1005.83  91.157 1.198e-05 ***
Residuals  8   88.27   11.03
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

회귀분산분석표

Source(요인)	DF(자유도)	SS(제곱합)	MS(평균제곱)	F(F값)
Reg(회귀)	K-1	SSR	MSR	F
Error(오차)	N-(K-1)-1	SSE	MSE	
total	N-1	SST		

회귀분산분석표

- SST(총제곱합, total sum of squares): Y의 관측값들이 가지는 총변동을 나타내는 제곱합
- SSE(오차제곱합, error sum of squares): 잔차들의 제곱합으로 Y의 총변동 중 설명이 안된 변동의 값
- SSR(회귀제곱합, regression sum of squares): Yi(Y의 개별값들)의 총변동 중 회귀식에 의해 설명된 변동값
- MS(평균제곱, mean square): 제곱합을 자유도로 나눈값
- MSR(회귀평균제곱, regression mean square)
- MSE(오차평균제곱, error mean square)
- F값

회귀분산분석표

Source(요인)	DF(자유도)	SS(제곱합)	MS(평균제곱)	F(F값)
Reg(회귀)	K-1	SSR	MSR	F
Error(오차)	N-(K-1)-1	SSE	MSE	
total	N-1	SST		

$$SST = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n}$$

$$SSR = \frac{\left(\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}\right)^2}{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}$$

SSE = SST - SSR

MSR = SSR / K-1

MSE = SSE / (n-(k-1)-1)

F = MSR / MSE

요인값 검증하기

SSR(회귀제곱합, regression sum of squares)

```
> #SSR(회귀제곱합, regression sum of squares)
> SSR<-function(X,Y){
+   XY<-NULL
+   X2<-NULL
+   for(i in 1:length(X)){
+     out1<-X[i]*Y[i]
+     out2<-X[i]*X[i]
+     XY<-c(XY,out1)
+     X2<-c(X2,out2)
+   }
+   outnum<-((sum(XY)-((sum(X)*sum(Y))/(length(X))))^2)/(sum(X2)-((sum(X)^2)/length(X)))
+   return(outnum)
+ }
> ssr<-SSR(X,Y)
> ssr
[1] 1005.828
```

요인값 검증하기

SST(총제곱합, total sum of squares)

```
> #SST(총제곱합, total sum of squares)
> SST<-function(Y){
+   YY<-NULL
+   for(i in 1:length(Y)){
+     out<-Y[i]*Y[i]
+     YY<-c(YY,out)
+   }
+   outnum<-(sum(YY)-(length(Y)*(mean(Y)^2)))
+   return(outnum)
+ }
> sst<-SST(Y)
> sst
[1] 1094.1
```

요인값 검증하기

SSR(회귀제곱합, regression sum of squares) & 회귀자유도 & 오차자유도

```
> #SSR(회귀제곱합, regression sum of squares)
> sse<-sst-ssr
> sse
[1] 88.27224
> #회귀자유도
> dfr <- 2-1
> dfr
[1] 1
> #오차자유도
> dfe <- 10-dfr-1
> dfe
[1] 8
```

요인값 검증하기

MSR(회귀평균제곱, regression mean square) &
MSE(오차평균제곱, error mean square) & F값

```
> #MSR(회귀평균제곱, regression mean square)
> msr<-ssr/dfr
> msr
[1] 1005.828
> #MSE(오차평균제곱, error mean square)
> mse<-sse/dfe
> mse
[1] 11.03403
> #F_value(F값)
> fvalue<-msr/mse
> fvalue
[1] 91.15688
```

결과해석

- F값(91.15688)이 F분포표의 F(1,8)=5.32보다 크므로 H0를 기각하고 H1을 받아들인다.
- H1 : 수은섭취량은 혈중수은량에 영향을 준다.

결과해석

F분포표

$\alpha = 0.05$

ν_2 \ ν_1	1
1	161.4
2	18.51
3	10.13
4	7.71
5	6.61
6	5.99
8	5.32

$\alpha = 0.05$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	99.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	9.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

회귀분석 결과

```
> #회귀분석 결과
> summary(lm.simple)

Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2792 -2.2541 -0.2594  1.5192  5.4872

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.14827    2.29187  -0.065     0.95
X            0.57710    0.06044   9.548  1.2e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.322 on 8 degrees of freedom
Multiple R-squared:  0.9193,    Adjusted R-squared:  0.9092
F-statistic: 91.16 on 1 and 8 DF,  p-value: 1.198e-05
```

결과해석

- 상수항의 p값은 0.95로 0.05 이상이고 X(수은섭취량)의 p값은 0.05이하이다.
- X(수은섭취량)은 95%의 신뢰수준으로 Y(혈중수은량)에 영향을 미친다.
- 회귀 모형에 대한 p값은 0.05이하 이므로 회귀모형은 유의하다.
- 설명력은 0.9193로 91%의 설명력을 나타낸다.