Advanced Course in Marketing 10 Analytical techniques in regression

Takumi Tagashira

Email: takumi.tagashira@r.hit-u.ac.jp

Review

- Overview of regression:
  - Estimation, tests, confidence interval
- Multiple regression model
  - Interpretation
  - Control variables

Analytical techniques on regression

- We will cover the following analytical techniques that are frequently used in the marketing context.
  - Dummy variable: using categorical independent variables
  - Interaction term: An effect of a variable depend on another variable
  - Log-linear model: Capturing the elasticity (skip)

Categorical variables for regression

- We are interested in the effects of categorical variables
  - E.g. gender, business sector.
- The most popular way to handle categorial variables is generating “dummy variables”.
  - Variables that take value of 1 if a sample belongs to the focal group, otherwise, 0.

Dummy variables

$$y_i = \alpha + \beta x_i + \gamma D_i + u_i$$

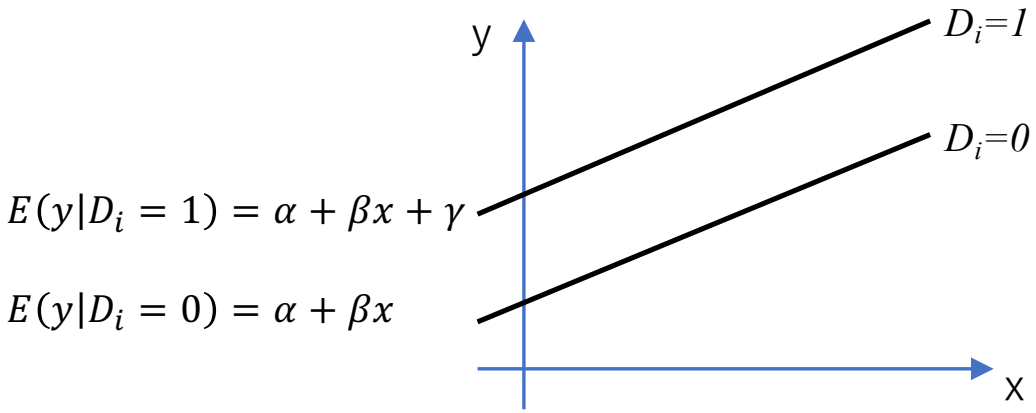
- D represents categorical variables (Dummy variables) can be employed as independent variables.
- e.g., Gender (Male dummy)
  - 1 if individual i is a male, 0 if i is a female.
- Interpretation of the result
  - $E(y|D_i = 1) = \alpha + \beta x + \gamma$
  - $E(y|D_i = 0) = \alpha + \beta x$
- The effects of dummy variable show the shift of intercept.

Dummy variables

$$y_i = \alpha + \beta x_i + \gamma D_i + u_i$$

- e.g., Suppose the slope is positive, and male has higher value for the outcome than females.

- Interpretation:
  - When D = 1 group has a higher value of y than D = 0 group on average, intercept shift represents the relative difference of y between groups (e.g. mean comparison)



Generating a categorical variable

- We use firmdata19.
- We aim to check whether the operating profit is different between retail and other industries.
  - We generate a categorical variable that states “Retail” if a sample belongs to “Retail Stores, NEC” or “Supermarket Chains”

```
library(tidyverse)
firmdata <- readxl::read_xlsx("data/MktRes_firmdata.xlsx")
retail <- c("Retail Stores, NEC", "Supermarket Chains")

firmdata19 <- firmdata %>%
  filter(fyear == 2019) %>%
   mutate(format = ifelse(ind_en %in% retail, "Retail", "Other"))

#Checking the sample frequency
with(firmdata19, table(format))
```

Regression with a categorical variable

- Just add the categorical variable in the model of lm() function.
- R automatically generate dummy variable based on used categorical variable (i.e., format).
  - Others group was set as the baseline this time by chance.

```
fit.d1 <- lm(op ~ mkexp + format, data = firmdata19)
summary(fit.d1)
```

- If you want to specify a particular group as 1, you should generate a dummy variable.

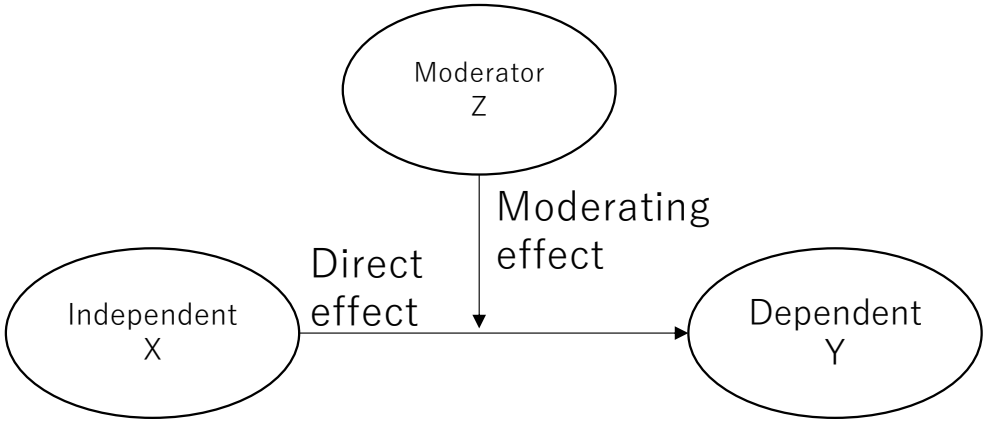
Generating a dummy variable and regression analysis

- Let's generate a dummy variable that takes 1 for retailers.
- Check the cross-tabulation table, then run a regression.
- Results?

```
firmdata19 <- firmdata19 %>%
  mutate(retail = ifelse(format == "Retail", 1, 0))
#Cross-tabulation table
with(firmdata19, table(retail, format))
#Regression
fit.d2 <- lm(op ~  mkexp + retail, data = firmdata19)
summary(fit.d2)
```

Interaction term

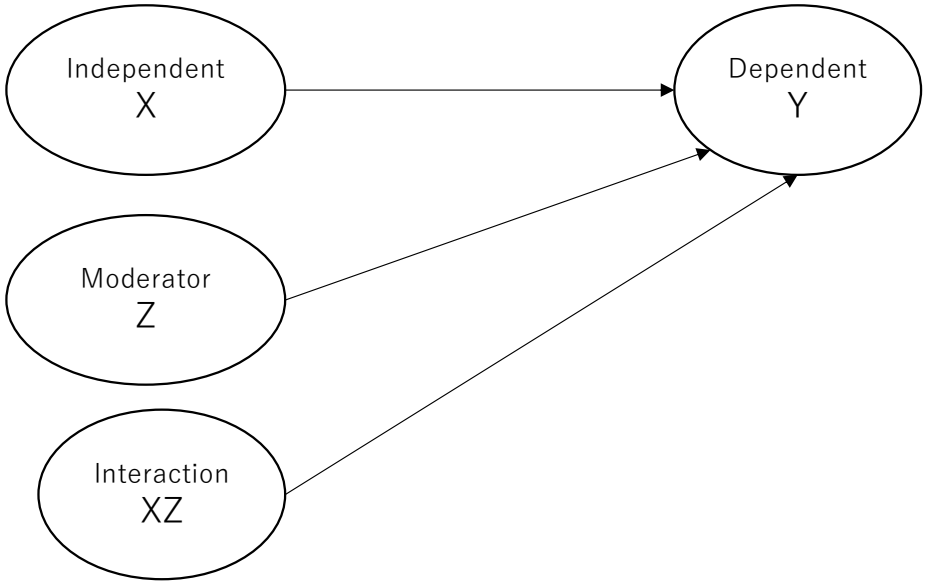
Moderating effect



- We sometimes are interested in the effect of an independent variable that depends on another variable.
- Marketing studies employ this type of hypothesis
  - How can we verify this relationship?

Interaction term

- We employ an interaction term (X*Z) to assess the moderating effect.



Things to consider

1. Interaction terms assess the conditional hypotheses.
   - E.g. When the value of Z changes, the effect of X on Y also changes.
2. A model with an interaction term should include the independent term of two variables used for the interaction.
3. The interpretation of coefficients are different from the ordinary regression models.

Slope dummy

- An interaction term with continuous variable and dummy variable indicates that the effect varies depending on groups.
  - i.e., slope dummy

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 (X \times D) + u$$

- *For D=1*:
  - $\frac{\Delta Y}{\Delta X} = \beta_1 + \beta_3$
- *For D=0*:
  - $\frac{\Delta Y}{\Delta X} = \beta_1$
- Thus, β_3 captures the difference in the slope of the linear model.

Intercept dummy in R

- Is the effect of marketing expense on profitability differs between retail and other sectors?
- The effect of **`mkexp:retail`** is positive, but the independent terms are negative.

```
fit.d3 <- lm(op ~  mkexp * retail, data = firmdata19)
summary(fit.d3)
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.108430   0.009217  11.764  < 2e-16 ***
## mkexp        -0.105662   0.027505  -3.842 0.000183 ***
## retail       -0.088147   0.023231  -3.794 0.000218 ***
## mkexp:retail  0.191432   0.063025   3.037 0.002837 **
```

Summary of the results

- View results more deeply.
- The predicted value $\hat{y}$ is shown as follows:

$$\hat{y}_i = 0.108430 - 0.105662 mkexpt_i - 0.088147 retail_i + 0.191432 mkexp_i * retail_i$$

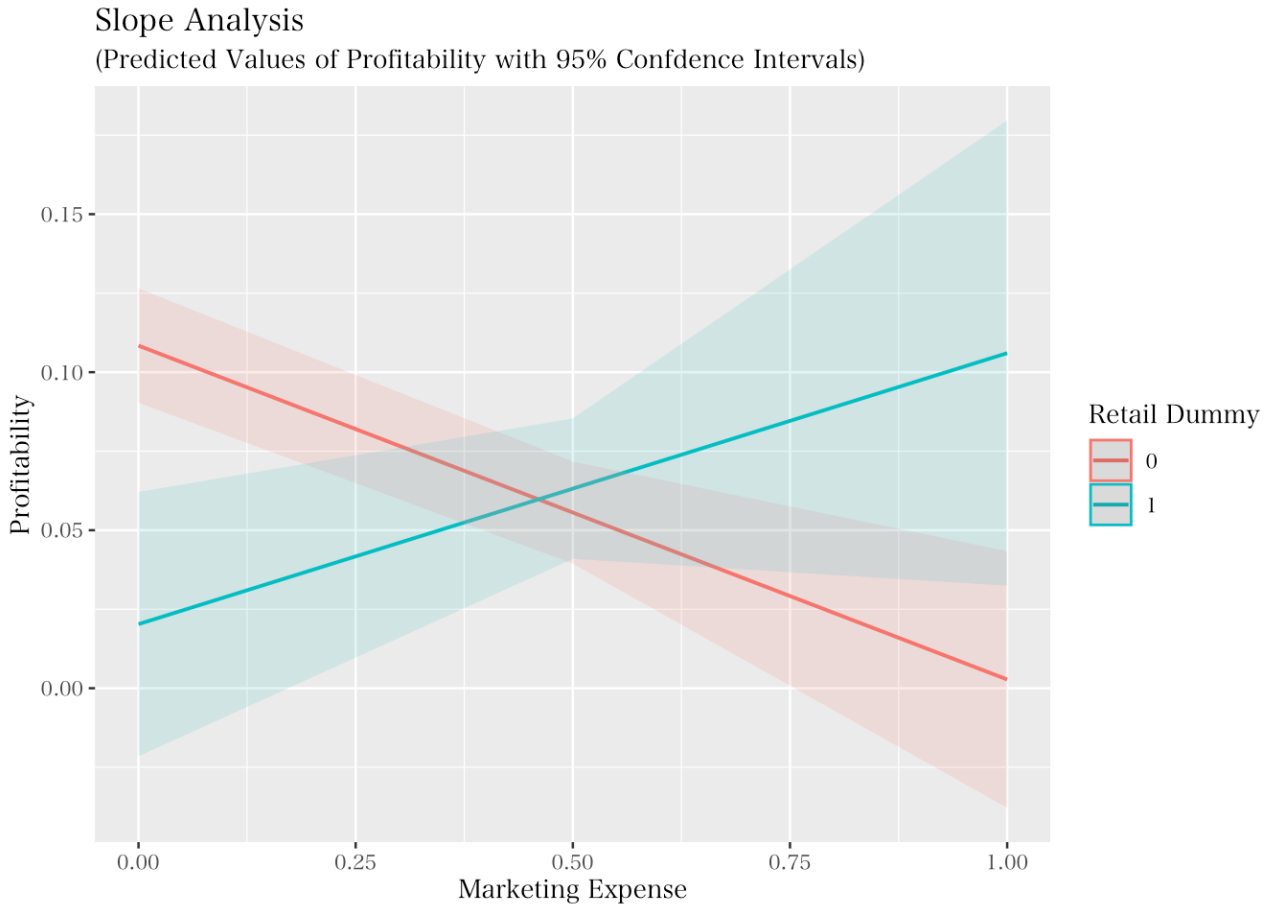
$$\begin{cases} Retail: \hat{y}_i = 0.020283 + 0.08577 mkexp_i \\ Others: \hat{y}_i = 0.108430 - 0.105662 mkexp_i \end{cases}$$

Visualization

- The results of slope dummy should be visualized.
- We use a package called “sjPlot”.
  - Please install it by “install.packages("sjPlot")”
  - Visualization can be executed by the following commands:

```
library(sjPlot)
pred <- plot_model(fit.d3, type = "pred", terms = c("mkexp","retail"),
ci.lvl = .95) +
  labs(title = "Slope Analysis",
       subtitle = "(Predicted Values of Profitability with 95% Confdence
Intervals)",
       x = "Marketing Expense", y = "Profitability") +
  scale_color_discrete(name = "Retail Dummy")
pred
```

- The figure shows that the positive effect of marketing expense for retailers.
  - But the effect seems negative for firms in other industries.



Interaction term with continuous variables

- We can generate interaction terms with continuous variables.

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 \underbrace{(X \times Z)}_{\text{Interaction}} + u$$

- However, we should be careful with the interpretation.

Interaction with continuous variables

- Interaction term can be applied to two continuous variables.

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 \underbrace{(X \times Z)}_{\text{Interaction}} + u$$

- Let $\frac{\Delta Y}{\Delta X}$ denote the effect of X on Y:

$$\frac{\Delta Y}{\Delta X} = \beta_1 + \beta_3 Z$$

- Let $\frac{\Delta Y}{\Delta Z}$ denote the effect of Z on Y:

$$\frac{\Delta Y}{\Delta Z} = \beta_2 + \beta_3 X$$

- The effects of X and Z on Y are interdependent each other.
- β_3 is known as the moderating effect.

Interpretation of the interaction model

- $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 (X \times Z) + u$

- The effect of X on Y:

$$\frac{\Delta Y}{\Delta X} = \beta_1 + \beta_3 Z$$

- The effect of Z on Y:

$$\frac{\Delta Y}{\Delta Z} = \beta_2 + \beta_3 X$$

- β_1 represents the effect of X on Y in what condition?

Interpretation of the interaction model II

- Suppose the Z in the previous model is personal height.
  - $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 (X \times Z) + u$

$$\frac{\Delta Y}{\Delta X} = \beta_1 + \beta_3 Z$$

- β_1 represents the effect of X on Y when the height is 0.

- Does it make sense?
- Is there any solution?

Mean centering

- A possible solution is **mean centering**:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 (X - \bar{X})(Z - \bar{Z}) + u$$

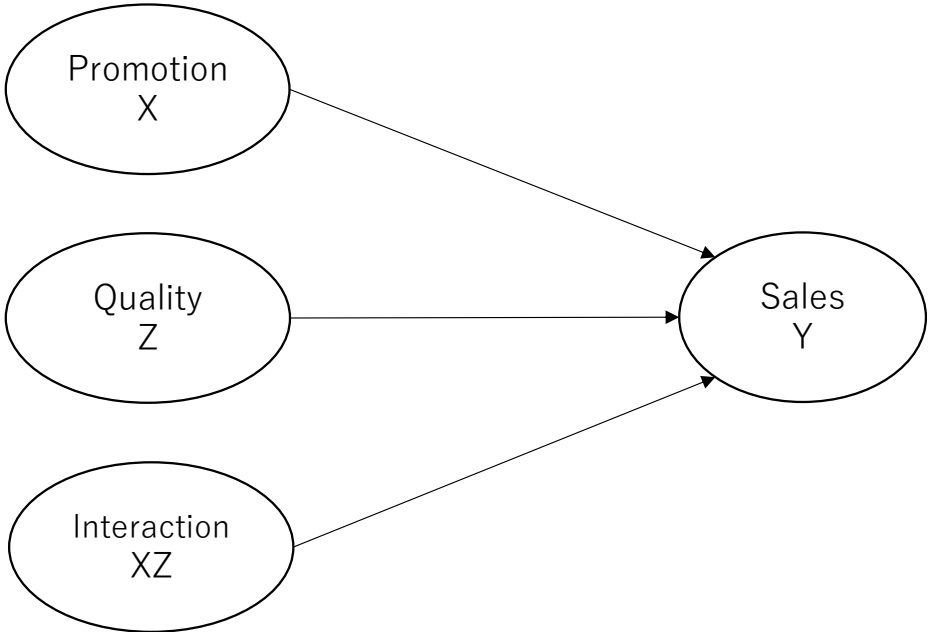
- The results can be considered as follows.

$$\frac{\Delta Y}{\Delta X} = \beta_1 + \beta_3 (Z - \bar{Z}); \ \frac{\Delta Y}{\Delta Z} = \beta_2 + \beta_3 (X - \bar{X})$$

- β_1 represents the effect of X on Y when Z equals mean.

Interaction model: Example

- Does the promotion cost moderate the effect of investment in product quality (R&D) on sales?
  - i.e., even if a firm develops a quality product, it should deliver information to customers.



Fictional data

- Headphone07 includes information about the annual sales of headphone brands in a particular year.
- The dataset includes the following variables:
  - Sales (million yen)
  - Promotion cost (million yen)
  - R&D investment (million yen)

- Suppose other things are equal.

Data preparation

- Importing “Headphone07.csv”

```
Headphone07 <- readr::read_csv("data/Headphone07.csv", na = ".")
```

- Create mean-centered variables

```
glimpse(Headphone07)

Headphone07 <- Headphone07 %>%
  mutate(promotion_c = promotion - mean(promotion, na.rm = TRUE),
         rd_c = rd - mean(rd, na.rm = TRUE))
```

Interaction without centering

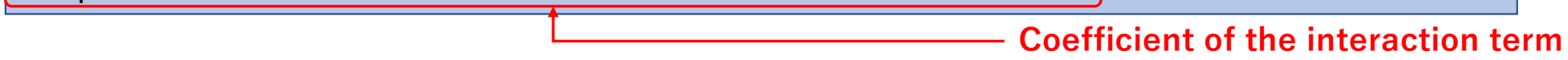
- Interaction and independent terms are automatically added with “var1*var2” specification in a reg model.

```
fit_int <- lm(sales ~ rd*promotion, data = Headphone07)
summary(fit_int)
```

Interaction without centering

- Interaction and independent terms are automatically added with “var1*var2” specification in a reg model.

```
fit_int <- lm(sales ~ rd*promotion, data = Headphone07)
summary(fit_int)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.033e+04  1.090e+02   186.5   <2e-16 ***
rd           -5.187e+01  2.844e-01  -182.4   <2e-16 ***
promotion    -1.914e+02  1.052e+00  -181.9   <2e-16 ***
rd:promotion  4.979e-01  2.730e-03   182.4   <2e-16 ***
```



- Both R&D and Promotion have significant negative effects.

With centering

- Using the mean centered variables.

```
fit_int_c <- lm(sales ~ rd_c*promotion_c, data = Headphone07)
summary(fit_int_c)
```

With centering

- Using the mean centered variables.

```
fit_int_c <- lm(sales ~ rd_c*promotion_c, data = Headphone07)
summary(fit_int_c)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      382.27812    1.51459  252.40   <2e-16 ***
rd_c              -0.91767    0.01196  -76.75   <2e-16 ***
promotion_c        0.69039    0.03972   17.38   <2e-16 ***
rd_c:promotion_c   0.49792    0.00273  182.37   <2e-16 ***
```

- Coefficient of R&D is negative.
- Promotion has a positive effect.
  - When the product quality is average, promotion increase sales.
  - When the promotion is average, R&D leads to lower sales.
  - R&D and price？ -> Further study is required.
- Variable transformation critically affect the results.

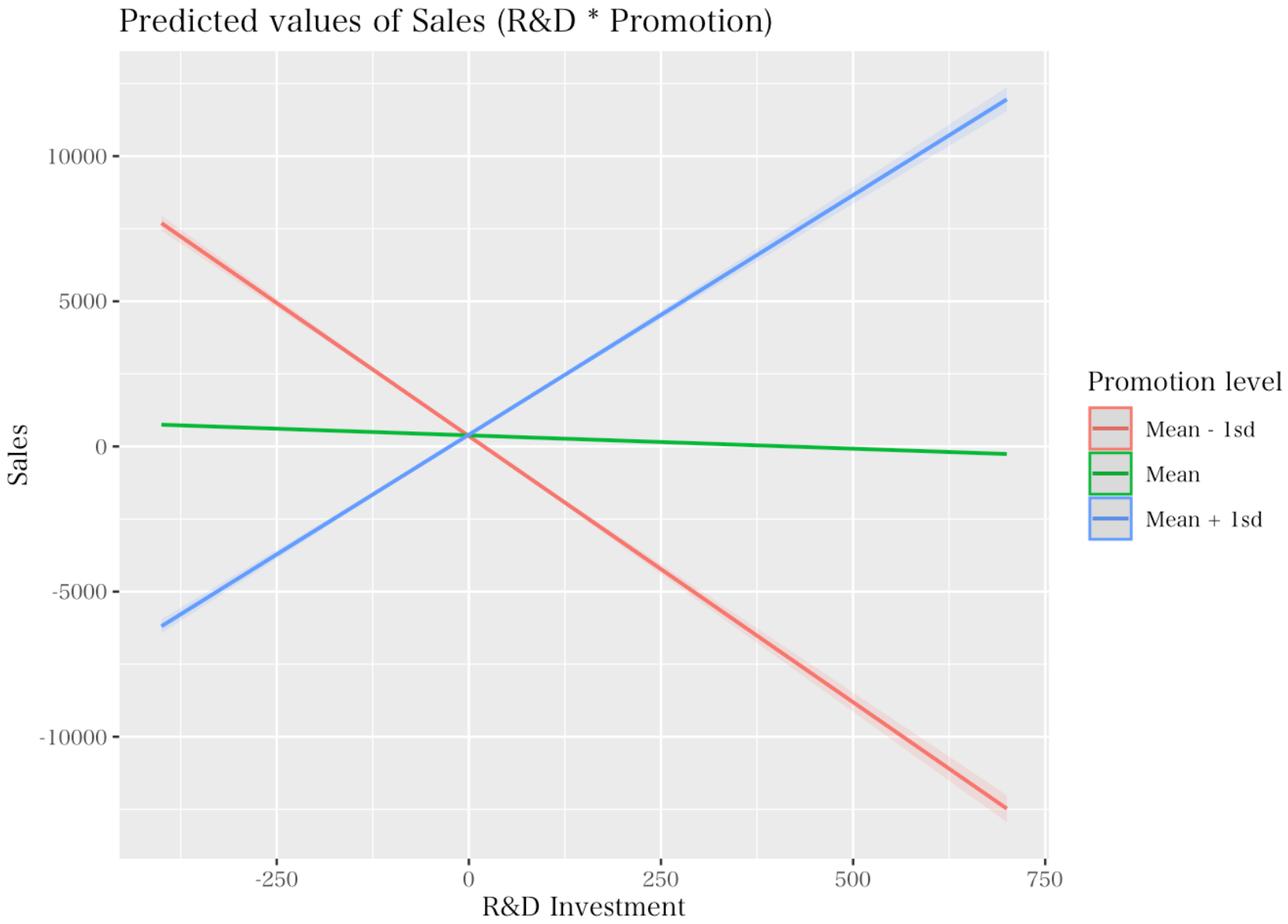
Visualization

- There are mainly two approaches to visualize interaction effects:
  1. Dividing the level of moderating variable into high (e.g., mean + 1 std. dev.) and low (e.g., mean - 1 std. dev.) and showing lines like those in the slope dummy example.
  2. Showing the continuous variation of effects of the main independent variable according to the moderating variable.

Visualization 1

- First approach can be done with the following commands
  - “leg” specifies the legend of the figure.

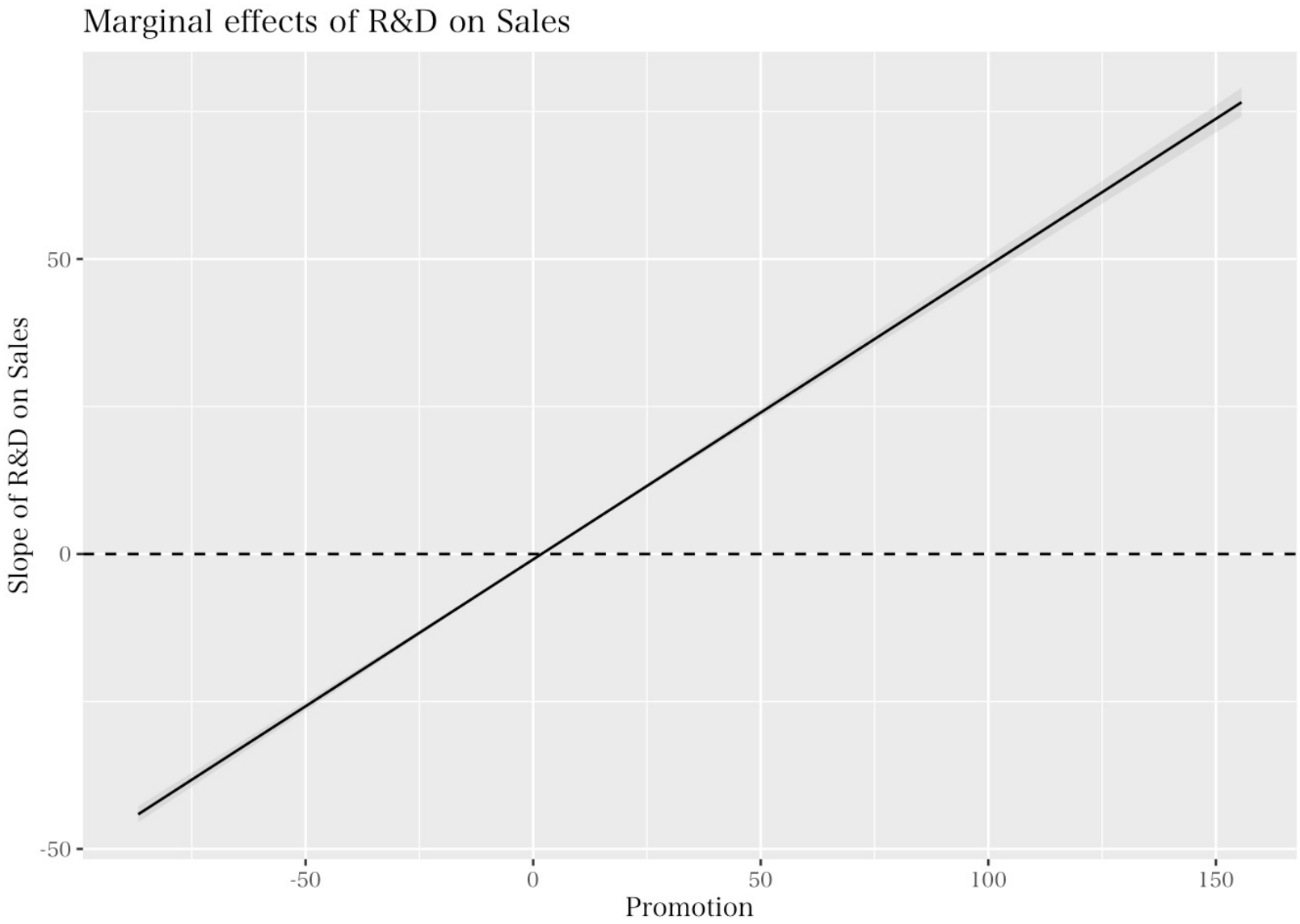
```
leg = c("Mean - 1sd", "Mean", "Mean + 1sd")

int_fig1 <- plot_model(fit_int_c, type = "int", mdrt.values = "meansd",
ci.lvl = .9999999999) +
  labs(title = "Predicted values of Sales (R&D * Promotion)",
       x = "R&D Investment", y = "Sales")+
  scale_color_discrete(name = "Promotion level", labels=leg)

int_fig1
```


Predicted values of Sales (R&D * Promotion)
Sales
10000
5000
0
-5000
-10000
-250
0
250
500
750
R&D Investment
Promotion level
Mean - 1sd
Mean
Mean + 1sd

Visualization 2

- Install a package called “marginaleffects”
  - We use “plot_slopes()” function in “marginaleffects”.
  - Details: https://vincentarelbundock.github.io/marginaleffects/dev/

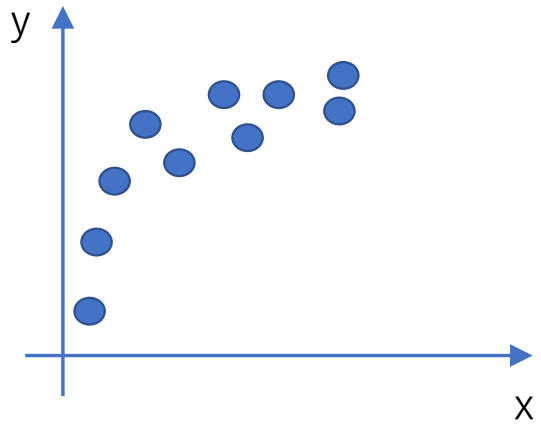
```
install.packages("marginaleffects")
library(marginaleffects)

int_fig2 <- plot_slopes(fit_int_c, variables = "rd_c", condition =
"promotion_c", conf_level = .99999999) +
  labs(title = "Marginal effects of R&D on Sales",
       x = "Promotion", y = "Slope of R&D on Sales") +
geom_hline(aes(yintercept=0), linetype = "dashed")
int_fig2
```


Marginal effects of R&D on Sales
Slope of R&D on Sales
50
0
-50
-50
0
50
100
150
Promotion

Log-linear model

Log-linear model

- Linear regression models do not fit well with non-linear relationships
  - e.g., the slope changes at a certain point.



Regression and linearity

- For example, a specification of the previous non-linear relationship is as follows.

$$Y = \alpha X^{\beta}$$

  - E.g., Cobb-Douglas production function ($Y = AL^{\alpha}K^{\beta}$)

- Regression models should be linear regarding parameters.

Log-linear model

- We can develop the following log-natural regression model.

$$\ln y_i = \beta_0 + \beta_1 \ln x_{1i} + \cdots + u_i$$

- The model keeps the linearity regarding the parameters.
  - The interpretation of parameters will be different from ordinary linear regression.

Long-transformation and interpretation

- Dependent and independent: $\ln y = \beta_0 + \beta_1 \ln x + u$
  - β_1 shows the percentile change of y when x changes 1%.

- Dependent: $\ln y = \beta_0 + \beta_1 x + u$
  - β_1 shows the percentile change of y when x changes 1 unit.

- Independent: $\mathrm{y} = \beta_0 + \beta_1 \ln x + u$
  - β_1 shows the change of y when x changes 1%.

Taking logarithm in R

- We can convert variables into logarithm by "log()" function in "lm()"
- Let's estimate the following model with "firmdata19".
  - Y: sales, L: labor cost, K: fixed assets
  - i.e., Cobb-Douglas production function

$$\ln Y_i = \ln A + \alpha \ln K_i + \beta \ln L_i + u_i$$

```
fit_prod <- lm(log(sales) ~ log(labor_cost) + log(ppent), data = firmdata19)

summary(fit_prod)
```

Log-linear model

- We can develop the following log-natural regression model.

$$\ln y_i = \beta_0 + \beta_1 \ln x_{1i} + \cdots + u_i$$

- The model keeps the linearity regarding the parameters.
  - The interpretation of parameters will be different from ordinary linear regression.
- This form is widely used in marketing.
- Why?

Elasticity

- The parameters imply elasticities.
  - i.e., Percentile change of y when x increase 1%.
- $\ln y = \beta_0 + \beta_1 \ln x + u$

$$\frac{\partial \ln y}{\partial \ln x} = \frac{\partial y}{\partial x} \times \frac{x}{y}$$

- Assume a demand function $q = f(p)$ and let η denote the price elasticity of demand.

$$\eta = \frac{dq}{dp} \times \frac{p}{q}$$

- (See handouts for details.)

Price elasticity of demand

- When other things equal, price is negatively associated with demand quantity.
- But how much does the price impact on demand?
  - Price sensitivity
- The unit of the product changes the slope.
  - e.g. Litre vs ml
- Measurement of price sensitivity should not depend on the unit of the products.
  - Price elasticity of demand

Price elasticity of demand and implication

- If the value (η) is greater than one is an important criteria.
- When $\eta = 1$:
  - It means that quantity increases 1% when the price reduces 1%. Price rate of change and demand rate of change are balanced.
- When $\eta < 1$:
  - When the elasticity is smaller than one, benefits from increased price (i.e. increased margin) surpasses the loss of demand quantity, and thus it is beneficial for firms to increase the price.
  - Increased price increases firm's revenue.
  - Benefits of increased price > Loss by increased price
- When $\eta > 1$:
  - Increased price reduces firm's revenue.
  - Benefits of increased price < Loss by increased price

Long-transformation and interpretation

- Dependent and independent: $\ln y = \beta_0 + \beta_1 \ln x + u$
  - β_1 shows the percentile change of y when x changes 1%.

- Dependent: $\ln y = \beta_0 + \beta_1 x + u$
  - β_1 shows the percentile change of y when x changes 1 unit.

- Independent: $\mathrm{y} = \beta_0 + \beta_1 \ln x + u$
  - β_1 shows the change of y when x changes 1%.

(Sup) Estimation of log-linear model I

- We demonstrate the log-linear model with a different dataset.
  - Download ‘price_data.csv’ via manaba.

```
price<- readr::read_csv("data/price_data.csv", na = ".")
#linear model
reg1 <-lm(q~p, data = price)
summary(reg1)

#log-linear model
reg2 <-lm(log(q)~log(p), data = price)
summary(reg2)
```

Results

- Log-linear results

```
lm(formula = log(q) ~ log(p), data = price)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.47781    0.08135   116.5   <2e-16 ***
log(p)      -0.18008    0.01354   -13.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2309 on 998 degrees of freedom
Multiple R-squared:  0.1506,     Adjusted R-squared:  0.1497
F-statistic: 176.9 on 1 and 998 DF,  p-value: < 2.2e-16
```

Estimation of log-linear model II

- Log-linear model
  - reg2 <-lm(log(q) ~log(p), data = price_data)
  - summary(reg2)
- We are interested in if the coefficient is higher/lower than **one**.
  - $H_0: \beta = 0$ is not consistent with our concern.
  - $H_0: \beta = 1$ should be tested.
- Use car::linearHypothesis()

```
install.packages("car")
library(car)
linearHypothesis(reg2, c("log(p) = 1"))
```

Results

- Results show that the coefficient is significantly different from 1.
  - More precisely, lower than 1 （-0.18）
  - i.e., Inelastic

```
linearHypothesis(reg2, c("log(p) = 1"))

Linear hypothesis test

Hypothesis:
log(p) = 1

Model 1: restricted model
Model 2: log(q) ~ log(p)

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    999 458.37
2    998  53.23  1    405.14 7596.7 < 2.2e-16 ***
```

Relationship between F and t statistics

- When we apply *F* statistic to the case of testing significance of a single independent variable, the *F* statistic is equal to the square of the corresponding *t* statistic.
  - t^2_{n-k-1} distribution is equal to $F_{1,n-k-1}$ distribution.
- However, *t* test is more flexible, and it can be used to one-tailed tests.

Interim exam

- Evaluating the understandings of research methods and interpretations of results
  - Take-home exam via manaba
  - Send your answers through the quiz tab on manaba by 23 October.
    - More precisely, by 23:59 on 22 October (JST).
  - A pdf version is uploaded on course news (for offline usage).
- You can discuss the questions with your classmates, but make sure you will answer them individually.

Exercise

- Use a data set “wage2” from the “wooldridge” package.
  - See a pdf for details.
- Use regression analysis and obtain an insight into people’s wages.

- Activation of the data set “wage2”.
  - install.packages("wooldridge")
  - library(wooldridge)
  - data('wage2’)
- No submission required.