Answer Key for Problem Set 4

남 준우 교수

1. Note that $R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$.

Since $\hat{Y}_i - \overline{Y} = b_2(X_i - \overline{X})$,

$$R^2 = \frac{b_2^2 \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \left[\frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (X_i - \vec{X})^2}\right]^2 \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$$

$$= \frac{\left\{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})\right\}^2}{\sum_i (X_i - \vec{X})^2 \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \frac{S_{XY}^{\ 2}}{S_X^{\ 2} S_Y^{\ 2}} = \gamma_{XY}^{\ 2}.$$

Since correlation coefficient (γ_{XY}) measures a linear relationship between X and Y, so does R^2.
Furthermore, since $-1 \le \gamma_{XY} \le 1$, $0 \le R^2 \le 1$ (Cauchy-Schwartz Inequality).

(Bonus) Note that when $\gamma_{XY} = \pm 1$ (perfect positive(or negative) relationship between X and Y), $R^2 = 1$.

(2) Done in the class.

(3) Note that $b = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (X_i - \vec{X})^2} = \frac{S_{XY}}{S_X^{\ 2}}$, and

similarly $d = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (Y_i - \overline{Y})^2} = \frac{S_{XY}}{S_Y^{\ 2}}$, $b \cdot d = \frac{S_{XY}^{\ 2}}{S_X^{\ 2} S_Y^{\ 2}} = \gamma_{XY}^{\ 2}$.(that is, $b \cdot d = R^2$)

Since $\left|\gamma_{XY}\right| \le 1$, $b \cdot d = \gamma_{XY}^{\ 2} \le 1$.

(Bonus) 1-(3)과 관련된 이야기: 그냥 한번 읽어볼 것.
◎ b is a slope estimate of 'regression of Y on X' and d is a slope estimate of 'regression of X on Y'. ⇒ 'regression of X on Y' is called as 'reverse regression').

(Bonus) Here we can easily conjecture that

R^2 of 'regression of Y on X'= R^2 of 'regression of X on Y'= $b \cdot d$.

Bonus 이상에서 Y를 X에 대해 회귀분석 했을 때의 R^2와 X를 Y에 대해 회귀분석했을 때의 R^2가 같다는 사실에서 우리는 R^2는 인과관계(causality)의 지표가 될 수 없음을 알 수 있다. 여기서 Causality 란 X, Y 두 변수가 있을 때 어느 변수가 원인이고 어느 변수가 결과인지를 판단하는 것을 말한다. 다시 말하면 Y를 X에 대해 회귀분석 했을 때의 R^2가 높다고 해서 X 가 원인이고 Y 가 결과임을 의미하는 것은 아니다는 것이다.

예를 들어

(1) 인간의 수를 캥거루의 수에 대해 회귀분석하면 높은 결정계수 값을 구하는데 이로부터 캥거루의 수가 증가하면 인간의 수가 증가한다고 할 수 있는가?

(2) 범죄 건 수를 경찰 수에 회귀분석하면 높은 결정계수 값을 구하는데 이는 경찰의 수가 증가하면 범죄 건 수가 증가한다는 의미인가?

이러한 예에서 나오는 문제점을 가성회귀(spurious regression)라 한다. 즉 이 경우 독립변수와 종속변수의 설정이 반대로 되었다는 것이다.

• 특히 여러 시계열 자료들은 같은 추세로 변하게 되는데 실제로 한 변수가 다른 변수에 대해 아무런 관계가 없다고 하더라도 같은 경기변동을 갖게 되어 회귀분석에서 높은 R^2값을 구하게 되는 문제점을 가성회귀라 하며 이로부터 인과관계를 잘못 판단할 수 있다.

이에 대해 자세한 사항은 시계열 분석에서 단위근 검정(unit root test) 등을 통해 설명되며 학부 수준을 상회하는 내용임.

2.

β_1에 대한 OLS 추정량을 구하면

$\hat{\beta}_1 = \bar{Y}$, $\hat{Y}_i - \bar{Y} = \hat{\beta}_1 - \hat{\beta}_1 = 0.$

따라서 결정계수 공식에서 분자의 값이 0이므로 $R^2 = 0$ 이다.

3,

(1) Since $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{52} e_i^2$, $\sum_{i=1}^{52} e_i^2 = (n-2)\hat{\sigma}^2 = 2.05 \times (52-2) = 102.50$.

(2) Since $s_{b_2} = \sqrt{\hat{V}(b_2)} = \sqrt{0.00088} = 0.0297$.

And since $\hat{V}(b_2) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{52}(X_i - \bar{X})^2}$, $\sum_{i=1}^{52}(X_i - \bar{X})^2 = \frac{\hat{\sigma}^2}{\hat{V}(b_2)} = \frac{2.05}{0.00088} = 2329.55$

(3) As the percentage of males18 years older who are high school graduates increase by 1 percentage point, the state's mean income increases by 0.15 (thousands of dollars) in average.

(4) Since $\bar{Y} = b_1 + b_2\bar{X}$, $b_1 = \bar{Y} - b_2\bar{X} = 14.07 - 0.15 \times 68.14 = 3.85$.

(5) Since $\sum_i (X_i - \bar{X})^2 = \sum_i X_i^2 - n\bar{X}^2$,

$\sum_i X_i^2 = \sum_i (X_i - \bar{X})^2 + n\bar{X}^2 = 2325.8 + 52 \times 68.14^2 = 243,764.90$

(6) Since $e_i = Y_i - (b_1 + b_2 X_i) = 12.28 - (3.85 + 0.15 \times 58.3) = -0.32$