Chapter 3. Regression

3.1 Exploratory Data Analysis

Objectives

- Examine that relationship between two continuous variables using a scatter plot.
- Quantify the degree of association between two continuous variables using correlation statistics.
- Understand potential misuses of the correlation coefficient
- Obtain Pearson correlation coefficients using the CORR procedure.

Scatter Plots (산점도)

Scatter plots are two-dimensional graphs produced by plotting one variable against another within a set of coordinate axes. The coordinates of each point correspond to the values of the two variables.

➔ 데이터의 특징을 미리 탐색하기에 용이.

Scatter plots are useful to

- explore the relationship between two variables
- locate outlying or unusual values
- identify possible trends
- identify a basic range of Y and X values
- communicate data analysis results.

Pearson Correlation Coefficient

Correlation statistic measure the degree of linear association between two variables. A common correlation statistic used for continuous variables is the Pearson correlation coefficient.

➔ 상관계수는 두 변수 간의 선형성만 파악할 수 있다.

Hypothesis Test for a Correlation

- The parameter representing a correlation is ρ.
- ρ is estimated by the simple statistic r.
- H0: $\rho = 0$
- Rejecting H0 indicates only great confidence that ρ is not exactly zero.

- A p-value does not measure the magnitude of the association.
- Sample size affects that p-value.

  → Small p-values can occur because of very large sample sizes. Therefore, it is important to also look at the value of r itself to see if it is a meaningfully large correlation.

Correlation versus Causation

Common errors can be made when interpreting the correlation between variables. One example of this is using correlation coefficients to conclude a cause-and-effect relationship.

➔ 상관관계가 존재한다는 것이 인과관계가 있다는 것을 의미하지는 않는다.

- A strong correlation between two variables does not mean change in one variable causes the other variable to change, or vice versa.
- Sample correlation coefficients can be large because of chance or because both variables are affected by other variables.
- "Correlation does not imply causation."

Missing Another Type of Relationship

➔ 상관계수는 선형성만 파악할 수 있으므로 이차함수 관계, 주기가 있는 관계 등은 파악할 수 없다.

Extreme Data Values

Correlation coefficients are highly affected by a few extreme values of either variable.

➔ 이상점 (outlier)에 영향을 받는다.

➔ 이상점을 포함한 경우와 포함하지 않은 경우의 상관계수를 비교해보고 상관계수에 큰 차이를 보인다면 그 점은 influential point (영향점) 이다.

The CORR Procedure

```
PROC CORR DATA=SAS-data-set <options>;
    VAR variables;
    WITH variables;
    ID variables;
RUN;
```

VAR specifies variables for which to produce correlations. If **WITH** statement is not specified, correlations are produced for each pair of variables in the **VAR** statement. If the **WITH** statement is specified, the **VAR** statement specifies the column variables in the correlation matrix.

WITH produces correlations for each variable in the **VAR** statement with all variables in the **WITH** statement. The **WITH** statement specifies the row variables in the correlation matrix.

ID The **ID** statement specifies one or more additional tip variables to identify observations in scatter plots and scatter plot matrix.

3.2 Simple Linear Regression

Overview

The *response variable* (y; 종속변수, 반응변수) is the variable of primary interest.

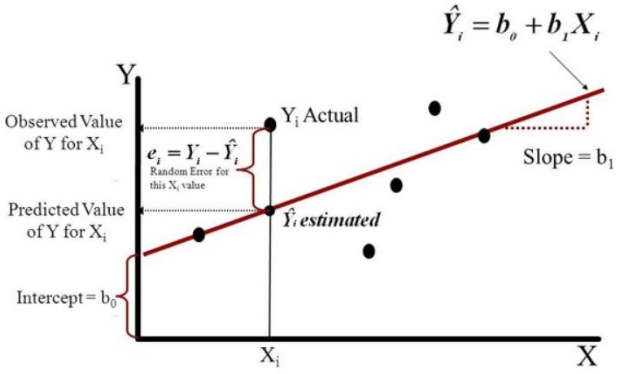
The *predictor variable* (x; 설명변수, 예측변수, 독립변수) is used to explain the variability in the response variable.

Simple Linear Regression

The objectives of simple linear regression are to

- assess the significance of the predictor variable in explaining the variability or behavior of the response variable
- predict the values of the response variable given the values of the predictor variable

Simple Linear Regression Model

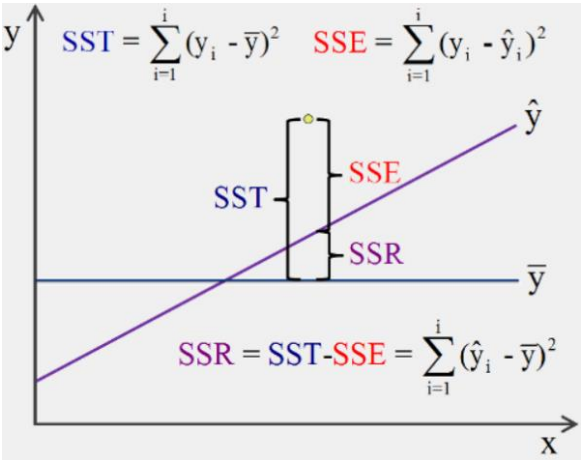


True Model: $Y = \beta_0 + \beta_1 X$

Fitted Model: $\hat{Y} = b_0 + b_1 X$

Estimates of the unknown population parameters β_0 and β_1 are obtained by the method of least squares. This method provides the estimates by determining the line that minimizes the sum of the squared vertical distances between the observations and the fitted line. In other words, the fitted or regression line is as close as possible to all the data points.

Explained versus Unexplained Variability



Explained variability (SSR)

- is related to the difference between the regression line and the mean of the response variable. The model sum of squares (SSR) is the amount of variability explained by your model.

Unexplained variability (SSE)

- is related to the difference between the observed values and the regression line. The error sum of squares (SSE) is the amount of variability unexplained by your model.

Total variability (SST)

- is related to the difference between the observed values and the mean of the response variable. The corrected total sum of squares is the sum of the explained and unexplained variability.

Model Hypothesis Test

Null Hypothesis

- The simple linear regression model does not fit the data better than the baseline model.
- $\beta_1 = 0$

Alternative Hypothesis

- The simple linear regression model does fit the data better than the baseline model.
- $\beta_1 \neq 0$

The REG Procedure

```
PROC REG DATA=SAS-data-set <options>;
    MODEL dependent(s)=regressor(s) </ options>;
RUN;
QUIT;
```

MODEL specifies the response and predictor variables. The variables must be numeric.

3.3 Concepts of Multiple Regression (다중회귀분석)

The Multiple Linear Regression Model

In general, you model the dependent variable Y as a linear function of k independent variables, (the Xs) as

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

Model Hypothesis Test

Null Hypothesis

- The simple linear regression model does not fit the data better than the baseline model.
- $\beta_1 = 0$

Alternative Hypothesis

- The regression model does fit the data better than the baseline model.
- $\beta_1 \neq 0$
- 적어도 1개의 설명변수는 반응변수를 설명한다.

Assumptions for Linear Regression

- The mean of Ys is accurately modeled by a linear function of the Xs.
- The random error term, ϵ, is assumed to have a normal distribution with a mean of zero.
- The random error term, ϵ, is assumed to have a constant variance, σ^2.
- The errors are independent.

Multiple Linear Regression versus Simple Linear Regression

Main Advantage

- Multiple linear regression enables you to investigate the relationship among Y and several independent variables simultaneously.

Main Disadvantage

- Increased complexity makes it more difficult to
  - ascertain which model is "best"
  - interpret the models.

3.4 Model Building and Interpretation

Model Selection

Eliminating one variable at a time manually for

- small data sets is a reasonable approach
- large data sets can take an extreme amount of time.

Mallows' C_p

$$C_p = p + \frac{(MSE_p - MSE_{full})(n-p)}{MSE_{full}}$$

where

MSE_p is the mean squared error for the model with p parameters.

MSE_{full} is the mean squared error for the full model used to estimate the true residual variance

n is the number of observations

p is the number of parameters, including an intercept parameter, if estimated.

- Mallows' C_p is a simple indicator of model bias. Models with a large C_p are biased.
- Look for models with $C_p \leq p$, where p equals the number of parameters in the model, including the intercept.

➔ Mallows recommends choosing the first (fewest variables) model where C_p approaches p.

Hocking's Criterion

Hocking (1976) suggests selecting a model based on the following:

- $C_p \leq p$ for prediction
- $C_p \leq 2p - p_{full} + 1$ for parameter estimation

Model Selection Methods

- Forward Selection
  - starts with an empty model.
  - computes an F statistic for each predictor variable not in the model and examines the largest of these statistics. If it is significant at a specified

significance level, the corresponding variable is added to the model. Once a variable is entered in the model, it is never removed from the model. The process is repeated until none of the remaining variables meet the specified level of entry.

- Backward Elimination
  - starts off with the full model.
  - Results of the F test for individual parameter estimates are examined, and the least significant variable that falls above the specified significance level is removed. After a variable is removed from the model, it remains excluded. The process is repeated until no other variable in the model meets the specified significance level for removal.
- Stepwise Selection
  - works like a combination of forward selection and backward elimination.
  - similar to forward selection in that it starts with an empty model and incrementally builds a model one variable at a time. However, the method differs from forward selection in that variables already in the model do not necessarily remain. The backward component of the method removes variables from the model that do not meet the significance criteria. The stepwise selection process terminates if no further variable can be added to the model or if the variable just entered into the model is the only variable removed in the subsequent backward elimination.

Conservative Significance Levels

Sample Size				
Evidence	**30**	**50**	**100**	**1000**
Weak	0.076	0.053	0.032	0.009
Positive	0.028	0.019	0.010	0.003
Strong	0.005	0.003	0.001	0.0003
Very Strong	0.001	0.0005	0.0001	0.00004

When you have a large number of variables, prespecifying the model is not feasible. Therefore, analysts use the results of the model selection techniques to select a model. However, the p-values calculated in the model selection techniques are not p-values in the traditional hypothesis-testing context. Instead, they should be viewed as indicators of relative importance among variables.

For large samples sizes, much smaller p-values are required to imply that the data provide evidence for the effect of interest.

Comparison of Selection Methods

- Stepwise regression
  - uses fewer computer resources.
  - has an advantage when there is a large number of independent variables.
- All-possible regression
  - generates more candidate models that might have nearly equal R^2 statistics and C_p statistics.
  - you can compare essentially equivalent models and use your knowledge of the data set and subject area to select a model that is more easily interpreted.