Factor Predictor

A factor predictor is a predictor that takes a discrete set values on "nominal" scale (non numerical)

e.g. Drug A, B, C

Femal, Male

Consider the case where the predictor takes M possible values.

$$X = \begin{cases} 1 \\ 2 \\ \vdots \\ M \end{cases}$$

We define M "dummy" or indicator variables

$$X_m = \begin{cases} 1 & X = m \\ 0 & \text{otherwise} \end{cases}$$

e.g.

	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$
Drug A	1	0	0
B	0	1	0
C	0	0	1

redundant

Remove last column and using only M−1 variables.

$$X^{(l)} = \begin{cases} 1 & X = l \\ 0 & \text{otherwise} \end{cases} \quad \text{for } l = 0 \cdots M-1$$

when $l = M \Rightarrow$ baseline level

We may write X_1 - categorial with M levels.

$$E[Y \mid X_1] = \beta_0 + \sum_{l=1}^{M-1} \beta_l X_1^{(l)}$$

(1) $X=1 \quad (X^{(1)}, X^{(2)} \cdots X^{(M-1)}) = (1\ 0\ 0 \cdots 0)$

$E[Y|X=1] = \beta_0 + \beta_1$

(2) $X=2 \quad (X^{(1)}, X^{(2)} \cdots X^{(M-1)}) = (0\ 1\ 0 \cdots 0)$

$E[Y|X=2] = \beta_0 + \beta_2$

$\vdots$

$(M-1) \quad X=M-1 \quad (X^{(1)}, X^{(2)} \cdots X^{(M-1)}) = (0\ 0 \cdots 0\ 1)$

$E[Y|X=M-1] = \beta_0 + \beta_{M-1}$

$(M) \quad X=M \quad (X^{(1)}, X^{(2)} \cdots X^{(M)}) = (0\ 0\ 0 \cdots 0)$

$E[Y|X=M] = \beta_0$

Therefore,

$$\beta_1 = E[Y|X=1] - E[Y|X=M]$$
$$\beta_2 = E[Y|X=2] - E[Y|X=M]$$
$$\vdots$$
$$\beta_{M-1} = E[Y|X=3] - E[Y|X=M]$$

} effect contrast between level $1 \cdots M-1$ and M

$$\beta_0 = E[Y|X=M] \quad \text{baseline level effect}$$

$$\Rightarrow \beta_l' = \begin{cases} \beta_0 & l=M \\ \beta_l + \beta_0 & l = 1 \cdots M-1 \end{cases}$$

$$\Rightarrow E[Y|X] = \sum_{l=1}^{M} \beta_l' X^{(l)} \qquad \text{b.c. } X = \begin{bmatrix} \\ \\ \\ \\ \\ \end{bmatrix}$$

e.g. $X=M \quad (X^{(1)} \cdots X^{(M)}) = (0\ 0 \cdots 0\ 1)$

$E[Y|X=M] = \beta_M' = \beta_0$

$X=1 \quad (X^{(1)} \cdots X^{(M)}) = (1\ 0 \cdots 0\ 0)$

$E[Y|X=1] = \beta_1' = \beta_1 + \beta_0$

In this case, β_m' can be interpreted directly as the effect of level m, for $m = 1 \cdots M$

Interaction

For two distinct predictors, we may consider the joint effect of the predictors simultaneously.

continuous predictors

We just include product term e.g. $X_1 \cdot X_2$

$$\Rightarrow E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 \cdot X_2$$

interaction

↓ modification of the effect of outcome of X_1 on the presense of X_2 (or vice versa)

$$E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 \cdot X_2$$

$$= \beta_0 + (\beta_1 + \beta_{12} X_2) X_1 + \beta_2 X_2 \quad \leftarrow [1 \; X_1 \; X_2] \begin{bmatrix} \beta_0 \\ \beta_1 + \beta_2 X_2 \\ \beta_2 \end{bmatrix}$$

$$= \beta_0 + \beta_1 X_1 + (\beta_2 + \beta_{12} X_1) X_2 \quad \leftarrow [1 \; X_1 \; X_2] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 + \beta_{12} X_1 \end{bmatrix}$$

interaction and correlation

correlation $X_1 \perp\!\!\!\perp X_2$ $Cov(X_1, X_2)$

interaction $D(Y, X_1) | X_2$

factor predictor

For the factor predictors with M_1 and M_2 levels, an interaction term consider all possible combinations of the level of each factor.

i.e. There are $M_1 \times M_2$ possible combinations

$$E[Y|X] = \underbrace{\sum_{m=1}^{M_1} \beta_m X^{(m)}}_{X_1} + \underbrace{\sum_{m'=1}^{M_2} \beta_{m'} X^{(m')}}_{X_2} + \sum_{m}^{M_1} \sum_{m'}^{M_2} \beta_{mm'} X^{(m)} X^{(m')}$$

If the intercept is used, then we could just use $(M_1 - 1)(M_2 - 1)$ terms to represent the interaction.

$$E[Y|X] = \beta_0 + \sum_{m=1}^{M_1 - 1} \beta_m X^{(m)} + \sum_{m'=1}^{M_2 - 1} \beta_{m'} X^{(m')} + \sum_{m}^{M_1 - 1} \sum_{m'}^{M_2 - 1} \beta_{mm'} X^{(m)} X^{(m')}$$

e.g. X_1 – categorical A, B, C

X_2 – categorical Female, Male

X_1	$X_1^{(1)}$	$X_1^{(2)}$	$X_1^{(3)}$
A	1	0	0
B	0	1	0
C	0	0	1

X_2	$X_2^{(1)}$	$X_2^{(2)}$
Female	1	0
Male	0	1

Model 1 (with main effect only)

$$E[Y|X] = \underbrace{\beta_1 X^{(1)} + \beta_2 X_1^{(2)} + \beta_3 X_1^{(3)}}_{X_1} + \underbrace{\beta_4 X_2^{(1)} + \beta_5 X_2^{(2)}}_{X_2}$$

Model 2 (with main effect only)

$$E[Y|X] = \beta_0' + \beta_1' X_1^{(1)} + \beta_2' X_1^{(2)} + \beta_4' X_2^{(1)}$$

X_1	X_2	$X_1^{(1)}$	$X_1^{(2)}$	$X_2^{(1)}$	Model
A	Female	1	0	1	$\beta_0' + \beta_1' + \beta_4'$
B	Female	0	1	1	$\beta_0' + \beta_2' + \beta_4'$
C	Female	0	0	1	$\beta_0' + \beta_4'$
A	Male	1	0	0	$\beta_0' + \beta_1'$
B	Male	0	1	0	$\beta_0' + \beta_2'$
C	Male	0	0	0	β_0'

β_2' circled; arrow from β_2' to the explanation below.

$\beta_2' = E[Y|X_1 = B, X_2 = \text{Male}] - E[Y|X_1 = C, X_2 = \text{Male}]$

the effect contrast between.

Model 2 (with interaction term)

$$E[Y|X] = \beta_0' + \underbrace{\beta_1' X_1^{(1)} + \beta_2' X_1^{(2)}}_{X_1} + \underbrace{\beta_4' X_2^{(1)}}_{X_2} + \underbrace{\beta_{14}' X_1^{(1)} X_2^{(1)}}_{A} + \underbrace{\beta_{24}' X_1^{(2)} \cdot X_1^{(1)}}_{B}$$

not allowed (?)

When there is β_0, total number of interactions between X_1 (M_1 levels) and X_2 (M_2 levels) is

$$(M_1 - 1)(M_2 - 1) = 2 \times 1 = 2$$

X_1	X_2	$X_1^{(1)}$	$X_1^{(2)}$	$X_2^{(1)}$	$E[Y \mid X_1, X_2]$
A	Female	1	0	1	$\beta_0' + \beta_1' + \beta_4' + \beta_{14}'$
B	Female	0	1	1	$\beta_0' + \beta_2' + \beta_4' + \beta_{24}'$
C	Female	0	0	1	$\beta_0' + \beta_4'$
A	Male	1	0	0	$\beta_0' + \beta_1'$
B	Male	0	1	0	$\beta_0' + \beta_2'$
C	Male	0	0	0	β_0'

Model selection

Criterion: Generalization error

We estimate the model by minimizing least-squares

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{n} (Y - X\beta)^T (Y - X\beta)$$

Let $\hat{m}(x) = x\hat{\beta}$

To measure how well $\hat{m}(x)$ can predict, we define generalization error

Imagine there are new data points (Y^0, X^0) $X^0 = (1, x_1^0 \cdots x_p^0)$

generalization error $G = E[(Y^0 - \hat{m}(X^0))^2]$

Training error $T = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}(x_i))^2$

T is a poor estimation of G, usually $T < G$

The training data is obtained from the model

$$y = X\beta + \varepsilon$$

From this data we obtain $\hat{\beta}$. Our fitted model is

$$\hat{m}(x) = x\hat{\beta}$$

The **training error** (in-sample error)

$$MSE = \frac{SSRes}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\beta)^2$$

Now imagine we get a new data at the same x_i

$$x_i^0 = x_i \qquad \underset{n\times 1}{y^0} = \underset{n\times p}{X^0}\underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon^0} = X\beta + \varepsilon^0$$

Testing error (out-of-sample prediction error)

$$\frac{1}{n}\sum_{i=1}^{n}(y_i^0 - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i^0 - x_i^0\hat{\beta})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i^0 - x_i\hat{\beta})^2$$

We can see that

$$E\Big[\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i^0 - \hat{y}_i)^2}_{\text{testing error}}\Big] = \frac{1}{n}\sum_{i=1}^{n}E[(y_i^0 - \hat{y}_i)^2] = \frac{1}{n}\cdot n\,\underbrace{E[(y^0 - \hat{m}(x^0))^2]}_{G}$$

$$\Rightarrow E\Big[\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i^0 - \hat{y}_i)^2}_{\text{testing error}}\Big] = E\Big[\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{training error}}\Big] + \frac{2}{n}\sigma^2 p = G$$

$$MSE = \frac{SSRes}{n}$$

proof

$$E[(y_i - \hat{y}_i)^2] = Var(y_i) + Var(\hat{y}_i) - 2Cov(y_i, \hat{y}_i) + (E(y_i) - E(\hat{y}_i))^2$$

$$E[(y_i^0 - \hat{y}_i)^2] = Var(y_i^0) + Var(\hat{y}_i) - 2Cov(y_i^0, \hat{y}_i) + (E(y_i^0) - E(\hat{y}_i))^2$$

y_i^0 is independent to y_i, but has the same distribution

$$\Rightarrow E(y_i^0) = E(y_i) \quad Var(y_i^0) = Var(y_i) \quad Cov(y_i^0, y_i) = 0$$

$$E((y_i^0 - \hat{y}_i)^2) = Var(y_i) + Var(\hat{y}_i) + (E(y_i) - E(\hat{y}_i))^2$$

$$= E[(y_i - \hat{y}_i)^2] + 2Cov(y_i, \hat{y}_i)$$

Average over data points for new observations.

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(y_i^0-\hat{y}_i)^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\right] + \frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}(y_i,\hat{y}_i)$$

For a linear model, $\mathrm{Cov}(y_i,\hat{y}_i) = \sigma^2 H_{ii}$

$$\Rightarrow E\left[\frac{1}{n}\sum_{i=1}^{n}(y_i^0-\hat{y}_i)^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\right] + \frac{2}{n}\cdot\sigma^2\,\mathrm{tr}H$$

$\downarrow$ C_p — testing error — training error — MSE — $\sum_{i=1}^{n} h_{ii} = p$

$$\Rightarrow C_p = \mathrm{MSE} + \mathrm{penalty}$$

Mallow's C_p

$$C_p = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \frac{2}{n}\hat{\sigma}^2 p$$

Smaller C_p are preferred.

R^2 / adjusted R^2

$$R^2 = 1 - \frac{SS_{Res}}{SST} = 1 - \frac{MSE}{SST/n}$$ ignore p

$$\text{adj } R^2 = 1 - \frac{MSE\cdot\frac{n}{n-p}}{SST/n} \rightarrow MSE\,\frac{n}{1-p/n} \approx MSE\left(1+\frac{p}{n}\right) = MSE + MSE\cdot\frac{p}{n}$$

$$\xrightarrow[n\to\infty]{} MSE + \sigma^2\frac{p}{n} = C_p$$

- Maximize R^2 / adj R^2

AIC & BIC

$$AIC = n\cdot\ln(MSE) + 2\cdot p = n\ln\frac{SS_{Res}}{n} + 2p$$

$$BIC = n\cdot\ln(MSE) + \ln(n)\cdot p$$

- Minimize AIC & BIC