Recap

Data $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n) \in \mathbb{R}^2$

$p = 2$

Best fit $y = b_0 + b_1 x$ where

$$\left.\begin{aligned} \sum y_i &= n b_0 + b_1 \sum x_i \\ \sum x_i y_i &= b_0 \sum x_i + b_1 \sum x_i^2 \end{aligned}\right\} \text{normal eq}^{\text{n}}\text{s}$$

Fitted value

$$\hat{y}_i = b_0 + b_1 x_i$$

Residual

$$e_i = y_i - \hat{y}_i$$

Sample mean

$$\bar{y} = \frac{1}{n} \sum y_i$$

$$SSTO = \sum (y_i - \bar{y})^2$$

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

Def$^{\text{n}}$ Coefficient of determination

$$R^2 := \frac{SSR}{SSTO}$$

Typically a good fit has R^2 close to 1.

Lemma i) $\sum e_i = 0$

ii) $\sum \hat{y}_i e_i = 0$

Proposition i) $SSTO = SSR + SSE$

ii) $0 \le R^2 \le 1$.

Proof of Prop (i) $\Rightarrow$ Prop (ii)

(i) implies $R^2 = \dfrac{SSR}{SSTO} = \dfrac{SSR}{SSR+SSE} = 1 - \dfrac{SSE}{SSTO}$

But $0 \le SSE, SSTO$. By (i), $0 \le SSE \le SSTO$,

So $0 \le R^2 \le 1$.

Proof of Prop (i)

$$\sum (y_i - \bar{y})^2 = \sum \left[(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \right]^2$$

$$= \sum \left\{ (\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i)^2 \right\} + 2 \underbrace{\sum (\hat{y}_i - \bar{y})(y_i - \hat{y}_i)}_{A}$$

$$A = \sum \hat{y}_i (y_i - \hat{y}_i) - \bar{y} \sum (y_i - \hat{y}_i)$$

$$= \sum \hat{y}_i e_i - \bar{y} \sum e_i$$

$= 0$ by Lemma 1.

So $SSTO = SSE + SSR$. $\boxed{\times}$

Proof of Lemma (i)

$$\sum e_i = \sum (y_i - b_0 - b_1 x_i)$$

$$= \sum y_i - n b_0 - b_1 \sum x_i$$

$$= 0$$ by first normal eq^n.

Proof of Lemma (i) use both normal eq^{n}s.

Matrix Notation $(p \geq 2)$

$$B = (X^t X)^{-1} X^t Y$$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1\,p-1} \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{n\,p-1} \end{pmatrix}$$

$$SSTO = Y^t Y - n\bar{y}^2$$

$$SSR = B^t X^t Y - n\bar{y}^2$$

$$SSE = Y^t Y - B^t X^t Y$$

$$R^2 = \frac{SSR}{SSTO}\,, \quad \text{Again } 0 \leq R^2 \leq 1.$$

Some statistics (skipping proofs)

Suppose

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i$$

where

$i = 1, 2, \cdots, n$

$x_{i1}, \cdots, x_{ip-1}$ are known constants

ε_i are independent $N(0, \sigma^2)$,

$\beta_0, \cdots, \beta_{p-1}$ parameters.

<u>Defn</u> $$MSR = \frac{SSR}{p-1}$$ regression mean square

$$MSE = \frac{SSE}{n-p}$$ Error mean square

$$F^* = \frac{MSR}{MSE}$$

<u>Theorem</u> If $\beta_1 = \beta_2 = \cdots = \beta_n = 0$ then F^* follows an F distribution with $p-1$ and $n-p$ degrees of freedom.

So to choose between the two hypotheses

$C_1 : \beta_1 = \beta_2 = \cdots = \beta_n = 0$

$C_2 : \beta_i \neq 0$ for at least one i

we use :

if $F^* \leq F(1-\alpha, p-1, n-p)$ then conclude C_1,

if $F^* > F(1-\alpha, p-1, n-p)$ then conclude C_2,

to control Type I errors at level α.

Example using the skin cream example

y : sales in district

x_1 : size of district

x_2 : per capita income of district

one can compute :

$$p = 3 \qquad B = (X^t X)^{-1} X^t Y = \begin{pmatrix} 3.4526 \\ 0.4960 \\ 0.0092 \end{pmatrix}$$

$$n = 15$$

$$MSR = \frac{1}{p-1}\left(Y^t Y - B^t X^t Y\right) = 26922.4$$

$$MSE = 4.74$$

$$F^* = \frac{MSR}{MSE} = 5680$$

Assuming α at 0.05 and assuming the ε_i are independ $N(0, \sigma^2)$, we require

$$F(0.95, 2, 12) = 3.89$$

Since F^* exceeds 3.89 we conclude

C_2: sales are related to population and income.

But is this relation useful for predictions,

well

$$R^2 = \frac{SSR}{SSTO} = 0.9989$$

so when the independent variables x_1 and x_2 are considered, the variation in sales is "99.9% explained".