

다중 및 가중 회귀모형

Park Beomjin¹

¹University of Seoul

1 예제1 : 주택 판매가격 데이터

- 주택 판매가격 데이터(houseprice.csv)는 주택 판매가격(price; 천만원)과 이에 영향을 미칠 것으로 판단되는 4가지 설명변수인 세금(tax; 만원), 대지평수(ground; 평), 건물층수(floor; 층), 그리고 주택 나이(year; 년)를 27가구에 대해 조사한 데이터이다.
- 이에 대한 회귀모형을 $Y = \beta_0 + \beta_1 x_{\text{tax}} + \beta_2 x_{\text{ground}} + \beta_3 x_{\text{floor}} + \beta_4 x_{\text{year}} + \epsilon$ 로 설정하였을 때, 다음과 같은 가설에 대한 검정을 시행하여라.

$$\begin{aligned} H_{01} : \beta_{\text{tax}} &= \beta_{\text{floor}} = 0 \\ H_{02} : \beta_{\text{year}} &= \beta_{\text{ground}} = 0 \\ H_{03} : \beta_{\text{tax}} &= 2\beta_{\text{ground}} \end{aligned} \quad (1)$$

```
PROC REG data = reg.houseprice;
model price = tax -- year;
H1 : test tax = year = 0;
H2 : test year = ground = 0;
H3 : test tax = 2 * ground;
RUN; QUIT;
```

2 예제2 : Wine 데이터

- Wine 데이터는 포르투갈의 북부에 있는 reg wine과 white wine 샘플 데이터이다.
- 데이터는 와인의 품질(quality) 변수와 11개의 화학 측정 결과를 나타내는 변수를 가지고 있다.
- 목표는 화학 측정 결과를 기반으로 와인 품질을 예측하는 모형을 모델링 하는 것이다.
- Q1 : red wine 데이터와 white wine 데이터를 결합하고 type이라는 새로운 변수를 생성하여 reg wine의 데이터인 경우 "red", white wine 데이터인 경우 "white" 값을 갖도록 하여라.

```

DATA reg.wine;
set reg.redwine reg.whitewine(in = x);
length type \$ 5;
if x = 0 then type = "red";
else type = "white";
RUN; QUIT;

```

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality	type
1590	6.6	0.725	0.2	7.8	0.073	29	79	0.9977	3.29	0.54	9.2	5	red
1591	6.3	0.55	0.15	1.8	0.077	26	35	0.99314	3.32	0.82	11.6	6	red
1592	5.4	0.74	0.09	1.7	0.089	16	26	0.99402	3.67	0.56	11.6	6	red
1593	6.3	0.51	0.13	2.3	0.076	29	40	0.99574	3.42	0.75	11	6	red
1594	6.8	0.62	0.08	1.9	0.068	28	38	0.99651	3.42	0.82	9.5	6	red
1595	6.2	0.6	0.08	2	0.09	32	44	0.9949	3.45	0.58	10.5	5	red
1596	5.9	0.55	0.1	2.2	0.062	39	51	0.99512	3.52	0.76	11.2	6	red
1597	6.3	0.51	0.13	2.3	0.076	29	40	0.99574	3.42	0.75	11	6	red
1598	5.9	0.645	0.12	2	0.075	32	44	0.99547	3.57	0.71	10.2	5	red
1599	6	0.31	0.47	3.6	0.067	18	42	0.99549	3.39	0.66	11	6	red
1600	7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6	white
1601	6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6	white
1602	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6	white
1603	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6	white
1604	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6	white
1605	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6	white
1606	6.2	0.32	0.16	7	0.045	30	136	0.9949	3.18	0.47	9.6	6	white
1607	7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6	white
1608	6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6	white
1609	8.1	0.22	0.43	1.5	0.044	28	129	0.9938	3.22	0.45	11	6	white
1610	8.1	0.27	0.41	1.45	0.033	11	63	0.9908	2.99	0.56	12	5	white
1611	8.6	0.23	0.4	4.2	0.035	17	109	0.9947	3.14	0.53	9.7	5	white
1612	7.9	0.18	0.37	1.2	0.04	16	75	0.992	3.18	0.63	10.8	5	white
1613	6.6	0.16	0.4	1.5	0.044	48	143	0.9912	3.54	0.52	12.4	7	white
1614	8.3	0.42	0.62	19.25	0.04	41	172	1.0002	2.98	0.67	9.7	5	white
1615	6.6	0.17	0.38	1.5	0.032	28	112	0.9914	3.25	0.55	11.4	7	white

- Q2 : wine 전체 데이터와 reg 그리고 white wine에 대해 quality 변수를 반응변수로 하고 화학 측정 결과를 나타내는 11개 변수를 설명변수로 하는 회귀모형을 각각 적합하고 변수들의 다중공선성을 확인하여라.

```
PROC REG data = reg.wine;
model quality = fixed_acidity -- alcohol / vif;
ods select ParameterEstimates;
RUN; QUIT;
```

```
PROC REG data = reg.wine (where = (type = "red"));
model quality = fixed_acidity -- alcohol / vif;
ods select ParameterEstimates;
RUN; QUIT;
```

```
PROC REG data = reg.wine (where = (type = "white"));
model quality = fixed_acidity -- alcohol / vif;
ods select ParameterEstimates;
RUN; QUIT;
```

```
PROC GLM data = reg.wine;
class type;
model quality = fixed_acidity -- alcohol type / tolerance solution;
RUN; QUIT;
```

- Q3 : vif가 10 이상인 변수를 제외한 후 회귀모형을 다시 적합하고 결과를 해석하여라.

$$\hat{V}(\hat{\beta}_j) = \frac{s^2}{(n-1)\hat{V}(X_j)} \frac{1}{1-R_j^2} \quad (2)$$

```
PROC REG data = reg.wine (where = (type = "white"));
model quality = fixed_acidity -- alcohol / vif;
delete density;
RUN; QUIT;
```

```
/* Variance proportion */
```

```
PROC REG data = reg.wine (where = (type = "white"));
model quality = fixed_acidity -- alcohol / vif collin;
RUN; QUIT;
```

Root MSE		0.75136	R-Square	0.2819
Dependent Mean		5.87791	Adj R-Sq	0.2803
Coeff Var		12.78272		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	150.19283	18.80418	7.99	<.0001	0
fixed_acidity	1	0.06552	0.02087	3.14	0.0017	2.69144
volatile_acidity	1	-1.86318	0.11379	-16.37	<.0001	1.14116
citric_acid	1	0.02209	0.09577	0.23	0.8176	1.16522
residual_sugar	1	0.08148	0.00753	10.82	<.0001	12.64406
chlorides	1	-0.24728	0.54654	-0.45	0.6510	1.23682
free_sulfur_dioxide	1	0.00373	0.00084415	4.42	<.0001	1.78788
total_sulfur_dioxide	1	-0.00028575	0.00037806	-0.76	0.4498	2.23923
density	1	-150.28417	19.07451	-7.88	<.0001	28.23254
pH	1	0.68634	0.10538	6.51	<.0001	2.19636
sulphates	1	0.63148	0.10039	6.29	<.0001	1.13854
alcohol	1	0.19348	0.02422	7.99	<.0001	7.70696

Root MSE	0.75604	R-Square	0.2727
Dependent Mean	5.87791	Adj R-Sq	0.2713
Coeff Var	12.86235		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	2.06364	0.34823	5.93	<.0001	0
fixed_acidity	1	-0.05032	0.01491	-3.38	0.0007	1.35613
volatile_acidity	1	-1.95834	0.11386	-17.20	<.0001	1.12830
citric_acid	1	-0.02895	0.09615	-0.30	0.7634	1.15988
residual_sugar	1	0.02564	0.00255	10.05	<.0001	1.43521
chlorides	1	-0.95253	0.54252	-1.76	0.0792	1.20365
free_sulfur_dioxide	1	0.00477	0.00083907	5.68	<.0001	1.74463
total_sulfur_dioxide	1	-0.00086970	0.00037303	-2.33	0.0198	2.15317
pH	1	0.16517	0.08254	2.00	0.0454	1.33091
sulphates	1	0.41934	0.09731	4.31	<.0001	1.05664
alcohol	1	0.36269	0.01127	32.19	<.0001	1.64712

3 예제3: 탑승객 데이터

- 탑승객 데이터(airline_passengers.csv)는 1949년 1월부터 1960년 12월까지의 어느 공항의 탑승객 수에 대한 데이터이다.

```
PROC IMPORT datafile = "data folder path\airline_passengers.csv"
dbms = csv replace out = reg.passengers;
getnames = yes;
RUN;
```

```
DATA reg.passengers2;
set reg.passengers;
ind = _n_;
RUN;
```

- Q1 : 연도를 X(설명변수) 탑승객 수를 Y(반응변수)로 하는 회귀모형을 적합 및 분산의 동질성(homogeneity of variance)에 대해 검정하고 적합한 모형에 대해 논의하여라.

```
/* SAS CODE for Q1 */
PROC REG data = reg.passengers2;
model passengers = ind / spec r;
output out = reg.passenger_out p = pred r = resid;
RUN; QUIT;
```

- Q2 : 데이터를 이용하여 이분산성을 해결할 수 있는 가중치를 산출하는 방법을 생각해 보고 가중치를 산출하여라.

```
/* SAS CODE for Q2 */
```

```
DATA reg.abs_resid;  
set reg.passenger_out;  
abs_resid = abs(resid);  
RUN; QUIT;
```

```
PROC REG data = reg.abs_resid NOPRINT;  
model abs_resid = ind;  
output out = reg.passenger_abs_resid p = pred;  
RUN; QUIT;
```

- Q3 : 계산한 가중치를 이용하여 가중 최소제곱법(weighted least square)을 통해 회귀 모형을 적합 및 분산의 동질성 검정을 하여라.

```
/* SAS CODE for Q3*/
```

```
DATA reg.weight_dat;  
set reg.passenger_abs_resid;  
w = 1 / (pred**2);  
PROC REG data = reg.weight_dat;  
model passengers = ind / spec r;  
weight w;  
output out = reg.fpassenger p = wp student = wr r = wr0;  
RUN; QUIT;
```