Model Testing

example, photography studios

X_1 = adults, X_2 = income, y = sales

$r_{12} = .78$ $\quad$ $r_{y1} = .945$

$r_{y2} = .836$

Approach 1: t-statistics

$H_0 : \beta_\ell = 0$, $\quad t_\ell^* = \dfrac{b_\ell}{SE(b_\ell)}$, $\quad$ null distribution $t(n-p)$

example: $\quad t_1^* = 6.87$ $(p = .000)$ $\quad t_2^* = 2.305$ $(p = .033)$ $\quad$ R: summary(reg.mod)

The data supports a model which includes both predictors.

(t_ℓ^* is testing the partial effect of input X_ℓ, accounting for the effects ~~These are partial effect tests~~ of all other inputs.
Recall that β_ℓ represents ℓ^{th} input effect, with all other inputs held fixed)

Approach 2: General Linear Test

example: $\quad$ candidate models

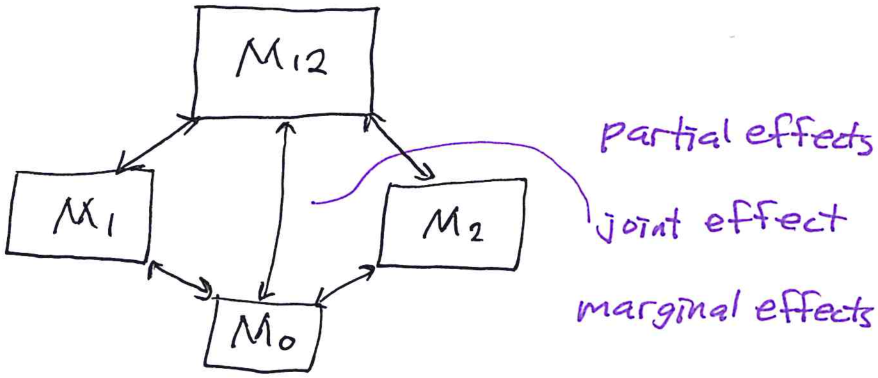
M_{12}, $\quad y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ $\quad$ both (both.mod)

M_1, $\quad y = \beta_0 + \beta_1 X_1 + \varepsilon$ $\quad$ (adults.mod)

M_2, $\quad y = \beta_0 + \beta_2 X_2 + \varepsilon$ $\quad$ (income.mod)

M_0, $\quad y = \beta_0 + \varepsilon$ $\quad$ (null.mod)

② testing a reduced model against a full model



(F): M_{12} , (R): M_0 (test for joint effect of (X_1, X_2) on response Y.)

$$SSE(F) = \sum_{i}^{n} (Y_i - \hat{Y}_i)^2 = SSE, \quad df_E(F) = n - p$$

$(\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2})$

$$SSE(R) = \sum_{i}^{n} (Y_i - \bar{Y})^2 = SSTO, \quad df_E(R) = n - 1$$

Define $SSR = SSTO - SSE$, $df = p - 1$

test statistic: $F^* = \dfrac{MSR}{MSE}$, null distribution $F(p-1, n-p)$

<u>example</u>: $F^* = 99.103$, $p = .000$ — R: anova(null.mod, both.mod)

The data supports models which include at least one of the predictors. $(R^2 = .9167)$

$SSR(X_1, X_2)$ measures variation in Y explained by inputs X_1, X_2.

(F^* is testing the joint effect of all inputs $(X_1, \ldots, X_r)$.)

③

Now consider testing for the individual effects of X_1, X_2.

(X_1 = adults (pop size) , X_2 = income (community wealth))

	test for X_1 effect	test for X_2 effect
marginal effect	(F): M_1 , (R): M_0	(F): M_2 , (R): M_0
partial effect	(F): M_{12} , (R) M_2	(F): M_{12} , (R): M_1

<u>example</u> : (F) M_1, (R) M_0

R

anova(null.mod, adults.mod)

$$SSE(R) = SSTO \quad , \quad SSE(F) = SSE(X_1) = \sum_{i}^{n} (Y_i - \hat{Y}_i(X_1))^2$$

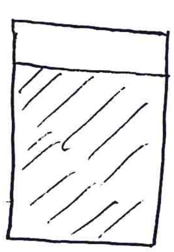
$$\hat{Y}_i(X_1) = b_0^{(1)} + b_1^{(1)} X_{i1}$$

$\Rightarrow$

$$SSR(X_1) = SSTO - SSE(X_1) \quad , \quad df = (n-1) - (n-2) = 1$$

[$SSR(X_1)$ measures the variation in response (sales) explained by X_1 (number of adults), not accounting for information on X_2.]

$F_1^* = 157.22$, $SSR(X_1) = 23372$



$SSTO = 26196.2$

$SSR(X_1) = 23372$

$R^2(X_1) = .89$

F_1^* is testing the marginal effect of X_1, not accounting for information on X_2.

④

R

anova(adults.mod, both.mod)

example : (F) M_{12} , (R) M_1

$$SSE(R) = SSE(X_1) \quad , \quad SSE(F) = SSE(X_1, X_2)$$

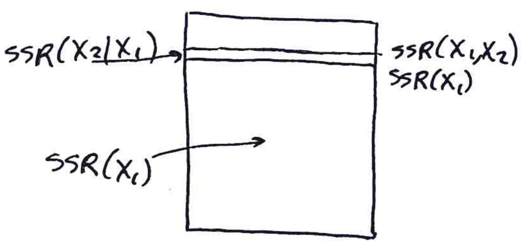
$\Rightarrow$

$$SSR(X_2 | X_1) = SSE(X_1) - SSE(X_1, X_2)$$
$$\vdots$$
$$= SSR(X_1, X_2) - SSR(X_1)$$

extra sum of squares $SSR(X_2|X_1)$ measures the variation in response (sales) explained by X_2 (community income) beyond that explained by X_1 (community size).

$F^*_{2|1} = 5.31$ ($p = .03$) , $SSR(X_2|X_1) = 643.5$

$F^*_{2|1}$ is testing the partial effect of X_2, after accounting for the effect of X_1



$SSR(X_2) = 18300$

$SSR(X_1|X_2) = 5715$

sequential sum of squares : $SSR(X_1, X_2) = SSR(X_1) + SSR(X_2|X_1)$

$= SSR(X_2) + SSR(X_1|X_2)$

R: lm(y ~ X1 + X2)

anova(mod)

thought experiments:

- burgers after pizza
- two good catchers on the same team

5

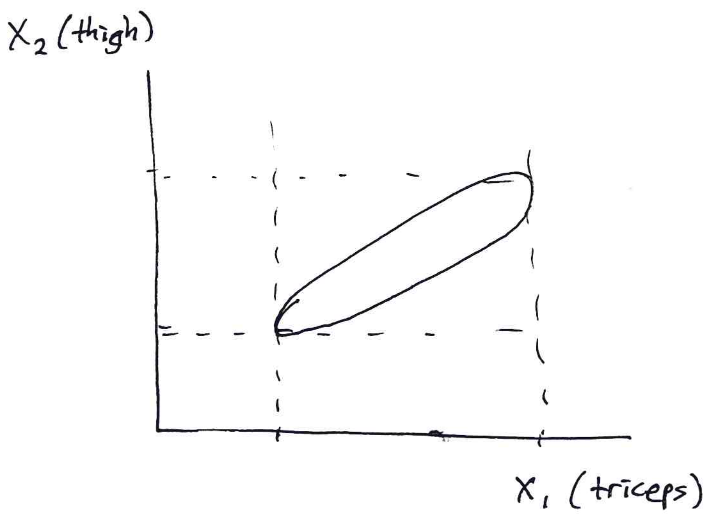
Multicollinearity exists when input variables are highly correlated among themselves

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

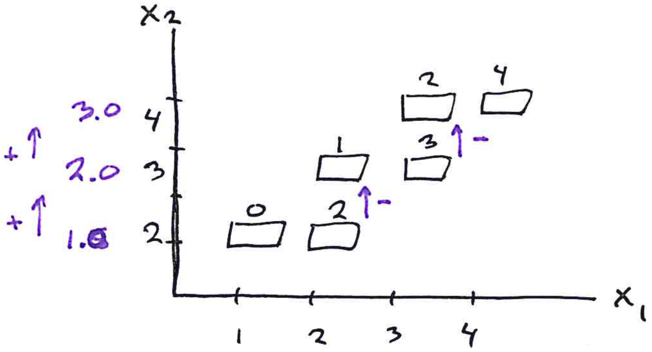
→ y = exam score, X_1 = difficulty, X_2 = ability

→ 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 $r_{Y1} = .843$, $r_{Y2} = .878$)

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

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

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

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

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

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

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