Advanced Course in Marketing
09 Regression I

Takumi Tagashira

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

Datasets for final presentation

- Financial data
  - MktRes_firmdata.xls
- Survey data
  - 2022_04_Yoshida_data.xlsx
- Open-sourced data
  - You can use accessible data

Yoshida Hideo foundation survey

- “2022_04_Yoshida_data.xlsx” is available.
  - Questionnaire survey from Yoshida Hideo memorial foundation (in Japanese) https://www.yhmf.jp/aid/data/
- They collected responses from 5172 people (age: 15-64)
  - The questionnaire includes many questions (consumer behavior, traits, reactions/attitudes to advertisements etc.)
- The next slide shows exampled (translated) items.
  - Q15
  - Q20
- The questionnaire was conducted in Japanese.
  - Please let me know if you need a help to translate specific items.
  - I cannot translate all items unfortunately.

Questionnaire items

Variable	Item
q15_1	Price is the most important factor when buying things.
q15_2	Price is very important when choosing a product.
q15_3	I usually buy the cheapest product.
q15_4	I don't care about price when buying food. *reverse coded
q20_1	Quality is more important than price when buying food.
q20_2	I am willing to pay a little more for a quality product.
q20_3	I always want the best quality.
q20_4	Quality is a deciding factor when purchasing products.

The items here are back-translated results from the Japanese items.

Open-sourced data

- There are accessible datasets.
  - E.g., Kaggle
- Using such data is another option for the final presentation.
- Please ensure to report the data source in your presentation if you use an open-sourced dataset.

Agenda

- Correlation and linear model
- Multiple regression model and interpretation

Data preparation

- We use a dataset “MktRes_firmdata.xlsx” in this session.
- This data include about 160 companies from 2010 to 2019 in the retail and service sector that are the subject of Japanese customer satisfaction index surveys at the Japan Productivity Center.

```
firmdata <- readxl::read_xlsx("data/MktRes_firmdata.xlsx")
```

Variables in firmdata

- fyear: Fiscal year
- legalname: Company name
- ind_en: Name of industry
- parent: Name of parent company (if any)
- fiscal_month: Fiscal month
- current_liability: Current liabilities
- ltloans: Long-term debt
- total_liability: total liabilities
- current_assets: current assets
- ppent: Property, plant and equipment
- total_assets: total assets
- net_assets_per_capital: Total net assets/total equity
- sales: Net sales
- sga: Selling, general and administrative expenses
- operating_profit: Operating profit
- net_profit: Net income
- pnet_profit: Net income attributable to owners of the parent (consolidated) / Net income (non-consolidated)
- re: Retained earnings
- adv: Advertising and promotion expenses
- labor_cost: Labor cost
- rd: Research and development expenses
- other_sg: Other selling, general and administrative expenses
- emp: Number of employees at end of period
- temp: Average number of temporary employees
- tempratio: temp/(emp+temp)
- indgrowth: Industry growth rate
- adint: Advertising concentration ratio (adv/sales)
- rdint: research concentration ratio (rd/sales)
- mkexp: (sga - rd) / sales
- op: operating_profit / sales
- roa: pnet_profit / total_assets

*Unit: “emp” and “temp”: number of people, Others: million yen

Data preparation II

- This dataset includes multiple samples and years.
  - i.e., panel dataset
  - We extract information from a specific year because panel data analysis is a more advanced topic than this course.
    - We focus on 2019.

```
library(tidyverse)

firmdata19 <- firmdata %>%
  filter(fyear == 2019)
```

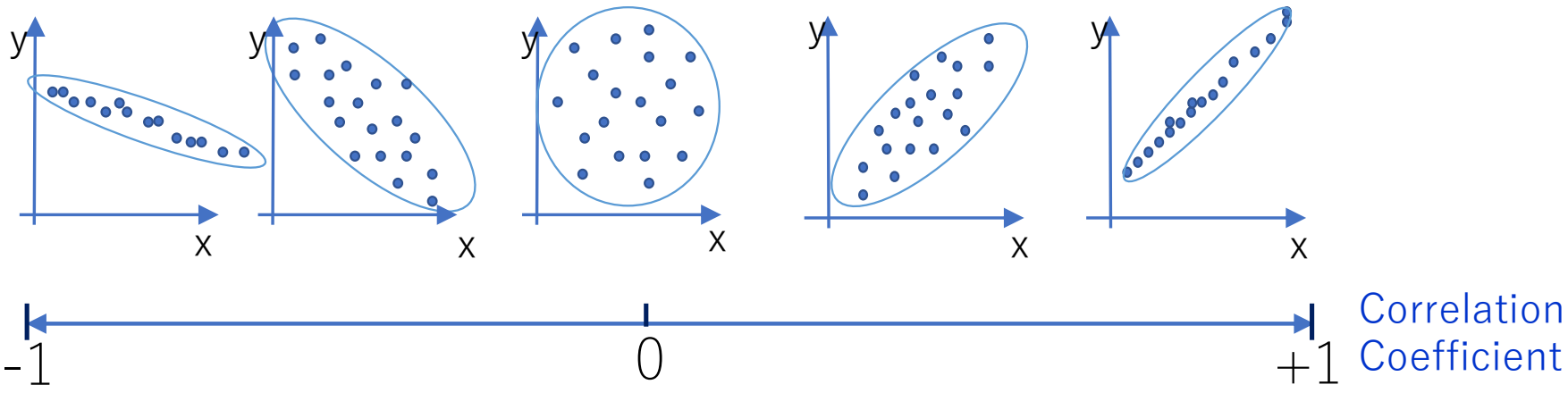
Marketing and prediction

- Firms want to improve their performances (e.g. sales or profit) by marketing practices or investments.

- It is better if we can predict the expected return by the decision making (practices/investment).
  - e.g. To what extent will an introduced new service technology increase sales?
- Verifying whether a practice has a significant effect on performance is also important.
  - This course mainly focuses on verification rather than prediction.

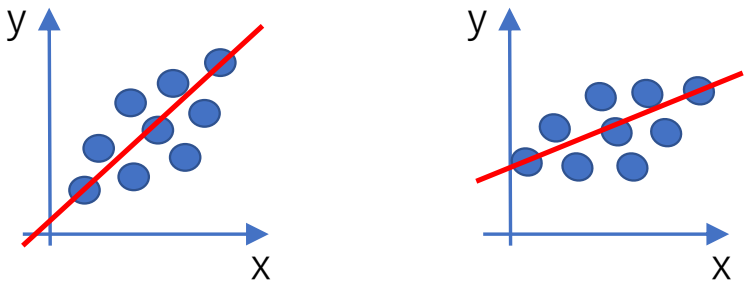
Correlation

- We have covered correlation coefficient as an indicator of the relationship between two variables.
- Correlation can be visualized by scatter plots



Linear relationship and correlation

- Correlation shows the strength of linearity.
- It does not assess the characteristics of the linear function.
  - e.g., Slope and intercept



Linear regression model

- The simplest model is to predict a dependent variable with following linear function.

- $y = \underset{\text{Intercept}}{\beta_0} + \underset{\text{Slope}}{\beta_1 x}$

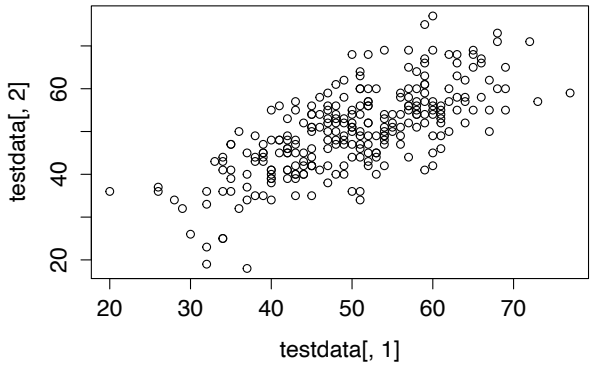
  - β_0, β_1 : Regression coefficients

- Linear regression model:
  - Regression models that employ linear function.
- Regression line:
  - Lines that show the relationship between dependent and independent variables and are derived by regression analysis.

Introduction to regression analysis

Linear regression model

- $y = \beta_0 + \beta_1 x$
- Actual datasets do not show the perfect linear relationship.
  - Linear relationship and error.
- Researchers can observe x and y but not regression coefficients.
  - We estimate coefficients from obtained data.



Estimation of coefficients

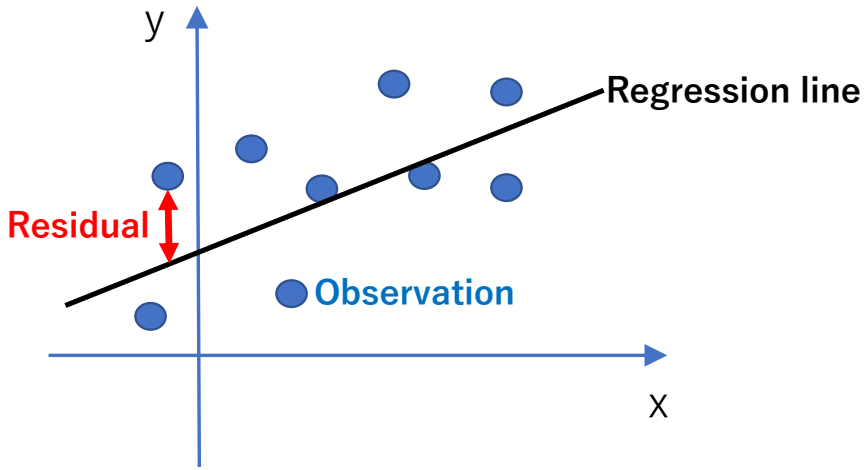
- We should find $\hat{\beta}_0$ and $\hat{\beta}_1$ which are estimates of $\beta_0$and β_1.
- Regarding the relationship between y and x, $\hat{\beta}_0$ and $\hat{\beta}_1$ define a fitted (or predicted) value of y ($\hat{y}_i$) when $x = x_i$.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- There are several methods to estimate $\hat{\alpha}$ and $\hat{\beta}$.
  - Ordinary Least Square (OLS)
  - Method of Moment
  - Method of Maximum likelihood

Intuition of OLS estimator

- OLSE is calculated by minimizing differences between regression line and observed value (sum of squared value of residuals).



OLSE

- OLSEs are defined as follows.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2}$$

Regression model in R

- Definition of regression model in R:

y tilde ~ x1+x2

- e.g., y: sales, x1: advertisement, x2: price

sales ~ advertisement + price

Regression analysis in R

- Using lm() function for linear regression analysis in R.
- Let's analyze the relationship between sales and the number of employees with firmdata19.
  - "coef()" function reports estimated coefficients.

```
reg1 <- lm(sales ~ emp, data = firmdata19)

coef(reg1)
```

Regression analysis in R

- Using lm() function for linear regression analysis in R.
- Let's analyze the relationship between sales and the number of employees with firmdata19.
  - "coef()" function reports estimated coefficients.

```
reg1 <- lm(sales ~ emp, data = firmdata19)

coef(reg1)
```

- The results shows that intercept and slope are 22809 and 58.1, respectively.
  - Thus, the "predicted value" of sales when the number of employees is 10, is 23390.7.

Properties of residuals

1. Sum of residual is 0:

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

2. Sum of products of residual and independent variable is 0:

$$\sum x_i \hat{u}_i = 0$$

- Based on (1) and (2):

$$\sum \hat{y}_i \hat{u}_i = 0$$

(Imp.) Properties of residuals

3. Mean of predicted value equals to predicted value of mean:

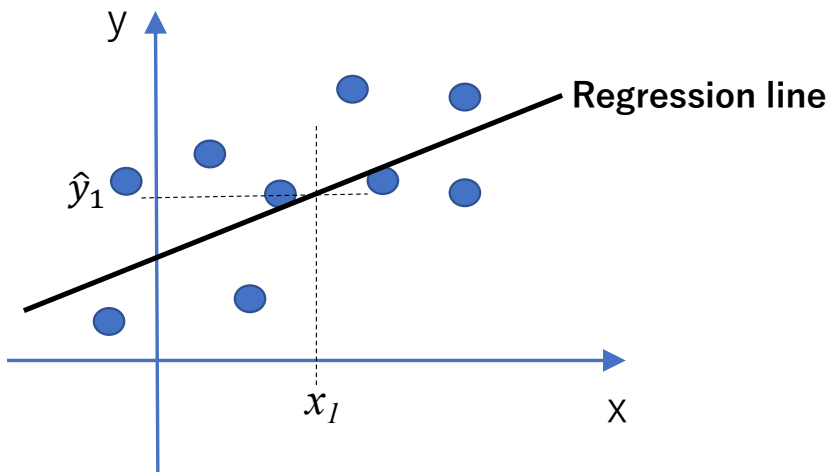
$$\bar{y} = \bar{\hat{y}}$$

4. OLS line passes $(\bar{x}, \bar{y})$:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

Interpretation of the results

- Predicted value captures a condition that the sum of residuals is 0 (i.e., gap toward positive and negative directions are balanced).
  - Predicted value $\hat{y}$ is interpreted as the mean value of y with given x.
  - The relationship between regression model and mean will be covered later.



Goodness of fit

- There is a measurement about the model fit.
- R^2 is used in general.
- R^2:
  - A percentage that the independent variable explains the variance of the dependent variable.
- $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = \frac{\sum(\hat{y}_i - \bar{y}_i)^2}{\sum(y_i - \bar{y})^2}$
- $0 \leq R^2 \leq 1$
- Interpretation:
- For example, when R^2 is 0.80, the model explains 80% of the dependent variable's variance.

Goodness of fit II

- R^2 is a measurement expressing how much the model explains/predicts the dependent variable well.
- This measurement becomes especially important for prediction.
  - However, advanced methods such as machine learning are more suitable for prediction.
- When we don't focus on the prediction, R^2 is less important.
  - e.g., Verification or interpretation of variables' effects
- The research and analytical designs to properly assess the effect of the focal variable become more prominent in verification.

Interpretation of R outputs

```
summary(fit_spent)
```

- More on that later

Estimated coefficients for each parameter (intercept and slope)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -76.48     167.76  -0.456    0.649
frequency    8277.01      41.11 201.356   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5624 on 1485 degrees of freedom
Multiple R-squared:  0.9647,    Adjusted R-squared:  0.9646
F-statistic: 4.054e+04 on 1 and 1485 DF,  p-value: < 2.2e-16
```

R2: The model explains 96% of the Monetary’s variance

Inference and test in regression

Practical procedures of an analysis

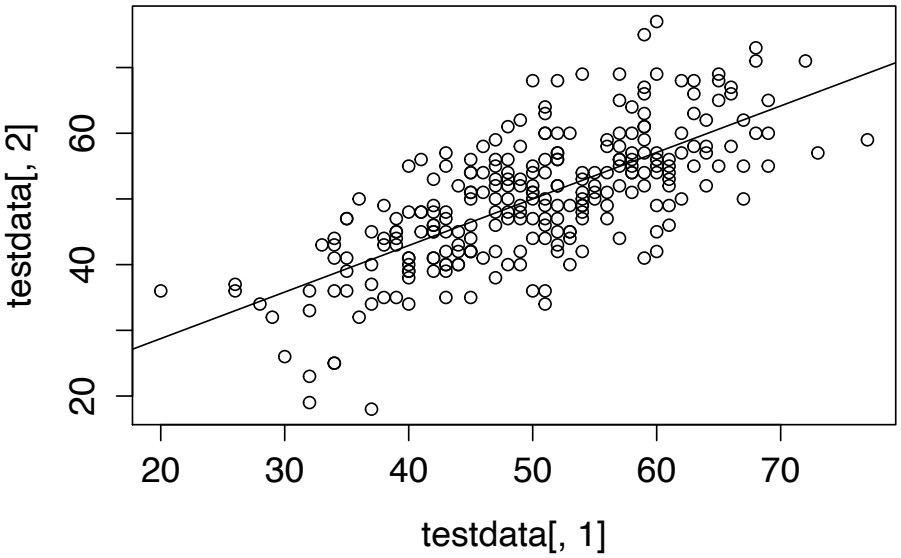
1. Define a model with lm() function.
2. Report the results with summary() function
3. Check the reported value and sign of the coefficients.
4. Check the test/inference results.
   - When p-value is lower than 0.05 (0.01), H0 can be rejected with 5% (1%) significance level.
   - i.e. The coefficient is significantly different from zero (+ or -).
   - -> We conclude that X has a significant effect on Y.
   - Confidence interval:
     - When the interval does not include zero, we often conclude that X has an impact on Y.

Estimated coefficients

- Some people misunderstand that the computed regression coefficient is a true value (i.e. parameter).
- Even if the estimate takes a positive value, it may be indifferent from zero in the true value (i.e., population parameter).
  - The population parameter is unknown.
- Thus, we employ statistical tests and interpret the coefficients.
- Understanding the theoretical aspects of regression analysis provides better implications.

Theoretical model of simple regression

- The following model explains y by x.
- $y = \beta_0 + \beta_1 x + u$.
  - where, u is an error term, β_0 is intercept, and β_1 is slope parameter.
  - We usually write theoretical models in analytical reports.
- Theoretical models include stochastic error term u to explain y.
  - y: random variable, but x: constant.
  - Thus, OLSEs are also random variables.
- Then, what does this linear function mean?



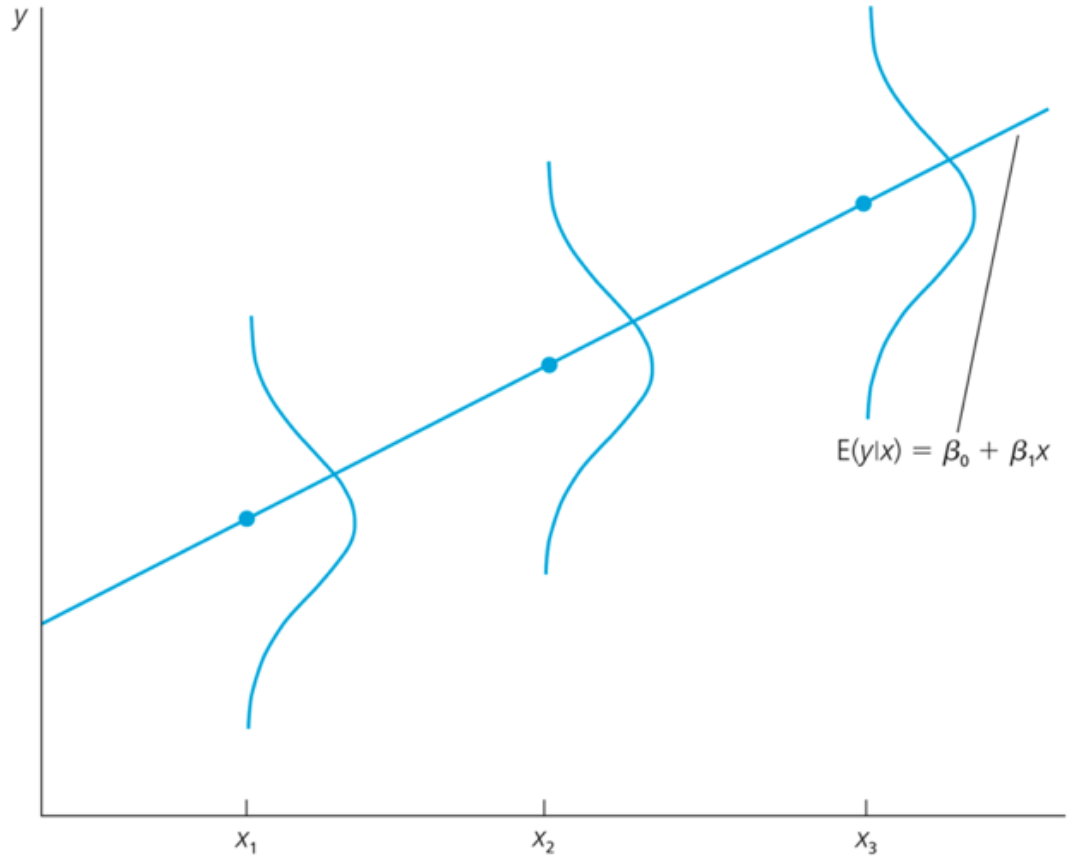
Regression model and conditional mean

- Implication of (theoretical) regression model
  - the conditional expectation of y given x is shown as a linear function of x. (population regression model).

$$E(y|x) = \beta_0 + \beta_1 x$$

- This specification means that a one-unit increase in x changes the expected value of y by amount β_1.
  - e.g. Tall people are heavier on average.
- Theoretical relationship (linear model) and observable data are not identical.

Regression model and conditional mean



Source: Wooldridge (2013) "Introductory Econometrics", p.24.

Assumptions

- $E(u) = 0$
- $E(u|x) = 0$
- $Var(u) = E(u^2) = \sigma^2$
- $Cov(u, x) = E(xu) = 0$
- Error term (u) follows a normal distribution.
  - This assumption is used for statistical tests.

Statistical properties of OLSE

- Unbiasedness:
  - $E(\hat{\beta}) = \beta$

- Asymptotic nature:
  - $\hat{\beta}$ follows $N\left(\beta, s.e.(\hat{\beta})^2\right)$ when sample size is sufficiently large.
    - where $s.e.(\hat{\beta})$ is standard error of OLSE.

Tests for regression coefficients

- We can check the test results by “summary()” function based on the “lm()” results.

```
summary(reg1)
```

Estimated results of each parameter (intercept and slope)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 22809.679  88556.098   0.258    0.797
emp            58.132      2.559  22.721   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Coefficient of intercept is not significant.
The effect of emp is positive and significant

R2: The model explains 78% of the sales

```
Residual standard error: 872800 on 145 degrees of freedom
Multiple R-squared:  0.7807,    Adjusted R-squared:  0.7792
F-statistic: 516.2 on 1 and 145 DF,  p-value: < 2.2e-16
```

Test of regression coefficient

- Statistical software automatically returns test results.
- Such tests employ the following hypotheses (without suffix).

$$H_0: \beta = 0,\ H_1: \beta \neq 0$$

- Test statistic of $\hat{\beta}$ and β can be defined as follows.

$$t = \frac{\hat{\beta} - \beta}{s.e.(\hat{\beta})}$$

- where, $s.e.(\hat{\beta})$ refers to standard error (details skipped).
- When H0 is true, the statistic $t = \hat{\beta}/s.e.\left(\hat{\beta}\right)$ follows t-distribution with degree of freedom of (n-2).
- Further procedure is consistent with other statistical tests.

Coefficient test

$$H_0: \beta = 0,\ H_1: \beta \neq 0$$

$$t = \frac{\hat{\beta} - \beta}{s.e.(\hat{\beta})}$$

$$\begin{cases} Reject\ H_0 \ \text{ if } |t| > t_{\alpha/2}(n-2) \\ Accept\ H_0 \text{ if } |t| \leq t_{\alpha/2}(n-2) \end{cases}$$

Confidence interval I

- The previous test confirms $\mathbf{H_0}$: $\beta_1 = 0$.
- The confidence interval of the coefficient parameters can be calculated based on a similar procedure to the one from the previous session.
  - E.g. confidence interval with 95% confidence level.
- Based on the asymptotic property of OLSE and central limit theorem, the following statistic follows standard normal distribution when the sample size is sufficiently large.

$$\frac{(\hat{\beta}_1 - \beta_1)}{s.e.(\hat{\beta}_1)}$$

Confidence interval II

- Let α denote the confidence level and we will obtain the following:

$$P\left(\left|\frac{\hat{\beta}_1 - \beta_1}{s.e.(\hat{\beta}_1)}\right| \leq z_{\alpha/2}\right) = 1 - \alpha$$

- Transforming the Inequalities and the interval is as follows:

$$P\left(\hat{\beta}_1 - s.e.(\hat{\beta}_1)z_{\alpha/2} \leq \beta_1 \leq \hat{\beta}_1 + s.e.(\hat{\beta}_1)z_{\alpha/2}\right) = 1 - \alpha$$

- Thus, the observable interval $\left[\hat{\beta}_1 \pm s.e.(\hat{\beta}_1)z_{\alpha/2}\right]$ represents the confidence interval of the unknown parameter.
- $z_{\alpha/2}$ should be computed based on the pre-determined confidence level.

(Sup.) Confidence interval III

- If you want to use the t-distribution, the interval can be specified as follows.
- With the confidence level α:

$$P\left(\left|\frac{\hat{\beta}_1 - \beta_1}{s.e.(\hat{\beta}_1)}\right| \leq t_{\alpha/2}(n-2)\right) = 1 - \alpha.$$

- The interval can be summarized as follows.

$$P\left(\hat{\beta}_1 - s.e.(\hat{\beta}_1)t_{\alpha/2}(n-2) \leq \beta_1 \leq \hat{\beta}_1 + s.e.(\hat{\beta}_1)t_{\alpha/2}(n-2)\right) = 1 - \alpha$$

- Thus, $\left[\hat{\beta}_1 \pm s.e.(\hat{\beta}_1)t_{\alpha/2}(n-2)\right]$ represents the confidence interval with t-distribution.

Confidence interval in regression

- confint() reports/extracts confidence interval of the coefficients.
- Default assumes the 95%confidence interval and normal distribution (Refer to ?confint for details)

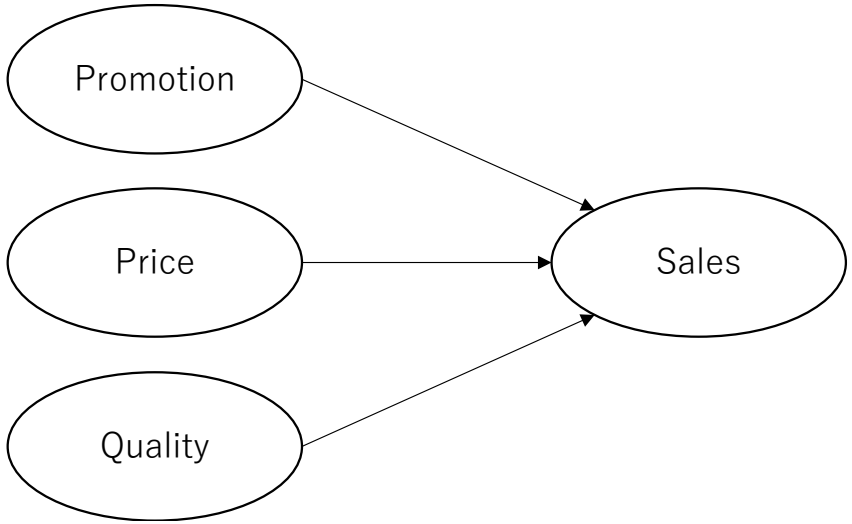
```
confint(reg1,level = 0.99)
```

- The results show that the confidence interval of the emp's coefficient is [51.45, 64.81].
  - If you want to compute the interval with different confidence level, specify the "level=" argument as: confint(fit_spent, level = 0.90)
- On the other hand, the confidence interval for intercept includes 0.

Multiple regression model

Using multiple independent variables

- We often want to explain a dependent variable by multiple independent variables.
- E.g. :
  - Is promotion the sole sufficient variable to explain firm sales?
  - What about the (product) quality and price ?



Multiple regression model

- We can employ multiple independent variables for a regression model.
  - i.e., Multiple regression model
- The following is a model with two independent variables.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

- OLSE can be estimated by matrix calculation.

$$Y = X'\beta + u$$
$$\hat{\beta} = (X'X)^{-1}X'Y$$

- t-tests should be conducted for each variable separately.

Practical instruction

- Principle:
  - We basically use multiple regression models rather than simple regression practically.
- We are usually interested in multiple explaining variables.
  - In most cases, we cannot ignore the other variables.
- We should employ a multiple regression analysis that includes various variables.
- Why?
  - The interpretation of coefficients is different.
  - We will cover later.

Multiple regression in R I

- We will use “firmdata19” to check the relationship between operating profit and advertising expense.
- We will start with a simple regression analysis (for an educational purpose).
  - You don’t need to follow this procedure in the actual analysis.

```
reg2 <- lm(operating_profit ~ adv, data = firmdata19)
summary(reg2)
confint(reg2)
```

Summary

- Simple regression analysis shows that advertising expense has a positive effect, and the confidence interval is [0.83, 1.68].
- Any other factors to control?
- We assess the following factors:
- Inter-personal service factor:
  - Number of employees and temporal workers, and labor costs
- Equipment and investment:
  - Total assets, R&D

Multiple regression in R II

- We can conduct multiple regression by adding variables:
  - E.g. “lm(y ~ x1 + x2, data = df).”

```
reg3 <- lm(operating_profit ~ adv + temp + emp +
             labor_cost + total_assets + rd, data = firmdata19)

summary(reg3)
confint(reg3)
```

- Result?

Goodness of fit

- R^2
- F-test

Adjusted R^2

- Ordinary R^2 increases as the number of independent variable increases (More strictly, its non-decreasing).

  → Use adjusted R-squared($\bar{R}^2$ or adj. R^2).
- $\bar{R}^2$ controls for the effect of the number of independent variables.

Goodness of fit and F-test

- We can test if all coefficients except β_0 (intercept) are zero or not (assume there are k independent variables).
- $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + u_i$ (Full model)

  $H_0: \beta_1 = \cdots = \beta_k = 0$, H_1: At least one coefficient is not 0.
- When H_0 is true, the model can be specified as follows.
- $y_i = \beta_0 + u_i$ (Model 0)
- The ratio between sum of squared residual (SSR) of full model and SSR of model 0 follows F-distribution with the degree of freedom of (k, n-k-1).
- Using this characteristics, checking if the all-slopes' coefficients are zero or not.

  →F-test.

Goodness of fit and F-test II

- Suppose SSR_1 and SSR_0 are SSR of full model and model 0, respectively.
  - $SSR_1 = \sum \hat{u}_i = \sum y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik})$
  - $SSR_0 = \sum \hat{e}_i = \sum y_i - \bar{y}$
- If the null hypothesis is true, the following F-value follows a F-distribution with (k, n-k-1) degree of freedom.

$$F = \frac{(SSR_0 - SSR_1)/k}{SSR_1/(n-k-1)} = \frac{SSR_0 - SSR_1}{SSR_1} \times \frac{n-k-1}{k}$$

Interpretation of multiple regression results

Aims of multiple regression model

- To capture multiple factors to explain the dependent variable.
  - e.g.,
    - Dependent variable: sales
    - Independent: advertisement, price, etc.

- To control the effects of other variables
  - A coefficient can be interpreted as an effect of independent variable while controlling for the effects of other independent variables in the model.
  - **Making other independent variables equal**.

Interpretation of multiple OLSE

- Each coefficient can be interpreted as a **partial effect**.
  - i.e., Partialled out nature of OLSE
- For a regression model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$, the predicted value of y is as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

- The change of predicted value $(\Delta\hat{y})$ by the changes of x_1 and x_2 can be defined as follows:

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2$$

Interpretation of multiple OLSE contd.

- When we fix x_2 (i.e., $\Delta x_2 = 0$),
$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1.$$
- Thus, $\hat{\beta}_1$ represents the effect of x_1 on y while treating the x_{2i} as a constant.
  - i.e., other variables are controlled.
- This property can be applied to the model with k independent variables.
- See the supplementary file for more detailed explanation.

Application of partialling out nature

- A coefficient shows the effect an independent variable while controlling for the effects of other independent variables.
- How can we use this property for analysis?
- Control variables.

Revisit the analysis

- Can you explain the result regarding the relationship between **operating_profit** and **adv**?

```
reg3 <- lm(operating_profit ~ adv + temp + emp +
             labor_cost + total_assets + rd, data = firmdata19)

summary(reg3)
confint(reg3)
```

Results

- Output format of the results is the same as simple regression.
- The results show that the coefficient of “adv” is negative and significant, and the confidence interval is [-2.01, -0.84].
- When we control for labor, assets, and R&D, advertising expense has a negative effect on operating profit.
- Other variables:
  - The number of laborers reduces profit.
  - The labor cost leads to higher profit when we control the number of labors.
  - Assets and R&D leads to higher profit.

Summary of multiple regression model

- We can employ more than two independent variables in regression models.
- Multiple regression analyses can analyze the effect of an independent variable on the dependent variable while controlling for other independent variables as constant.
  - This is a very important property.
- Thus, we usually employ multiple regression models.
  - For example, when we want to analyze the effects of x1 and x2 on y, we **DONOT** employ the following two-separated simple regression models.
    1. $y = \alpha_0 + \alpha_1 x_1 + u$
    2. $y = \beta_0 + \beta_1 x_2 + e$
- We usually refer to existing research to specify the relevant control variables.

Exercise

- Conduct regression analysis with firmdata19.
- Pick up a dependent variable and a main independent variable.
- Specify a multiple regression model
  - i.e., identify the relevant control variables
- No submission required.