MLR with Qualitative Predictors

Dummy Variable and Baseline Category

An economist conducting a study on the insurance industry aimed to establish a relationship between the adoption speed of a specific insurance innovation (Y) and the size of the insurance firm (X1) as well as the type of firm (X2), which could be either a stock company or a mutual company. Notably, X2 is a qualitative variable.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$X_1 = \text{size of firm}$$
$$X_2 = \text{mutual company or stock company}$$

Comment:

- Indicator variables with c classes will be represented by c-1 indicator variables, each taking on binary values of either 0 or 1. The 0 status is generally considered the "**baseline**" by default. In R, it is based on the alphabetical order by default.
- Indicator variables are frequently called ***dummy variables***, or ***binary variables.***

$X_2 = 0\ (The\ mutual\ company)$. This is the baseline.
$X_2 = 1\ (The\ stock\ company)$

- In situations where there are three kinds of companies, namely mutual, stock, and other (i.e., c=3), it is necessary to define two indicator variables (i.e., c-1) to represent them.

$$X_2 = \begin{matrix} 1 & if\ mutual\ company \\ 0 & Elsewise \end{matrix}$$
$$X_3 = \begin{matrix} 1 & if\ stock\ company \\ 0 & Elsewise \end{matrix}$$

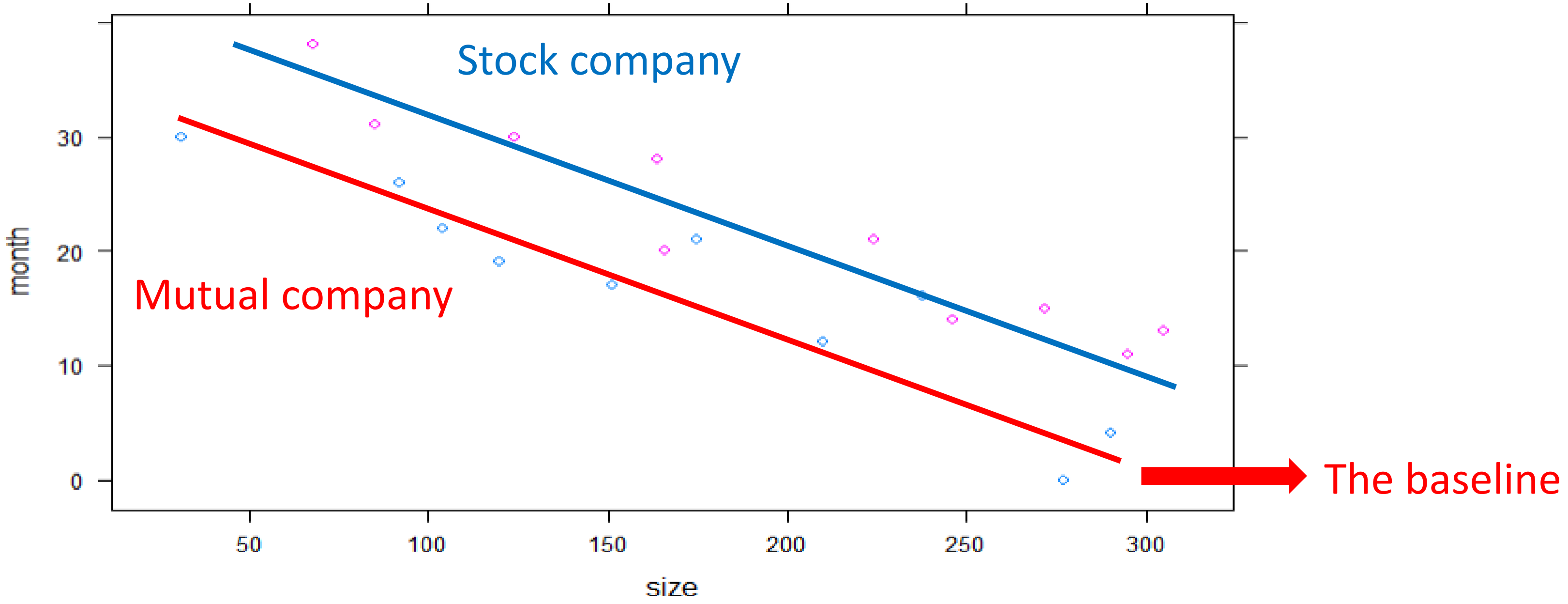
$X_2 = 0, X3 = 0\ (The$ "elsewise") This is the baseline.
$X_2 = 1, X3 = 0\ (The\ mutual\ company)$
$X_2 = 0, X3 = 1\ (The\ Stock\ company)$

The Scatter Plot of the MLR with Qualitative Predictors

$$X_1 = size\ of\ firm$$

$$X_2 = \begin{cases} 0 & if\ mutual\ company \\ 1 & if\ stock\ company \end{cases}$$

Month (Y)	Size (X1)	Type (X2)
17	151	0
26	92	0
28	164	1
11	295	1



The response function: E(Y)= $\beta_0 + \beta_1 X_1 + \beta_2 X_2$

When $X_2 = 0$ (mutual company), the model becomes the baseline function

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_2 (0) + \varepsilon \\
&= \beta_0 + \beta_1 X_1 + \varepsilon \qquad (1)
\end{aligned}
$$

When $X_2 = 1$ (stock company)

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_2 (1) + \varepsilon \\
&= \beta_0 + \beta_1 X_1 + \beta_2 + \varepsilon \\
&= (\beta_0 + \beta_2) + \beta_1 X_1 + \varepsilon \qquad (2)
\end{aligned}
$$

- The response function for the baseline (1) and the next category (2) exhibit an equivalent slope denoted by β_1. This implies that the adoption time(Y) changes uniformly with a change in the company size (X1).
- The difference in intercepts, β_2, reveals the duration difference in adopting a new technology between a stock company (x2=1) and a mutual company (x2=0), considering any given firm size (X1). If $\beta_2 < 0$, it indicates a shorter adoption time for stock companies than mutual companies.
- The effect of company size (x1) on Y is similar for both mutual and stock companies (x2). This characteristic is commonly referred to as the absence of an **interaction effect** between x1 and x2 on Y.
- In the absence of an interaction effect, the distinction in the mean adoption time (Y) between the two types of companies for any specific X1 value is denoted as the ***main effect***, β_2 .

Estimate the Coefficients for the MLR

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.874069   1.813858  18.675 9.15e-13 ***
size        -0.101742   0.008891 -11.443 2.07e-09 ***
type         8.055469   1.459106   5.521 3.74e-05 ***
```

The 95% CI for β_1 is, $\mathrm{b}_1 \pm ts\{b_1\} = -0.102 \pm 2.11(0.00889) = -0.102 \pm 0.0188 = (-0.12, \quad -0.08)$

The 95% CI for β_2 is, $\mathrm{b}_2 \pm ts\{b_2\} = 8.06 \pm 2.11(1.46) = 8.06 \pm 3.08 = (5, \quad 11)$

With 95% confidence level, we conclude that

- For both types of companies, the average adoption time decreases by at least 0.08 and at most 0.12 when the company size increases by 1 unit.
- Additionally, at any given level of company size, we observe that stock companies tend to adopt the innovation at least 5 months and at most 11 months later than mutual companies.

Adding the Interaction Term, X_1X_2

$lm(Y \sim X_1 + X_2 + X_1 * X_2)$

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 33.8383695  2.4406498  13.864 2.47e-10 ***
size        -0.1015306  0.0130525  -7.779 7.97e-07 ***
type         8.1312501  3.6540517   2.225   0.0408 *  
size:type   -0.0004171  0.0183312  -0.023   0.9821    
```

```
Analysis of Variance Table

Response: month
          Df  Sum Sq Mean Sq  F value    Pr(>F)    
size       1 1188.17 1188.17 107.7819 1.627e-08 ***
type       1  316.25  316.25  28.6875 6.430e-05 ***
size:type  1    0.01    0.01   0.0005    0.9821    
Residuals 16  176.38   11.02                       
```

$Ho: \beta_3 = 0, Ha: \beta_3 \neq 0$

$$t_s = \frac{b_3}{s\{b_3\}} = -\frac{0.0004171}{0.01833} = -0.02$$

- Do not reject Ho (p-value =0.9821)
- The interaction is insignificant
- Can also do a GLT test.

Understand the Coefficients in the MLR with a categorical variable

$$\mathrm{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

To comprehend the coefficients, we break down the equation:

$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1$ (Mutual)

$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$ (Stock)

- Firstly, $\beta_0 + \beta_1 X_1$ describes the linear model for the baseline category (i.e., the mutual company), where the linear impact of X1 on Y is β_1 for this baseline.

- Secondly, β_2 describes the main category effect difference between the other category (i.e., the stock company) and the baseline. This main effect difference is associated with the category (X2), not with the other predictor (i.e., X1).

- Lastly, β_3 describes the interaction effect between X1 and X2, which is associated with X1. The linear impact of X1 on Y is $\beta_1 + \beta_3$ in this category.

- To define the linear model for the other category (i.e., the stock company), we can sum up the above three points and write:

$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$$

Understand the Coefficients in the MLR with a categorical variable

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 33.8383695  2.4406498  13.864 2.47e-10 ***
size        -0.1015306  0.0130525  -7.779 7.97e-07 ***
type         8.1312501  3.6540517   2.225   0.0408 *  
size:type   -0.0004171  0.0183312  -0.023   0.9821    
```

$\hat{Y} = b_0 + b_1 X_1 = 33.8 - 0.1X_1$ For mutual firms ($X_2 = 0$)

$\hat{Y} = b_0 + b_2 + (b_1 + b_3)X_1 = (33.8 + 8.1) - (0.1 + 0.0004)X_1$ For stock firms ($X_2 = 1$)

- When X1 increases by 1 unit, the mutual firm experiences a decrease of 0.1 in Y,
while the stock firm experiences a reduction of 0.0004 more than the mutual firm.

- For a given value of X1, the average Y response for the mutual firm is $33.8 - 0.1X_1$, while the stock firm's mean Y response is (33.8+8.1)-(0.1+0.0004) X_1 for the stock firm. The difference is 8.1-0.0004 X_1 = $b_2 + b_3 X_1$

- β_3 , or the interaction effect is insignificant.

Construct the regression model with qualitative predictors **with three (or more) categories**

Example (insurance): in a study of insurance industry, an economist wished to relate the speed with which a particular insurance innovation is adopted (Y) to the size of the insurance firm (X1) and the type of firm (type 1, 2 and 3)

Comment:

- Indicator variables with c classes will be represented by c-1 indicator variables, each taking on the values 0 and 1.
- Two dummy variables, X2 and X3 are required to describe the categorical variable. X2=1 only for type 2, and X3=1 only for type 2. The baseline is type 1 (X2=0, X3=0).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \varepsilon$$

The first category is treated as a base line for other categories to compare to

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \varepsilon$$

$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1$ For type 1
$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_{12} X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) X_1$ For type 2
$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_3 + \beta_{13} X_1 = (\beta_0 + \beta_3) + (\beta_1 + \beta_{13}) X_1$ For type 3

1. $\beta_0 + \beta_1 X_1$ describe the linear model for the baseline category (type1). The linear impact of X1 on Y is β_1 in the baseline.
2. β_2 describes the main category effect difference between the second category (type 2) and the baseline.
3. β_{12} describes the interaction effect and represents the linear impact difference between type 2 and the baseline.

The response function for type 2 finally sums up to: $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) X_1$

Similarly, the linear model for the third category is

$$Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_{13}) X_1$$

Set up hypotheses to compare the two company

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$, where $\beta_3 \neq 0$

$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1$ For mutual firm, $X_2 = 0$

$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$ For stock firm, $X_2 = 1$

Question 1 “the mutual firm and the stock firm have the same average adopt time for any firm size. “

Can be tested by

$Ho: \beta_2 = \beta_3 = 0$

Reduced model $Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$

$Ha: Not\ Ho$

Full model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$

Set up hypotheses to compare the two company

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$, where $\beta_3 \neq 0$

$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1$ For mutual firm, $X_2 = 0$

$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$ For stock firm, $X_2 = 1$

Question 2 “the firm size (X1) has no impact on the adopt time in mutual firm and stock firm.”

Can be tested by

$Ho: \beta_1 = \beta_3 = 0$

Reduced model $Y_i = \beta_0 + \beta_2 X_{i2} + \epsilon_i$

$Ha: Not\ Ho$

Full model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$

Set up hypotheses to compare the two company

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$, where $\beta_3 \neq 0$

$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1$ For mutual firm, $X_2 = 0$

$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$ For stock firm, $X_2 = 1$

Question 3 "the firm size (X1) has the same impact on the adopt time in mutual firm and stock firm."

Can be tested by

$Ho: \beta_3 = 0$

Reduced model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$

$Ha: Not\ Ho$

Full model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$

Set up hypotheses to compare the two company

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$, where $\beta_3 \neq 0$

$$\begin{cases} \mathrm{E}(Y) = \beta_0 + \beta_1 X_1 & \text{For mutual firm, } X_2 = 0 \\ \mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 & \text{For stock firm, } X_2 = 1 \end{cases}$$

Question 4 "Given the firm size (X1) has the same impact on the two companies (i.e., $\beta_3 = 0$) , the average adoption time for the stock firm, at any given firm size, is also the same as the mutual firm."

Can be tested by

$Ho: \beta_2 = 0$

Reduced model $Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$

$Ha: Not\ Ho$

Full model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$

Some considerations in using indicator variables

Many different coding of indicator variables are possible. For example, consider a variable X be the “frequency of product use”

Code1	X
Frequent user	3
Occasional user	2
Nonuser	1

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Mean	E{Y}=$\beta_0 + \beta_1 X_i$
Frequent user	$\beta_0 + 3\beta_1$ (1)
Occasional user	$\beta_0 + 2\beta_1$ (2)
Nonuser	$\beta_0 + 1\beta_1$ (3)

Or

Code2	X
Frequent user	6
Occasional user	3
Nonuser	1

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Mean	E{Y}=$\beta_0 + \beta_1 X_i$
Frequent user	$\beta_0 + 6\beta_1$ (4)
Occasional user	$\beta_0 + 3\beta_1$ (5)
Nonuser	$\beta_0 + 1\beta_1$ (6)

Or

Code3	X1	X2
Frequent user	1	0
Occasional user	0	1
Nonuser	0	0

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

Mean	E{Y}=$\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$
Frequent user	$\beta_0 + \beta_1$ (7)
Occasional user	$\beta_0 + \beta_2$ (8)
Nonuser	β_0 (9)

Note the key implication:

	Allocation code 1	Allocation code 2	Allocation code 3
$E(Y\|frequent\ user) - E(Y\|occasional\ user)$	(1)-(2)=β_1	(4)-(5)=$3\beta_1$	(7)-(8)=$\beta_1 - \beta_2$
$E(Y\|occasional\ user) - E(Y\|non\ user)$	(2)-(3)=β_1	(5)-(6)=$2\beta_1$	(8)-(9)=β_2

Only code 3 makes no assumption about the spacing of the categories.