Topic 17 Explain the Betas in a MLR model with Qualitative X through simulation

Summary

1. Simulate a data set with interaction and main effect
    - Compare the result when fitting the data set with two models
2. Simulate a data set without interaction, only the main effect
    - Compare the result when fitting the data set with two models.
3. Extension: model categorical variable with three levels

The simulated data set

- Y, "grade", is continuous, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$
- X1, "time", is continuous
- X2, "type", is categorical with two levels, type0 and type1.
  - <u>X2=0 for type0</u>, and <u>X2=1 for type1</u>
  - Each type has 1000 observations
  - n=2000.

Assumptions

- When time increases by 1 unit, Y increases by 0.8 for type0, and by 0.3 for type1. I.e., the time and type have an interaction effect on Y.
- Standardized the variable "time" such that
  - The mean of the time is 0 and the standard deviation of time is 1
  - For type 0 and type 1 respectively
- The random error, $\varepsilon \sim Normal(0, 1)$
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$
  - For type0, X2=0, then $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ **(1)**
    - $\mathrm{E}(Y) = E(\beta_0 + \beta_1 X_1 + \varepsilon) = \beta_0 + \beta_1 E(X_1) + E(\varepsilon) = \beta_0$
    - $\sigma(Y) = \sigma(\varepsilon) = 1$
  - For type1, X2=1, then $Y = \beta_0 + \beta_1 X_1 + \beta_2(1) + \beta_3 X_1(1) + \varepsilon = \beta_0 + \beta_2 + (\beta_1 + \beta_3) X_1 + \varepsilon$ **(2)**
    - $\mathrm{E}(Y) = \beta_0 + \beta_2 + (\beta_1 + \beta_3) E(X_1) = \beta_0 + \beta_2$
    - $\sigma(Y) = \sigma(\varepsilon) = 1$
  - This is the main effect of type (X2) on the grade (Y), under average time X1=0
  - Question 1: what is the meaning of β_0 *and* β_2, β_1 *and* β_3?

Recall from previous topic that

- $r = b_1 \frac{S_Y}{S_X}$
- $r = b_1$ when $S_Y = 1 \; and \; S_X = 1$
- Can simulate data with a given linear impact (b_1)by setting linear correlation coefficient (r) of <u>an equal value</u>.
  - $cor(Y, X_1) = b_1 = 0.8$, for type0, and
  - $cor(Y, X_1) = b_1 = 0.3$, for type1

1. Data with both interaction and main effect

Simulating (Y, X1) for X2=type0

```
# simulate (Y, X1) data for X2=type0
qdata_type0 <- data.frame(mvrnorm(n=1000,mu=c(7,0),Sigma=rbind(c(1,.8),c(.8,1)),empirical=TRUE ) )
colnames(qdata_type0)<-c('Grade','Time')
qdata_type0$Type = '0'
```

Simulating (Y, X1) for X2=type1

```
# simulate(Y, X1) data for X2=type1
qdata_type1 <- data.frame(mvrnorm(n=1000,mu=c(9,0),Sigma=rbind(c(1,.3),c(.3,1)),empirical=TRUE ))
colnames(qdata_type1) <- c('Grade','Time')
qdata_type1$Type = '1'
```

Stack them to form the whole data set

```
# Combine data
qdata <-rbind(qdata_type0,qdata_type1)
qdata$Type<-as.factor(qdata$Type)
```

Review the data

```
The linear impact of time on grade, i.e., the correlation between Grade & Time for Type0: 0.8
The linear impact of time on grade, i.e., the correlation between Grade & Time for Type1: 0.3
```

```
Mean Grade for Type0: 7
Mean Grade for Type1: 9
Mean Grade for all types: 8
Mean Time for Type0: -3.455576e-18
Mean Time for Type1: 1.340724e-17
Total sample size n= 2000
```

Interpret the beta (b) on the actual linear regression model

```
Call:
lm(formula = Grade ~ Time + Type + Time * Type, data = qdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7779 -0.4956 -0.0018  0.4897  3.4010 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.00000    0.02521  277.65   <2e-16 ***
Time         0.80000    0.02522   31.71   <2e-16 ***
Type1        2.00000    0.03565   56.09   <2e-16 ***
Time:Type1  -0.50000    0.03567  -14.02   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7973 on 1996 degrees of freedom
Multiple R-squared:  0.6827,	Adjusted R-squared:  0.6822 
F-statistic:  1431 on 3 and 1996 DF,  p-value: < 2.2e-16
```

$$Y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

$\beta_0 = 7, the\ mean\ Y\ on\ the\ type0\ (baseline)$=7

$\beta_1 = 0.8, the\ linear\ impact\ of\ X1\ on\ Y\ for\ type0\ (baseline)$

$\beta_2 = 2, the\ main\ effect\ difference\ between\ type1\ and\ the\ baseline$

$the\ mean\ Y\ on\ the\ type1 = 7 + 2 = 9$

$\beta_3 = -0.5, the\ linear\ impact\ difference$ between type1 and the baseline

$the\ impact\ of\ time\ on\ Y\ on\ the\ type1 = 0.8 - 0.5 = 0.3$

Note: R usually determines the baseline according to the alphabetical Order on the variable

- “Type0” “Type1”, then the default baseline=“Type0”
- “N”, “E”, “S”, “W”, then the default baseline=“E”

When a wrong model is fit on the data

```
Call:
lm(formula = Grade ~ Time + Type, data = qdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9678 -0.5481 -0.0054  0.5514  4.2563 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.00000    0.02642  264.98   <2e-16 ***
Time         0.55000    0.01869   29.43   <2e-16 ***
Type1        2.00000    0.03736   53.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8354 on 1997 degrees of freedom
Multiple R-squared:  0.6514,	Adjusted R-squared:  0.6511 
F-statistic:  1866 on 2 and 1997 DF,  p-value: < 2.2e-16
```

$$Y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$\beta_0 = 7, the\ mean\ Y\ on\ the\ type0\ (baseline) = 7$$

$$\beta_1 = 0.55 = \frac{0.8 + 0.3}{2}, the\ time\ has\ the\ same\ impact\ on\ grade\ .$$

$$\beta_2 = 2, the\ main\ effect\ difference\ on\ type1\ from\ the\ baseline$$

$$the\ mean\ Y\ on\ the\ type1 = 7 + 2 = 9$$

2. Data without the interaction, only main effect

Simulating (Y, X1) for X2=type0

```
qdata_type0 <- data.frame(mvrnorm(n=1000,mu=c(7,0),Sigma=rbind(c(1,.8),c(.8,1)),empirical=TRUE ) )
colnames(qdata_type0)<-c('Grade','Time')
qdata_type0$Type = '0'
```

Simulating (Y, X1) for X2=type1

```
qdata_type1 <- data.frame(mvrnorm(n=1000,mu=c(9,0),Sigma=rbind(c(1,.8),c(.8,1)),empirical=TRUE ))
colnames(qdata_type1) <- c('Grade','Time')
qdata_type1$Type = '1'
```

Stack them to form the whole data set

Interpret the beta (b) on the actual linear regression model

```
Call:
lm(formula = Grade ~ Time + Type, data = qdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.05381 -0.39636  0.00376  0.42229  2.20170 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.00000    0.01898  368.84   <2e-16
Time         0.80000    0.01343   59.58   <2e-16
Type1        2.00000    0.02684   74.52   <2e-16
```

$Y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2$

$\beta_0 = 7, the\ mean\ Y\ on\ the\ type0\ (baseline)$=7

$\beta_1 = 0.8, the\ linear\ impact\ of\ X1\ on\ Y\ for\ type0\ (baseline)$

$\beta_2 = 2, the\ main\ effect\ difference\ on\ type1\ from\ the\ baseline$

$the\ mean\ Y\ on\ the\ type1 = 7 + 2 = 9$

When a wrong model is fit on the data

```
Call:
lm(formula = Grade ~ Time + Type + Time * Type, data = qdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.05381 -0.39636  0.00376  0.42229  2.20170 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.000e+00  1.898e-02  368.75   <2e-16 ***
Time         8.000e-01  1.899e-02   42.12   <2e-16 ***
Type1        2.000e+00  2.685e-02   74.50   <2e-16 ***
Time:Type1  -9.389e-16  2.686e-02    0.00        1    
```

$Y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$

$\beta_0 = 7, the\ mean\ Y\ on\ the\ type0\ (baseline)$=7

$\beta_1 = 0.8\ , the\ linear\ impact\ of\ time\ on\ grade\ for\ type0$

$\beta_2 = 2, the\ main\ effect\ difference\ on\ type1\ from\ the\ baseline$

$the\ mean\ Y\ on\ the\ type1 = 7 + 2 = 9$

$\beta_3 = 0, the\ linera\ impact\ of\ time\ on\ grade\ for\ type1$
$has\ no\ difference\ from\ type0$, or **the interaction effect is insignificant**

Note: it looks okay if the model include more terms than actual, because the extra term could be proved to be insignificant in the data. But if there is any assumption violation, this might not be true.

Define hypotheses based on

$$Y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

1. The linear impact of X1 on Y is the same for both types.

   Ho: Ha:

2. The mean Y for type0 is the same as type1, given X1=0.

   Ho: Ha:

3. X1 has no impact on Y for both types.

   Ho: Ha:

Define hypotheses based on

$$Y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

4. The interaction effect between X1 and X2 on Y is insignificant.

Ho: Ha:

5. Given an insignificant interaction effect, the mean Y for type0 is the same as type1

The full model should be $Y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2$, and
The reduced model should be $Y \sim \beta_0 + \beta_1 X_1$

Ho: Ha:

Extension

- Y, “grade”, is continuous,
- X1, “time”, is continuous
- X2 and X3 are dummy variables for the categorical variable, “type” with three levels, i.e., c=3 (type0, type1, and type2)
  - X2=0 and X3=0 for type0 (baseline)
  - X2=1 and X3=0 for type1
  - X2=0 and X3=1 for type2
  - Each type has 1000 observations, n=3000.

Assumptions

- When time increases by 1 unit, Y increases by 0.8 for type0, 0.3 for type1 and 0.9 for type2. I.e., the time and type have an interaction effect on Y.
- Standardized the variable "time" such that
  - The mean of the time is 0 and the standard deviation of time is 1
  - For type 0 and type 1 respectively
- The random error, $\varepsilon \sim Normal(0, 1)$
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \varepsilon$
  - For type0, X2=0 and X3=0 , then $Y = \beta_0 + \beta_1 X_1 + \varepsilon$,
  - For type1, X2=1 and X3=0 , then $Y = \beta_0 + \beta_1 X_1 + \beta_2(1) + \beta_4 X_1(1) + \varepsilon$
    $= \beta_0 + \beta_2 + (\beta_1 + \beta_4) X_1 + \varepsilon$
  - For type2, X2=0 and X3=1 , then $Y = \beta_0 + \beta_1 X_1 + \beta_3(1) + \beta_5 X_1(1) + \varepsilon$
    $= \beta_0 + \beta_3 + (\beta_1 + \beta_5) X_1 + \varepsilon$
  - Question 2: what is the meaning of β_0 $\beta_2, \beta_{3,}$ β_4 *and* β_5?
  - Question 3: predict the values of the coefficients and verify with R