Regression HW 2

CH03,04,05

1. 단순회귀에서 회귀제곱합,
$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
을 이차형식 $\boldsymbol{y}^\top B\boldsymbol{y}$로 표현하시오. 이이차형식의 분포를 구하고, 또한 기대치를 ⟨정리 5.1⟩에 의하여 구하시오.

2. 만약
$$y_1 = \beta_0 + \epsilon_1$$
$$y_2 = 2\beta_0 - \beta_1 + \epsilon_2$$
$$y_3 = \beta_0 + 2\beta_1 + \epsilon_3$$
이고, $\mathrm{E}(\epsilon_i) = 0, i = 1, 2, 3$이라면 β_0와 β_1의 최소제곱추정값은 무엇인가? $y_i, i = 1, 2, 3$의 함수로써 나타내어라. 그리고 이 경우의 잔차제곱합(residual sum of squares)을 구하시오.

3. 단순회귀모형
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ \ \epsilon_i \sim N(0, \sigma^2), \ i = 1, 2, \ldots, n$$
에서 각각의 x_i가 $cx_i\,(c \neq 0)$로 대체된다고 가정하자. $\hat{\beta}_0, \hat{\beta}_1, s^2_{y\cdot x}, R^2$과 $H_0 : \beta_1 = 0$에 대한 $t-$검정 결과는 어떤 영향을 받는가?

4. 생선을 잡아서 얼음창고에 일주일 동안 보관한 후에 생선의 신선도가 어느 정도 변하는가를 실험하였다. 신선도를 y로 놓고 10점 만점으로 하여 0점이 신선도가 전혀 없는 것이고 10점이 가장 좋은 경우이다. 설명변수 x는 생선을 잡은 지 x시간이 경과한 후에 얼음창고에 넣는 것을 가리킨다. 실험으로 10개의 데이터를 얻었다.

y(신선도)	8.5	8.4	7.9	8.1	7.8	7.6	7.3	7.0	6.8	6.7
x(경과시간)	0	0	3	3	6	6	9	9	12	12

(1) 선형회귀모형 ($y = \beta_0 + \beta_1 x + \epsilon$)이 타당한가를 유의수준 $\alpha = 0.05$를 사용하여 적합결여 검정을 행하라.

(2) 선형회귀모형이 타당한 경우, 신선도의 점수가 시간당 얼마만큼이나 떨어지는가를 95% 신뢰계수를 가지고 구간추정하라(즉, β_1의 구간추정).

5. 두 타이어회사 A, B에서 생산되는 타이어를 비교하기 위하여 고속도로에서 트럭이 달리는 상황을 모의실험(simulated experiment)하여 다음의 데이터를 얻었다. x는 트럭이 달리는 속도이고 y는 타이어가 마모되기까지의 총 주행거리이다.

$x_{1j} = x_{2j}$	10	20	30	40	50	60	70
$y_{1j}(A)$	9.8	12.5	14.9	16.5	22.4	24.1	25.8
$y_{2j}(B)$	15.0	14.5	16.5	19.1	22.3	20.8	22.4

(1) 산점도를 그리시오.

(2) 각 회사별로 속도와 총주행거리 간의 회귀모형을 구한다면, 두 개의 직선이 동일하다고 볼수 있는가? 유의수준 $\alpha = 0.05$로 가설검정하시오.

(3) 관심의 대상이 x가 증가함에 따라 y가 얼마나 증가하는가에 있다. 두 회사의 타이어에 대하여 각각 회귀모형을 적합했을 때, 기울기가 같은지 유의수준 5%로 검정하시오.

6 . **R 실습**. 아마존 강 수위 문제 아마존 강 유역은 지구상의 가장 큰 열대림 지역이지만 대부분의 다른 자연자원과 마찬가지로 개발의 손길이 미치면서 열대림이 급속히 파괴됐다. 1970년대 이후 아마존 상류지역에 도로가 건설되면서 인구가 빠르게 증가되었고 대규모의 삼림파괴가 이뤄졌다. 강수량과 유수량이 모두 영향을 받을 수 있기 때문에 이것은 결국 아마존 강 전체에 영향을 미치는 심각한 기후학적 및 수문학적 변화를 가져왔다. 다음의 표는 페루 이키토스(Iquitos)에서 1962년부터 1978년까지 기록한 아마존 강 최고수위 (High)와, 최저수위 (Low)를 기록한 것이다 (단위: 미터).

1962년부터 1969년까지의 데이터는 개발 이전에 수집된 데이터이고, 1970년부터 1978년까지의 데이터는 개발이후에 관측된 데이터를 나타낸다. 이 데이터는 아마존 상류지역의 삼림파괴가 아마존 유역의 강 수위에 변화를 일으켰는지 분석하고자 한다. 우리의 관심은 시간에 따른 아마존 강 수위 변화여부이다. 예를 들어, 우리가 다음을 적합한다면,

$$\texttt{High} = \beta_0 + \beta_1 \times \texttt{Year} + \epsilon$$

Table 1: 아마존 강 데이터 (Amazon River data)

Year	High (m)	Low (m)	Year	High (m)	Low (m)
1962	25.82	18.24	1971	27.36	21.91
1963	25.35	16.50	1972	26.65	22.51
1964	24.29	20.26	1973	27.13	18.81
1965	24.05	20.97	1974	27.49	19.42
1966	24.89	19.43	1975	27.08	19.10
1967	25.35	19.31	1976	27.51	18.80
1968	25.23	20.85	1977	27.54	18.80
1969	25.06	19.54	1978	26.21	17.57
1970	27.13	20.49			

(a) $\beta_1 = 0$은 시간에 따른 아마존 강의 최고수위에 아무런 (선형)변화가 없다는 것을 의미하고, (b) $\beta_1 > 0$은 아마존 강의 최고수위가 증가된 것을 의미하는데, 이것은 해마다 아마존 강의 흐르는 물이 늘어난 것을 나타낼 수 있다. (c) $\beta_1 < 0$은 시간에 따라 아마존 강의 최고수위가 낮아 진 것을 의미하는데, 이것은 해마다 아마존 강의 흐르는 물이 줄어든 것을 의미한다. 다음의 물음에 답하시오.

(1) `High`와 `Year`, `Low`와 `Year`, 그리고 `High`와 `Low`에 대해 산점도를 그리시오.

(2) `Year`에 대한 `High`, `Year`에 대한 `Low`, 그리고 `Low`에 대한 `High`의 회귀모형을 구하시오. 3개 회귀모형의 결과를 요약하고, 각 모형별로 회귀계수의 의미를 설명하시오.

(3) 이 자료를 근거로 우리는 삼림파괴가 아마존 강 수위의 변화를 일으킨다고 할 수 있는가?

(4) 아마존강의 최저수위와 최고수위와의 산점도를 1960년대, 1970년대 자료별로 다르게 그리고, 각각의 회귀선을 적합하시오.

(5) 아마존강의 최저수위와 최고수위와의 관계가 1960년대와 1970년대에 따라 차이가 있는가? 두 회귀모형의 동일성 여부를 유의수준 $\alpha = 0.01$에서 검정하시오.

(6) (4)에서 구한 두 회귀모형의 기울기가 같은지 유의수준 $\alpha = 0.01$에서 검정하시오.