5

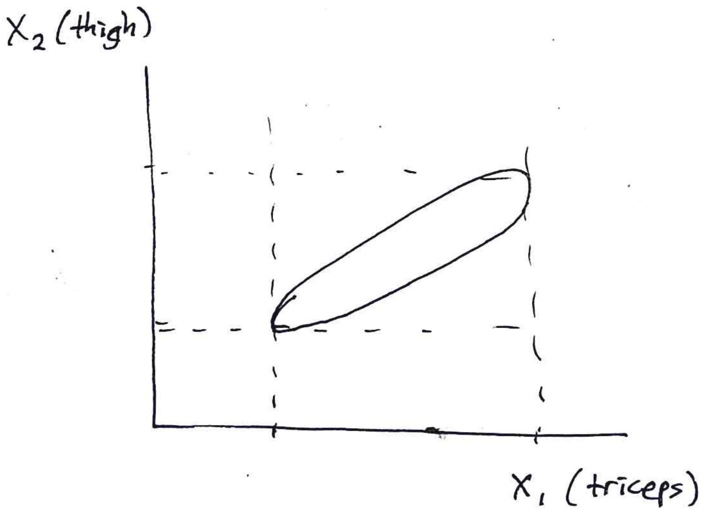
Multicollinearity exists when input variables are highly correlated among themselves

example : X_1 = triceps, X_2 = thigh, X_3 = midarm, y = body fat

(X_1, X_2, X_3 are easy measurements, y is cumbersome and expensive)

correlation matrix $r_{12} = .924$, $r_{13} = .458$, $r_{23} = .085$

$r_{y1} = .843$, $r_{y2} = .878$, $r_{y3} = .142$



Input space is multi-dimensional, not necessarily rectangular.

Analysis must include an understanding of how input variables are related

Thoughts :

(1) Interpretation of a regression coefficient depends on which other inputs are in the model

examples → Y = body fat, X_1 = weight, X_2 = abdomen size

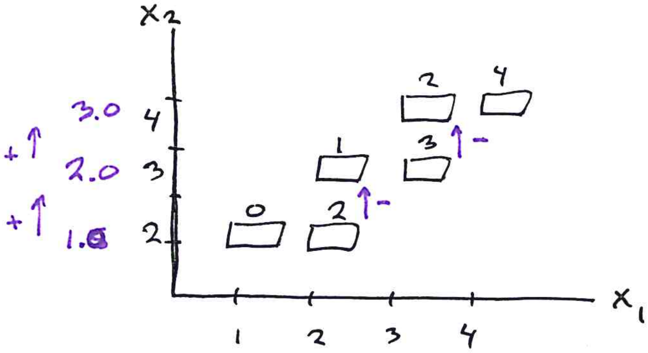
→ y = exam score, X_1 = difficulty, X_2 = ability

→ y = school quality, X_1 = money spent, X_2 = teacher quality

→ y = wins, X_1 = shots on goal, X_2 = goals scored

⑥

simple example: $E(Y) = 2X_1 - X_2$, $r_{12} >> 0$.



X_2 has a negative partial effect, given a fixed level of X_1

X_2 has a positive marginal effect, not accounting for the level of X_1

example: $\hat{Y} = 117.085 + 4.334X_1 - 2.857X_2 - 2.186X_3$

(recall that $r_{Y2} >> 0$)

---

(2) If an input is highly correlated with response, additional inputs are limited in how much new information can be provided.

example:

(recall that $r_{Y1} = .843$, $r_{Y2} = .878$)

$SSR(X_1) = 352.27$ $\quad R^2(X_1) = .711$

$SSR(X_2|X_1) = 33.17$ $\quad R^2(X_1, X_2) = .778$

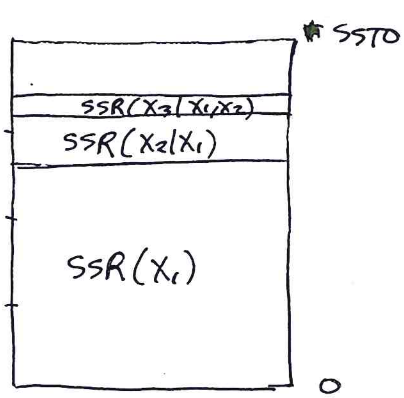
$SSR(X_3|X_1, X_2) = 11.55$ $\quad R^2(X_1, X_2, X_3) = .801$

$F^* = 21.52$ ($p = .000$)

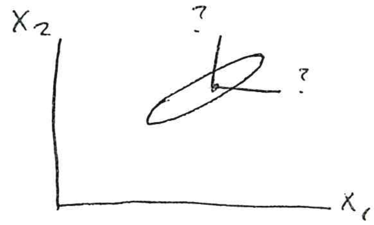
$t_1^* = 1.437$ ($p = .170$)

$t_2^* = -1.106$ ($p = .285$)

$t_3^* = -1.370$ ($p = .190$)



(3) Estimated regression coefficients have large sampling variance.



idea: We have little information on the effect of X_1 (X_2) with X_2 (X_1) held fixed.

example:

Model	b_1 ($SE(b_1)$)	b_2 ($SE(b_2)$)
X_1	.8572 (.1288)	—
X_2	—	.8565 (.1100)
X_1, X_2	.2224 (.3034)	.6594 (.2912)
X_1, X_2, X_3	4.334 (3.016)	-2.857 (2.582)

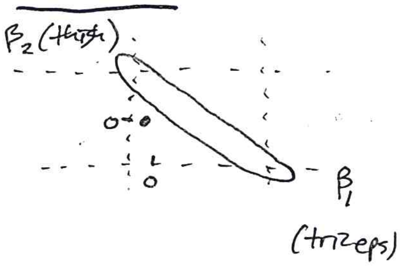
(4) Prediction and mean estimation (fitted regression function) is not affected by multicollinearity within the input space

example:

X_1	X_2	X_3	CI for μ_h
25	·	·	[18.6, 21.3]
25	50	·	[18.0, 20.7]
25	50	29	[17.9, 20.5]

(5) Regression coefficient estimates are highly correlated when input variables are highly correlated.

example:



$\beta_1 > b_1$ is compatible with $\beta_2 < b_2$

$\beta_1 < b_1$ is compatible with $\beta_2 > b_2$.

(0,0) is contained in $CI_{\beta_1} \times CI_{\beta_2}$

but (0,0) not contained in $CE(\beta_1, \beta_2)$

[illustrates the need to think in higher dimensions]

⑧

Investigating relationships in higher dimensions requires higher level statistical methods, such as regression analysis. Two-dimensional methods and graphs are insufficient.

example: (X_1 = triceps (P1), X_2 = thigh (P2), X_3 = midarm (W))

$r_{1W} = .458$, $r_{2W} = .085$ (moderate, small correlations between W, X_1 and W, X_2)

$r^2_{1W} = .2096$, $t_1^* = 2.185$ ($F_1^* = 4.77$), $p = .04$

$r^2_{2W} = .0072$, $t_2^* = 0.361$ ($F_2^* = 0.13$), $p = .72$

marginal effect tests
$M_2: w = \beta_0 + \beta_2 X_2 + \varepsilon$
$M_0: w = \beta_0 + \varepsilon$

The observed data is compatible with the reduced model. It is not necessary to add thigh measurement to the no effects model for predicting midarm measurement.

Now look at what happens when both X_1, X_2 are used to model w.

$r^2_{3,12} = .99$, $F^*_{12} = 880.7$, $P_{12} = .000$ (very high correlation between X_3 and (X_1, X_2))

$t^*_{1|2} = 41.82$ ($p = .000$) ($F^*_{1|2} = 1748.6$), $t^*_{2|1} = -37.26$ ($p = .000$) ($F^*_{2|1} = 1388.6$)

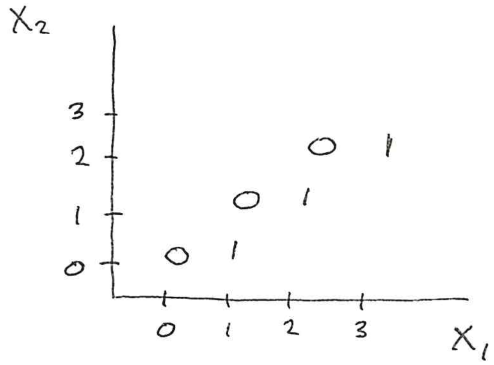
partial effect test
$M_{12}: w = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
$M_1: w = \beta_0 + \beta_1 X_1 + \varepsilon$

The observed data is not compatible with the reduced model.

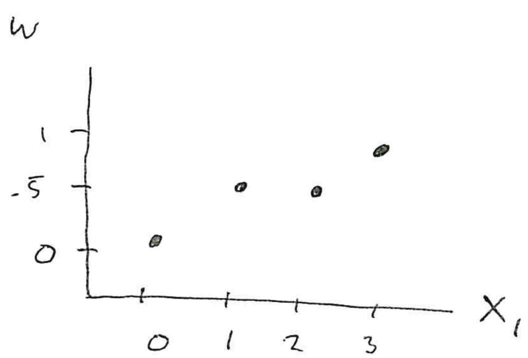
we accept the addition of thigh measurement as a predictor for midarm to the model which already includes triceps.

8 (b)

simple example: $w = X_1 - X_2$, X_1, X_2 are positively correlated



$Corr(w, (X_1, X_2))$ is large



$Corr(w, X_1)$ is moderate

$Corr(w, X_2)$ is small

example: $\widehat{midarm} = 62.33 + 1.88\ triceps - 1.6\ thigh$

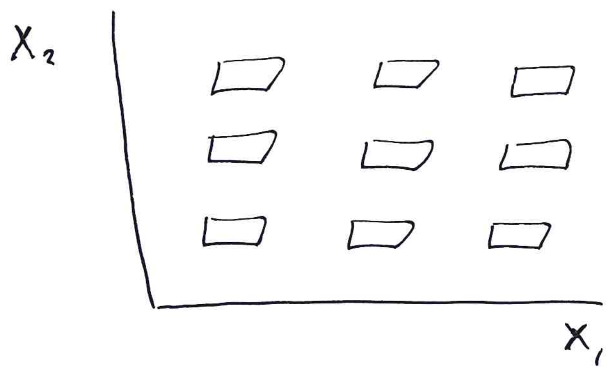
Orthogonal Design Matrix :

Def:

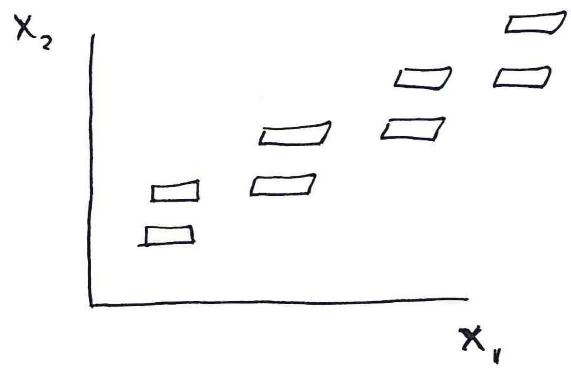
[A design is orthogonal if $X'X$ is diagonal.]

idea:

Orthogonal Design



Correlated Inputs



marginal effect of X_1: average $\updownarrow \updownarrow$, take differences

partial effect of X_1: differences $\leftrightarrow \leftrightarrow$, then average

partial effect of X_1 is determined for constant X_2 conditions

under orthogonality, levels of X_1 are the same across X_2.

example: study of coal conservation into oils, Canadian lignite

y = conversion percentage

X_1 = temperature, X_2 = molar ratio $\left(\frac{CO}{H_2}\right)$
(380, 460) °C $(\frac{1}{4}, \frac{3}{4})$

X_3 = pressure, X_4 = contact time
(7.1, 11.1) MPa   (10, 50) min

$n = 16$, $r_{x_i, x_k} = 0$ (2^4 Design)

To get $X'X$ diagonal, we use coded variables:

$$X_i = \begin{cases} +1, & \text{input at high level} \\ -1, & \text{input at low level} \end{cases}$$

Note that $Cov(\underline{b}) = \sigma^2 (X'X)^{-1}$ is diagonal. ($Cov(b_\ell, b_k) = 0$ for all $\ell \neq k$)

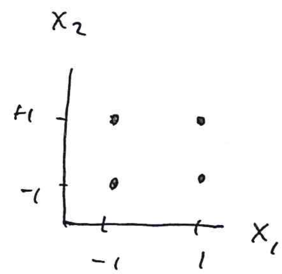
Also, $SSR(X_\ell \mid X_k, k \neq \ell) = SSR(X_\ell)$

(Thus, there is no ambiguity in defining the effect of an input)

For orthogonal designs, regression coefficient estimates and variation explained by an input do not depend on which other inputs are included in the model.

<u>example</u>: 2^k design, $b_\ell = \frac{1}{2}(\bar{Y}_{hi} - \bar{Y}_{\ell o})$

$$t^*_\ell = \frac{b_\ell}{SE(b_\ell)} \qquad SE(b_\ell) = \sqrt{\frac{MSE}{n}}$$



$b_1 = 5.0$, $b_2 = 3.15$, $b_3 = 3.975$, $b_4 = 1.862$, $SE(b_\ell) = 1.408$

($p_1 = .005$), ($p_2 = .047$), ($p_3 = .017$), ($p_4 = .213$)

selected model includes X_1, X_2, X_3.

$b_1 = 5.0$, $b_2 = 3.15$, $b_3 = 3.975$, $SE(b_\ell) = 1.451$ (MSE increases when X_4 is removed)

($p_2 = .051$)

Do not get stuck in dichotomous thinking in model selection.