# APSTA-GE 2003: Intermediate Quantitative Methods

Office Hour Notes

Tong Jin

**Created on: 10/04/2020**

**Modified on: 10/05/2020**

## Topics

- Answer questions in **Assignment 5** and clarify concepts

**Note:** Don't worry about the duplicate choices in Question 1. We will manually grade and adjust this.

## Question 1

Additive means that each independent variables are treated as independent with each other. We assume that they are not correlated.

Interactive means that we assume that there exists correlation among independent variables. Therefore, for independent variables with interactive effects, the predicted values of one variable will change as the values of the other change.

### Example

Let's examine interactive effects using the `marketing` dataset in the `datarium` package.

We want to examine if investing in online marketing leads to sale increases. Additionally, we would like to see if youtube budget and facebook budget are correlated.

```
dat_sim <- datarium::marketing
str(dat_sim)
```

```
## 'data.frame':    200 obs. of  4 variables:
##  $ youtube  : num  276.1 53.4 20.6 181.8 217 ...
##  $ facebook : num  45.4 47.2 55.1 49.6 13 ...
##  $ newspaper: num  83 54.1 83.2 70.2 70.1 ...
##  $ sales    : num  26.5 12.5 11.2 22.2 15.5 ...
```

```
# Create a dummy variable that marks those higher than 170 youtube sales as 1.
dat_sim$youtube_170 <- ifelse(dat_sim$youtube >= 170, yes = 1, no = 0)

lm_sim <- lm(sales ~ facebook * youtube_170, data = dat_sim)
summary(lm_sim)
```

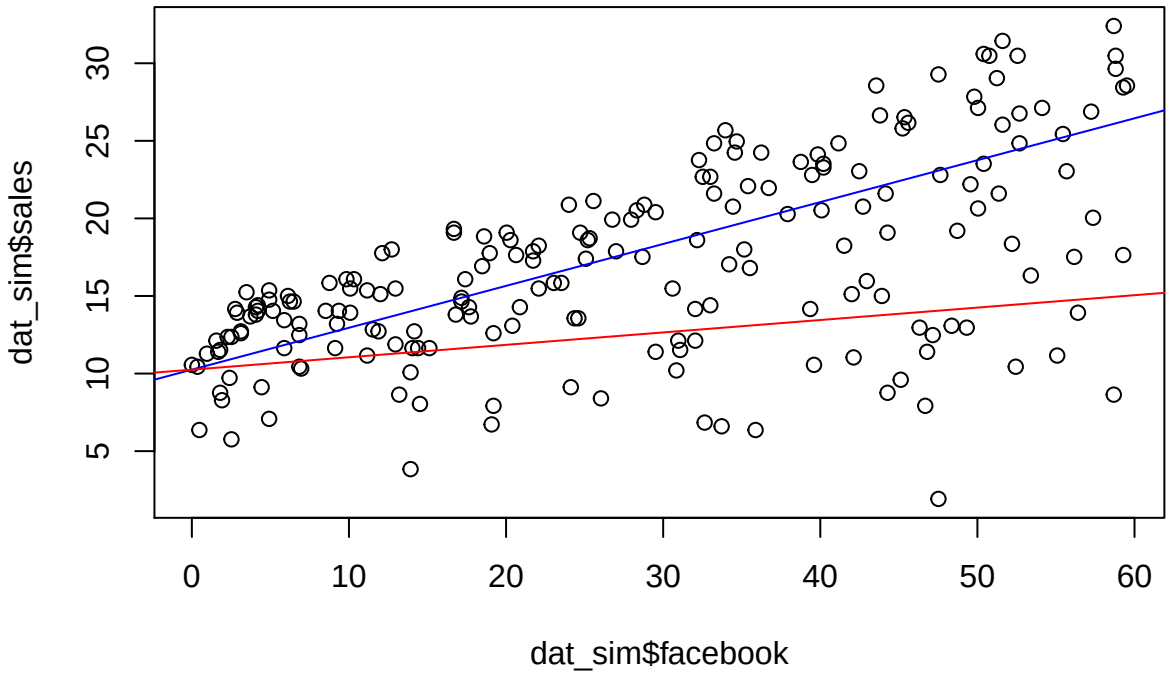
```
##
## Call:
## lm(formula = sales ~ facebook * youtube_170, data = dat_sim)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -12.3292  -1.2557   0.1728   1.6300   8.1034
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          10.24590    0.50482  20.296  < 2e-16 ***
## facebook              0.08424    0.01590   5.298 3.13e-07 ***
## youtube_170           2.49415    0.73916   3.374 0.000892 ***
## facebook:youtube_170  0.19386    0.02237   8.666 1.67e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.799 on 196 degrees of freedom
## Multiple R-squared:  0.8031, Adjusted R-squared:  0.8001
## F-statistic: 266.5 on 3 and 196 DF,  p-value: < 2.2e-16
```

From the model result, we can write down the model equation:

$$\begin{aligned} sales &= 10.25 + 0.08 \cdot facebook + 2.49 \cdot youtube_{170} + 0.19(facebook \cdot youtube_{170}) \\ &= 10.25 + (0.08 + 0.19 \cdot youtube_{170}) \cdot facebook + 2.49 \cdot youtube_{170} \\ &= 10.25 + (2.49 + 0.19 \cdot facebook) \cdot youtube_{170} + 0.08 \cdot facebook \end{aligned}$$

```
# Visualize
plot(dat_sim$facebook, dat_sim$sales)
abline(a = 10.25, b = 0.08 + 0.19, col = "blue") # youtube_170 == 1
abline(a = 10.25, b = 0.08, col = "red")         # youtube_170 == 0
```




Interpretation:

the regression coefficient of `facebook` is $0.08+0.19 \cdot youtube_{170}$. It depends on the status of `youtube_170` (whether or not youtube budget is higher than 170). Because `youtube_170` is binary (0 or 1), the regression coefficient of `facebook` is different for different `youtube_170` status.

Because for both `youtube_170` status, the slope (regression coefficient of `facebook`) is positive, we can say that, on average, the higher the facebook budget, the higher the sales.

For higher facebook budget, the distance between two lines increases. This indicates that the influence of heavily investing in youtube marketing will increase as the facebook budget increases.

**Question 9**

Through stratification using a binary variable, we create two separte dataset and fit two models.

- sample size: the interaction model includes all sample points while the stratified models use separated sample points ($n_{s1} < n, n_{s2} < n$).
- sample size affects variance (error). Genearlly, the larger the sample size, the smaller the variance.
- because the estimation of regression model is based on the data sample, given an equation, different samples generate different estimations. Therefore, the ability of explaining data variance varies.
- Stratification applies an additional centerization to the data because it tries to estimate using two models instead of one.