Regression and its Applications

하지운

Sungkyunkwan University

통계학과

Contents

Basic Statistics

Basic Statistics

1. t - distribution

Let X be a random variable. Then we say $X \sim t_n$ if its pdf is

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}, x \in \mathbb{R}, n > 0$$

2. F - distribution

Let X be a random variable. Then we say $X \sim F_{n_1,n_2}$ if its pdf is

$$f(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}$$

3. Basic Terms

Let $(\Omega, \mathcal{Y}, \mathbf{P})$ be a probability space where Ω is the sample space, $\mathcal{Y}$ is a σ-field on Ω, and $\mathbf{P}$ is a probability measure defined on $\mathcal{Y}$.
Let $\mathbf{X} : \Omega \to \mathcal{X}$ be a random variable (or random vector), where $\mathcal{X} \subseteq \mathbb{R}^n$ is the observation space.
Let $\mathcal{P}$ be a statistical model defined as a family of probability distributions of $\mathbf{X}$, parameterized by $\theta \in \Theta$, such that $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$.

4.Hypothesis Test

Let $\Theta = \Theta_0 \bigsqcup \Theta_1$ be a measurable parameter space.
We test the null hypothesis $H_0 : \theta \in \Theta_0$ against the alternative hypothesis $H_1 : \theta \in \Theta_1$.
A test procedure is typically defined by a rejection region $R \subset \mathcal{X}$ or a measurable test function $\phi : \mathcal{X} \to [0, 1]$.

Simple Linear Regression Model

What is Regression?

- Response variable(Dependent Variable): the outcome variable on which comparisons are made (we usually denote response variable by y.)
- Explanatory variable(Independent variable): defines the groups to be compared with respect to values on the response variable (we usually denote explanatory variable by x.)
- Example: Response/Explanatory
  - Blood alcohol level/number of beers consumed
  - Grade on test/Amount of study time
  - Yield of corn per bushel/Amount of rainfall
- The main purpose of regression analysis with two variables is to investigate whether there is an association and to describe that association.
- An association exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable.

Simple Linear Regression Model

- Data: $(x_1, y_1), \cdots, (x_n, y_n)$
- Model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \cdots, n$
  - x_i : regressor or predictor variable or explanatory variable
  - y_i : response variable
  - β_0 : unknown intercept, β_1 : unknown slope
  - ϵ_i : random error with mean 0 and variance σ^2
- Conventionally, we assume x_i's are fixed constants and Y_i's are independent random variables.
- Meaning of β_0:
  - The predicted value for Y when $x = 0$
  - Helps in plotting the line
  - May not have any interpretative value if no observations had x values near 0
- Meaning of β_1 : measures the change in the predicted variable(Y) for a 1 unit increase in the explanatory variable in x

Simple Linear Regression Model

- Data: $(x_1, y_1), \cdots, (x_n, y_n)$
- Model: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \cdots, n$
  - x_i : regressor or predictor variable or explanatory variable
  - Y_i : response variable
  - β_0 : unknown intercept, β_1 : unknown slope
  - ϵ_i : random error with mean 0 and variance σ^2
- Conventionally, we assume x_i's are fixed constants and Y_i's are independent random variables.
- Meaning of ϵ: The error variable. Due to this random error, the response variable can be different even with the same x.
- Meaning of σ^2 : The error variance $\rightarrow$ magnitude of uncertainty.
- Q. How to find regression line and estimate σ^2?

Simple Linear Regression Model

1. Least square estimation

Let $L(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ be a loss function.
The least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize L are given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Simple Linear Regression Model

- Residual: $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, i = 1, \cdots, n$
  - We can think residuals represent errors ϵ_i's because we have $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$ from $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - Measures the size of the prediction errors, the vertical distance between the point and the regression line.
  - Each observation has a residual by calculating $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
  - A large residual indicates an unusual observation.
- Property of residual:

$$\sum_{i=1}^{n} e_i = 0$$

- Residual sum of squares:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SSE$$

$\Rightarrow$ In fact, the least square estimators of β_0 and β_1 are the minimizers of SSE. For this reason, the regression line is also called as least squares regression line.

2. Point Estimator of σ^2

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2}$$

where $n-2$ is the degree of freedom, which makes $\hat{\sigma}^2$ is an unbiased estimator of σ^2.

Simple Linear Regression Model

- Coefficient of Determination R^2
  - Total sum of squares:

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

  - Decomposition of SST:

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = SSE + SSR$$

  - SSE (Error Sum of Squares): $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
  - SSR (Regression Sum of Squares): $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
- Coefficient of Determination R^2:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Note that $R^2 = (\hat{r}_{xy})^2$.

Simple Linear Regression Model

- Interpretation of R^2
  - $0 \leq R^2 \leq 1$
  - the proportion of variability in the response that is accounted for by the model.
  - A large value of R^2 suggests that the model has been successful in explaining the variability in the response.
  - A small value of R^2 suggests that the model does not explain the variability in the response. So we may think about alternative models.

Simple Linear Regression Model

3. Distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ when ϵ_i's are normal.

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)\right), \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$$

Simple Linear Regression Model

- Testing Hypothesis
  - $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$

$$\text{Test statistic:} \quad t_0 = \frac{\hat{\beta}_0}{\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2})}} = \frac{\hat{\beta}_0}{s.e.(\hat{\beta}_0)}$$

Reject H_0 if $|t_0| > t_{\alpha/2, n-2}$

  - $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$

$$\text{Test statistic:} \quad t_0 = \frac{\hat{\beta}_1}{\sqrt{\frac{\sigma^2}{\sum(x_i - \bar{x})^2}}} = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}$$

Reject H_0 if $|t_0| > t_{\alpha/2, n-2}$

Simple Linear Regression Model

- Confidence intervals (C.I.s) of parameters in simple linear regression
  - $(1-\alpha) \times 100\%$ C.I. for β_0:

$$\hat{\beta}_0 - t_{\alpha/2,n-2} s.e.(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2} s.e.(\hat{\beta}_0)$$

  - $(1-\alpha) \times 100\%$ C.I. for β_1:

$$\hat{\beta}_1 - t_{\alpha/2,n-2} s.e.(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2} s.e.(\hat{\beta}_1)$$

- Now we estimated the β_0 and β_1. Next: How to predict y_i?

Simple Linear Regression Model

4. Distribution of $\hat{y}_0$ and $\hat{\mu}_0$

$$\hat{y}_0 - y_0 \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)\right)$$

$$\hat{\mu}_0 - \mu_0 \sim N\left(0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)\right)$$

5. Distribution of $\hat{y}_0$ and $\hat{\mu}_0$

- $(1-\alpha) \times 100\%$ C.I. for y_0:

$$\hat{y}_0 - t_{\alpha/2,n-2} s.e.(\hat{y}_0 - y_0) \leq y_0 \leq \hat{y}_0 + t_{\alpha/2,n-2} s.e.(\hat{y}_0 - y_0)$$

- $(1-\alpha) \times 100\%$ C.I. for μ_0:

$$\hat{\mu}_0 - t_{\alpha/2,n-2} s.e.(\hat{\mu}_0 - \mu_0) \leq \mu_0 \leq \hat{\mu}_0 + t_{\alpha/2,n-2} s.e.(\hat{\mu}_0 - \mu_0)$$

How to interpret the result?

We will use data `mtcars`.

Motor Trend Car Road Tests

Description

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Usage

```
mtcars
```

Figure: `mtcars` information 1

How to interpret the result?

We will use data `mtcars`.

Format

A data frame with 32 observations on 11 (numeric) variables.

[, 1]	`mpg`	Miles/(US) gallon
[, 2]	`cyl`	Number of cylinders
[, 3]	`disp`	Displacement (cu.in.)
[, 4]	`hp`	Gross horsepower
[, 5]	`drat`	Rear axle ratio
[, 6]	`wt`	Weight (1000 lbs)
[, 7]	`qsec`	1/4 mile time
[, 8]	`vs`	Engine (0 = V-shaped, 1 = straight)
[, 9]	`am`	Transmission (0 = automatic, 1 = manual)
[,10]	`gear`	Number of forward gears
[,11]	`carb`	Number of carburetors

Figure: `mtcars` information 2

How to interpret the result?

```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Figure: `mtcars` information 3

How to interpret the result?

```r
head(mtcars)

# simple linear regression
fit_simple <- lm(mpg ~ wt, data = mtcars)

# summary
summary(fit_simple)
```

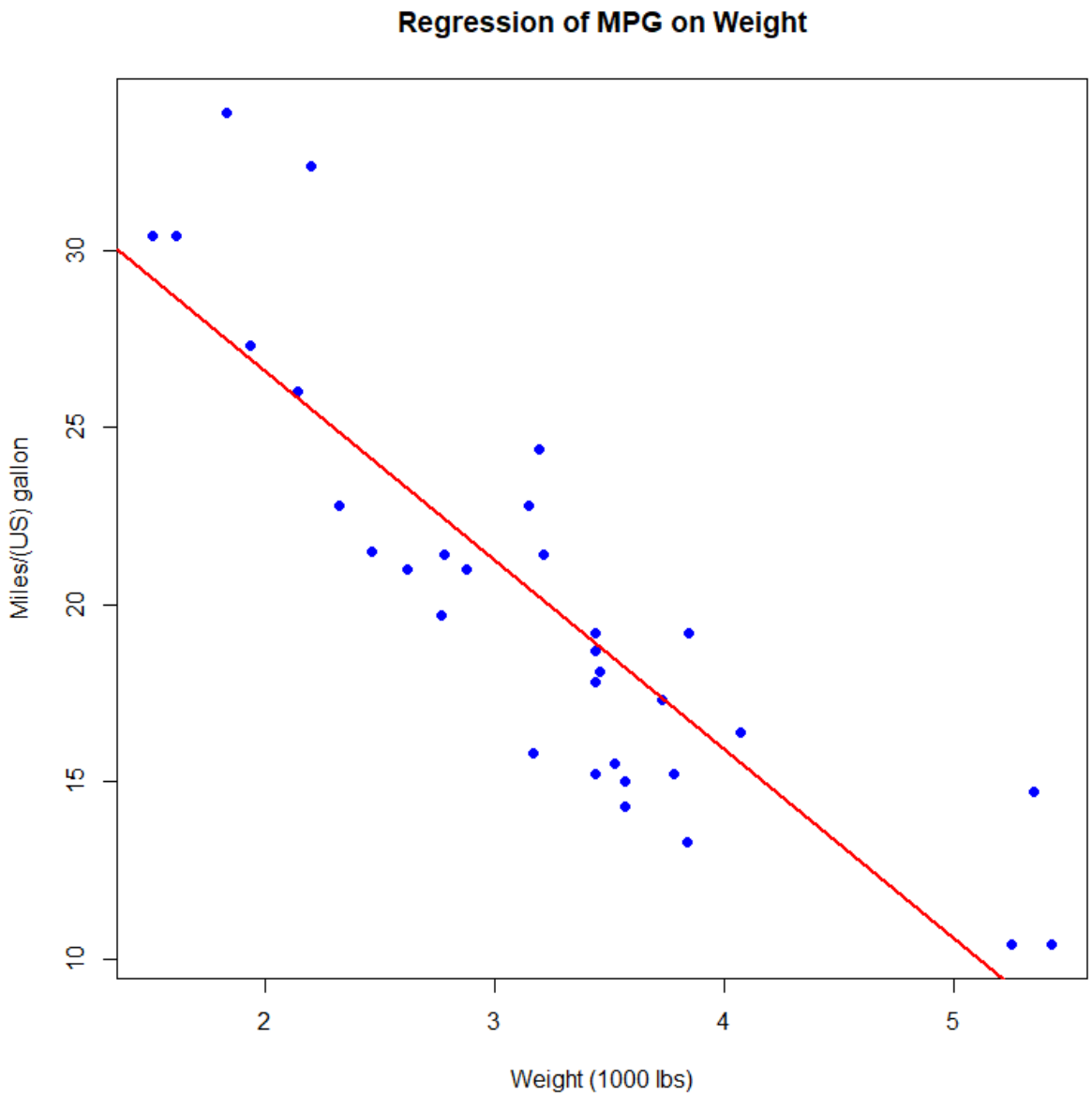
Figure: `mtcars` Simple Linear Regression 1

How to interpret the result?

```
> # simple linear regression
> fit_simple <- lm(mpg ~ wt, data = mtcars)
>
> # summary
> summary(fit_simple)

Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Figure: `mtcars` Simple Linear Regression 2

How to interpret the result?



Multiple Linear Regression Model

Multiple Linear Regression Model

- Data: $(x_{11}, \cdots, x_{p1}, y_1), \cdots, (x_{n1}, \cdots, x_{pn}, y_n)$
- Model: $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i, i = 1, \cdots, n$
  - $x_{1i}, \cdots, x_{pi}$: regressor or predictor variable or explanatory variable (Assuming fixed constants)
  - y_i : response variable
  - $\beta_0, \cdots, \beta_p$: unknown regression coefficients
  - ϵ_i : random error with mean 0 and variance σ^2
- Interpretation of regression coefficients β_j : The change of Y corresponding to a unit change of X_j when other covariates are fixed.

Multiple Linear Regression Model

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Using this matrix notation, the multiple linear regression model can be expressed in matrix form as $\mathbf{Y} = \mathbf{X}\beta + \epsilon$.
Moreover, the loss function L, the square error loss function can be expressed as

$$L = \|\mathbf{Y} - \mathbf{X}\beta\|^2 \Rightarrow \mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\beta)$$

and the normal equation is

$$(\mathbf{X}^\top\mathbf{X})\beta = \mathbf{X}^\top\mathbf{Y}$$

Now the least square estimator for β is simply

$$\hat{\beta} = \hat{\beta}^{LSE} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$$

1. Distribution of $\hat{\beta}$

Given $\mathbf{Y} \sim MVN(\mathbf{X}\beta, \sigma^2 I)$, $\sigma^2 > 0$, then

$$\hat{\beta} \sim MVN(\beta, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$$

pf)

$$E(\hat{\beta}) = E((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E(\mathbf{Y}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\beta = \beta$$

$$Var(\hat{\beta}) = Var((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Var(\mathbf{Y})((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)^\top$$
$$= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} \quad \square$$

Multiple Linear Regression Model

- Residual: $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_p x_{pi}, i = 1, \cdots, n$
  With matrix notation,

$$\mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y} = (I - H)Y$$

  Denote $H = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$ as Hat Matrix.

- Property of residual:

$$\sum_{i=1}^{n} e_i = 0$$

- Residual sum of squares:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \mathbf{e}^\top \mathbf{e} = SSE$$

Multiple Linear Regression Model

2. Lemma

Let $H = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$. Then

1. $HX = X$
2. $tr(H) = p + 1$
3. rank $(I - H) = tr(I - H) = n - p - 1 = df$ (degree of freedom)

3. Point estimator of σ^2

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{df} = \frac{SSE}{n - p - 1}$$

Multiple Linear Regression Model

- Multiple correlation coefficient
  - When we have only one predictor X_1, the correlation between the response variable Y and X_1 is a good measure to see the strength of overall linear relationship.
  - When we have multiple predictors $X_1, \cdots, X_p$, we may need to see the correlations of $(Y, X_1), \cdots, (Y, X_p)$. This does not give a good summary for the strength of overall linear relationship between Y and the whole set of predictors.
  - A better way to see such strength is to consider the correlation of Y and $\hat{Y}$. This correlation is called *multiple correlation coefficient*:

$$Corr(Y, \hat{Y}) = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum(y_i - \bar{y})^2}\sqrt{\sum(\hat{y} - \bar{y})^2}}$$

Multiple Linear Regression Model

- Coefficient of Determination R^2
  - Total sum of squares:

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \|Y - \bar{Y}\|^2$$

  - Decomposition of SST:

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = SSE + SSR$$

  - SSE (Error Sum of Squares): $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \|Y - \hat{Y}\|^2$
  - SSR (Regression Sum of Squares): $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \|\hat{Y} - \bar{Y}\|^2$
- Coefficient of Determination R^2:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Note that $R^2 = (Corr(Y, \hat{Y}))^2$.

Multiple Linear Regression Model

- ANOVA table for regression model

Source of Variation	Sum of squares	Degrees of Freedom	Mean square	F_0
Regression	SSR	p	$\frac{SSR}{p} = \mathrm{MSR}$	$\frac{\mathrm{MSR}}{\mathrm{MSE}}$
Error	SSE	$n-p-1$	$\frac{SSE}{n-p-1} = \mathrm{MSE}$	
Total	SST	$n-1$		

- Distribution of F_0:

$$F_0 = \frac{MSR}{MSE} \sim F_{p,n-p-1}$$

Example

Recall the data of `mtcars`.

```
 9  # multiple linear regression
10  fit_multiple1 <- lm(mpg ~ wt + hp, data = mtcars)
11
12  summary(fit_multiple1)
13
14  fit_multiple2 <- lm(mpg ~ wt + hp + cyl + disp, data = mtcars)
15
16  summary(fit_multiple2)
```

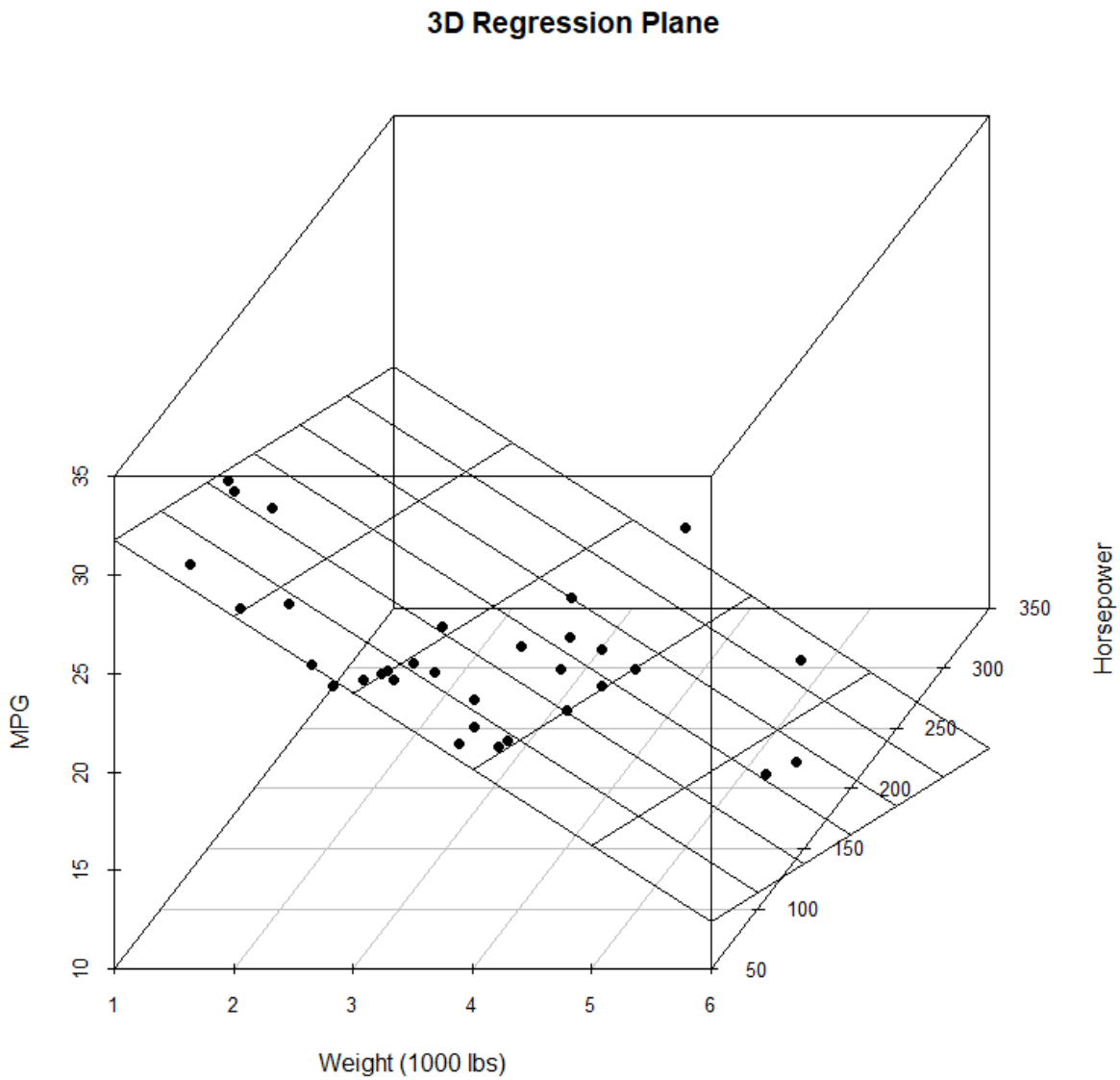
Figure: `mtcars` Multiple Linear Regression 1

Example

```
> fit_multiple1 <- lm(mpg ~ wt + hp, data = mtcars)
>
> summary(fit_multiple1)

Call:
lm(formula = mpg ~ wt + hp, data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max
-3.941 -1.600 -0.182  1.050  5.854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
hp          -0.03177    0.00903  -3.519  0.00145 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8148
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

Figure: `mtcars` Multiple Linear Regression 2

Example

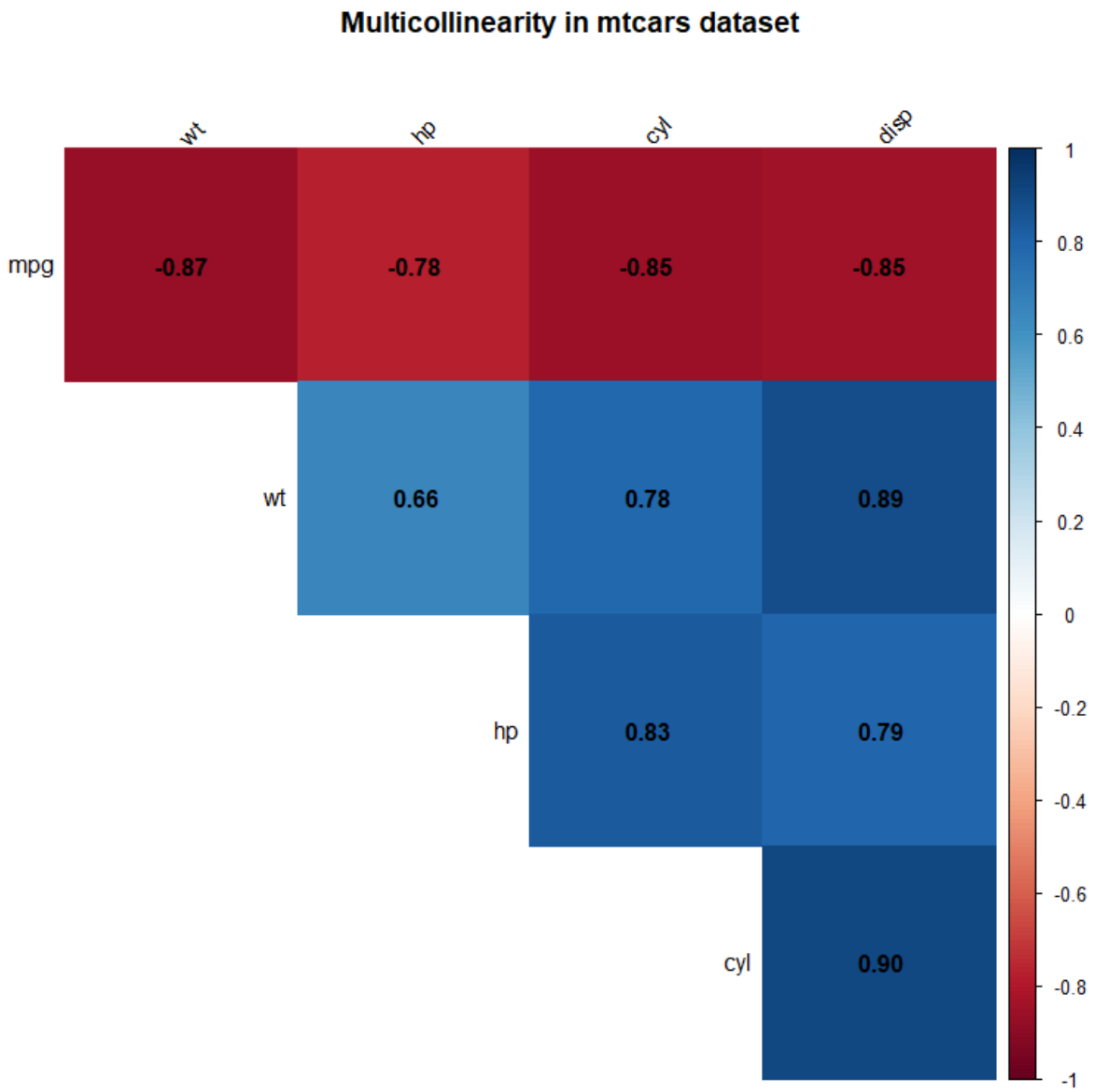

3D Regression Plane
MPG
10
15
20
25
30
35
Weight (1000 lbs)
1
2
3
4
5
6
Horsepower
50
100
150
200
250
300
350

Example

```
> fit_multiple2 <- lm(mpg ~ wt + hp + cyl + disp, data = mtcars)
>
> summary(fit_multiple2)

Call:
lm(formula = mpg ~ wt + hp + cyl + disp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0562 -1.4636 -0.4281  1.2854  5.8269

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.82854    2.75747  14.807 1.76e-14 ***
wt          -3.85390    1.01547  -3.795 0.000759 ***
hp          -0.02054    0.01215  -1.691 0.102379
cyl         -1.29332    0.65588  -1.972 0.058947 .
disp         0.01160    0.01173   0.989 0.331386
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.513 on 27 degrees of freedom
Multiple R-squared:  0.8486,    Adjusted R-squared:  0.8262
F-statistic: 37.84 on 4 and 27 DF,  p-value: 1.061e-10
```

Figure: `mtcars` Multiple Linear Regression 3

Example


Multicollinearity in mtcars dataset
wt
hp
cyl
disp
mpg
-0.87
-0.78
-0.85
-0.85
wt
0.66
0.78
0.89
hp
0.83
0.79
cyl
0.90
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Several Regression Models

Ridge Regression Model

- Let us consider $\hat{\beta}(\lambda) = (X^\top X + \lambda I)^{-1} X^\top Y$ instead of $(X^\top X)^{-1} X^\top Y$
- Ridge estimator: $\hat{\beta}(\lambda) = (X^\top X + \lambda I)^{-1} X^\top Y$
- Ridge trace: A graph λ versus $\hat{\beta}(\lambda)$
- If there is a big change around $\lambda = 0$, we may suspect *multicollinearity*.
- Using ridge trace, one can detect multicollinearity and choose the ridge coefficient λ.
- Ridge regression gives biased estimator unless $\lambda = 0$ but it can give more accurate estimate for regression coefficients.

Ridge Regression Model

- The ridge regression added a penalty (L_2 penalty):

$$\hat{\beta}_{\text{Ridge}} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta^\top X_i)^2 + \lambda \|\beta\|_2^2$$

where $\|\beta\|_2^2 = \sum_{j=1}^{d} \beta_j^2$. $\lambda\|\beta\|_2^2$ is called the L_2 penalty.

LASSO Regression Model

- One of the most famous penalized parametric regression model: it has revolutionized the modern statistical research because of its attractive properties.

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta^\top X_i)^2 + \lambda \|\beta\|_1$$

where $\|\beta\|_1 = \sum_{j=1}^{d} |\beta_j|$. $\lambda\|\beta\|_1$ is called the L_1 penalty.

- If we normalized the covariates so that $X^\top X = I_n$, the LASSO estimates can be written as

$$\hat{\beta}_{\text{LASSO},j} = \hat{\beta}_{\text{LSE},j} \times \max\{0, 1 - \frac{n\lambda}{|\hat{\beta}_{\text{LS},j}|}\}$$

for $j = 1, \cdots, n$

Logistic Regression Model

- We use Logistic Regression Model if Y is a categorical r.v.
- Let us define $\pi_i = P(Y_i = 1|X_i) = \frac{\exp(\beta_0+\beta_1 X_i)}{1+\exp(\beta_0+\beta_1 X_i)}$.
- Then the logistic model is nonlinear model where π_i is the response variable and X_i is the explanatory variable.
- However, if you transform π_i to $\log\left(\frac{\pi_i}{1-\pi_i}\right)$, the model can be expressed as a linear model. That is,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X$$

  This transformation is called *logit transformation*.
- $\frac{\pi_i}{1-\pi_i}$ is often called *odds* and $\log\left(\frac{\pi_i}{1-\pi_i}\right)$ is called *log odds*.
- Therefore, the logistic model is a linear regression model that regresses log odds on X.

Logistic Regression Model

- The logistic regression provides a simple and elegant way to characterize the function $q(x)$ in a 'linear' way. First consider

$$odds(x) = \frac{q(x)}{1 - q(x)} = \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)} \in [0, \infty)$$

  This is the *odds* that measures the contrast between the event $Y = 1$ versus $Y = 0$.

Simple Logistic Regression

- Let us consider a simple logistic regression model with a dummy explanatory variable X_i.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$$

- Interpretation of β_1: Unlike the usual linear regression model, the regression coefficient β_1 does not have the straightforward interpretation. However, we can make some valuable interpretation using Odds.

Multiple Logistic Regression

- Multiple Logistic regression model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$$

  or equivalently

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi})}{1 + \exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi})}$$

- Estimation and other inference can be made similar to the simple logistic regression model.
- We can also make similar interpretation for regression coefficients $\beta_1, \ldots, \beta_p$ to that of the simple logistic regression model.

Multinomial Logistic Regression

- What if the response variable has more than two outcome? $\rightarrow$ We can not apply logistic regression.
- Suppose that the response variable Y has three possible outcomes, say $1, 2, 3$, and X is an explanatory variable. Let $\pi_j(x_i)$ be the probability that $Y = j$ when $X = x_i, j = 1, 2, 3$.
- Multinomial Logistic Regression model:

$$\log\left(\frac{\pi_j(x_i)}{\pi_3(x_i)}\right) = \beta_{0j} + \beta_{1j}x_i, \quad j = 1, 2$$

This is equivalent to

$$\pi_j(x_i) = \frac{\exp(\beta_{0j} + \beta_{1j}x_i)}{1 + \exp(\beta_{01} + \beta_{11}x_i) + \exp(\beta_{02} + \beta_{12}x_i)}, \quad j = 1, 2$$

Multinomial Logistic Regression

- In general, when Y has k categories and there are p explanatory variables, multinomial logistic regression model is

$$\log\left(\frac{\pi_j(x_i)}{\pi_k(x_i)}\right) = \beta_{0j} + \beta_{1j}x_{1i} + \cdots + \beta_{pj}x_{pi}, \quad j = 1, 2, \ldots, (k-1)$$

or

$$\pi_j(x_i) = \frac{\exp(\beta_{0j} + \beta_{1j}x_i)}{1 + \sum_{j=1}^{k-1}\exp(\beta_{0j} + \beta_{1j}x_i)}, \quad j = 1, 2, \ldots, (k-1)$$

Theory of Statistics

We define regression model($\mathbb{E}(Y|X)$) as follows.

1. Data (Y_i, X_i) are mutually indep.

$$\{(Y_i, X_i) : Y_i \in \mathbb{R}, X_i \in \mathbb{R}^p, i = 1, \cdots, n\}$$

2. Purpose: inference of conditional probability distribution of $f(y|x)$
3. (Model 1) $Y_i|X_i \sim f(y|x), i = 1, \cdots, n,$

$$\mathcal{P} = \{f(y|x) : y \in \mathbb{R}, x \in \mathbb{R}^p\}$$

We define regression model($\mathbb{E}(Y|X)$) as follows.

1. Data (Y_i, X_i) are mutually indep.

$$\{(Y_i, X_i) : Y_i \in \mathbb{R}, X_i \in \mathbb{R}^p, i = 1, \cdots, n\}$$

2. Purpose: inference of conditional probability distribution of $f(y|x)$
3. (Model 2) $Y_i = g(X_i) + \epsilon_i, \epsilon_i \overset{\text{iid}}{\sim} F, g : \mathbb{R}^p \to \mathbb{R}, F$: symmetric at 0.

$$\mathcal{P} = \{g : \mathbb{R}^p \to \mathbb{R}\} \times \{F : \text{ symmetric at } 0\}$$

$$\mathcal{P} = \{g : \mathbb{R}^p \to \mathbb{R}\} \times \{F = N(0, \sigma^2) : \sigma > 0\}$$

4. (Model 3; linear model)
$Y_i = X_i^\top \beta + \epsilon_i, \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \beta \in \mathbb{R}^p, \sigma^2 > 0, i = 1, \cdots, n,$

$$\Theta := \{(\beta, \sigma^2) : \beta \in \mathbb{R}^p, \sigma > 0\} \quad \mathcal{P} = \{f(y|x) : y \in \mathbb{R}, x \in \mathbb{R}^p\}$$

Decision Theory

1. Let $\mathcal{P} := (\mathcal{X}, \mathcal{Y}, (P_\theta)_{\theta \in \Theta})$ be a model where (P_θ) is a collection of distributions.
2. Let $\mathcal{A}$ be an *Action space* if it is a collection of all possible and available actions. ex) In Hypothesis Testing, $\mathcal{A} = \{H_0, H_1\}$.
3. Loss Function.

$$L : \Theta \times \mathcal{A} \to \mathbb{R}_+$$

The loss that we get when we choose action $a \in \mathcal{A}$ and the truth value is $\theta \in \Theta$.
Ex) (Estimation) $L(\theta - a) = (\theta - a)^2 = \|\theta - a\|^2$ (quadratic loss)

Decision Rule/Decision Procedure

1. Let δ be a decision rule if the action mapped to observation x.

$$\delta : x \to \mathcal{A}$$

Ex) (Estimator) $\mathbf{X} = (x_1, \cdots, x_n) \mapsto \bar{X}$
Ex) (Test) $\mathbf{X} \mapsto \phi(\mathbf{x}) = I(\mathbf{x} \in C)$ (rejecting probability), where C is a rejection region.

Then $\delta(x) = \begin{cases} H_0, & \frac{\sqrt{n}}{\sigma}(\bar{X} - \theta_0) > z_{\alpha/2}, \\ H_1, & o.w. \end{cases}$

2. Let $R(\theta, \delta)$ be a Risk function if

$$R(\theta, \delta) = \mathbb{E}_\theta L(\theta, \delta(\mathbf{x}))$$

ex) (estimation) $L(\theta, \delta) = (\theta - \delta)^2$,
$R(\theta, \delta) = \mathbb{E}_\theta(\theta - \delta(x))^2 = Var_\theta \delta(x) + (\mathbb{E}_\theta \delta(x) - \theta)^2$

Prediction

1. Best predictor under squared error.
2. We want to find the prediction function $g(x)$ to minimize MSPE $= \mathbb{E}(Y - g(X))^2$
3. In other words, we want to find $g^* = \arg\min_{g \in \mathcal{G}} \mathbb{E}(Y - g(X))^2$, where $\mathcal{G}$: all function space.
   Note that
   1. To make it easier to find the predictor, we constrain $\mathcal{G}$ as linear function.
   2. Instead of squared error, we use other distance measure.

Prediction

1. Optimization of Prediction

Let $\mathbb{E}Y < \infty$

- $\mathbb{E}Y = \arg\min_c \mathbb{E}(Y - c)^2$
- $\mathrm{Cov}(Y - \mathbb{E}(Y|X), g(X)) = 0, {}^{\forall} g(x) \in L_2$
- $\mathbb{E}(Y|X) = \arg\min_{g(X) \in L_2} \mathbb{E}(Y - g(X))^2$

2. Best Linear Predictor

Let $\mathbb{E}Y < \infty$, $Var(X)$, $Var(X)^{-1}$ exist.

- For $X \in \mathbb{R}$,

$$b_0 X = \underset{bX, b \in \mathbb{R}}{\arg\min}\, \mathbb{E}(Y - bX)^2, \quad b_0 = \frac{\mathbb{E}XY}{\mathbb{E}X^2}$$

- For $X \in \mathbb{R}$,

$$a_1 + b_1 X = \underset{a+bX, a, b \in \mathbb{R}}{\arg\min}\, \mathbb{E}(Y - (a + bX))^2,$$

$$b_1 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}, a_1 = \mathbb{E}Y - b_1 \mathbb{E}X$$

Prediction

2. Best Linear Predictor

Let $\mathbb{E}Y < \infty$, $Var(X)$, $Var(X)^{-1}$ exist.

- For $X \in \mathbb{R}^k$,

$$\mu_L(X) = \underset{\mu(X)=\mu_Y+(X-\mu_X)^\top\beta}{\arg\min} \mathbb{E}(Y-\mu(X))^2,$$

$$\mu_L(X) = \mu_Y + (X-\mu_X)^\top\beta, \beta = \Sigma_{XX}^{-1}\sigma_{XY}$$

3. Multiple Correlation Coefficient

$$\rho_{XY} = \text{Corr}(Y, \mu_L(X)) = \sqrt{\frac{\sigma_{YX}\Sigma_{XX}^{-1}\sigma_{XY}}{\sigma_Y^2}}$$

$$R^2 = \rho_{XY}^2 = \frac{Var(\mathbb{E}[Y|X])}{Var(Y)}$$

Example (Multivariate Normal Distribution)

Let $Y \in \mathbb{R}, X \in \mathbb{R}^p$ and

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \sigma_{YX} \\ \sigma_{XY} & \sigma_X^2 \end{pmatrix} \right)$$

Then, $[Y|X = x] = N(\mu_Y + \sigma_{YX}\Sigma_{XX}^{-1}(x - \mu_X), \sigma_Y^2 - \sigma_{YX}\Sigma_{XX}^{-1}\sigma_{XY})$

- Note that

$$\mathbb{E}[Y|X = x] = \mu_Y + (x - \mu_X)^\top \Sigma_{XX}^{-1}\sigma_{XY} = \mu_Y + (x - \mu_X)^\top \beta$$

- $\beta = \Sigma_{XX}^{-1}\sigma_{XY}$: Regression coefficient.
- Minimum of MSE:

$$\mathbb{E}((Y - \mathbb{E}(Y|X))^2) = \mathbb{E}(Var(Y|X)) = \sigma_Y^2 - \sigma_{YX}\Sigma_{XX}^{-1}\sigma_{XY}$$

- Multiple Corr. Coeffi. or Coeffi. of Determination:

$$R^2 = \frac{Var(\mathbb{E}[Y|X])}{Var(Y)} = \frac{\sigma_{YX}\Sigma_{XX}^{-1}\sigma_{XY}}{\sigma_Y^2}$$

Applications

Times Series Data Analysis

- Time Series is a collection of observations recorded sequentially in time.
- Each observation is a r.v., which can be either univariate (i.e., scalar, int, float) or multivariate (i.e., vector).
- The primary objective of time-series modeling is to develop mathematical models that provide plausible descriptions for the observed data.
  - Past observations often contain useful information about future values.
  - Time-series models explicitly represent temporal dependence.
- There are two representative classical time-series models: AR and MA.
  - Modern AI models do not explicitly use AR or MA equations.
  - However, ARMA models are still useful because they are simple, fast, and interpretable.

Autoregressive (AR) Model

- **Autoregressive (AR)** model assumes that the current x_t can be explained as a function of p past values.
  - p is the number of steps needed to forecast the current value.
  - p is sometimes referred to as the *window size*.
- **AR(p)** models the current value as a linear function of past values:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3} + \cdots + \phi_p x_{t-p} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

  - The model is called *auto* because it relies on its own data.
  - The model is called *regression* because it predicts the value directly.
  - We assume ϵ_t is zero-mean i.i.d. normal RV (i.e., white noise).
- AR is useful when the present tends to continue patterns from the recent past.
- AR model is inherently *generative* and can recursively predict future values.

Example: AR(1) Model

- **AR(1)** model is the simplest:

$$x_t = \phi x_{t-1} + \epsilon_t$$

  - AR is a *recursive* structure.
  - $\phi = 0$: pure noise (unpredictable)
  - $0 < \phi < 1$: decaying memory
  - $\phi \approx 1$: very strong persistence

- Intuitively, the current value is partially explained by previous value(s), together with unexpected shocks (innovations).

Moving Average (MA) Model

- **Moving Average (MA)** model assumes that the current value depends on the current and previous q noise (shock) terms.
- **MA(q)** models the current value as a linear function of recent shocks:

$$x_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

  - Short-term memory: only the most recent q noise terms matter.
  - In other words, the effect of the noise lasts for q timesteps.
  - There is no direct dependence on past observed values $(x_{t-1}, x_{t-2}, \dots)$.

Example: MA(1) Model

- **MA(1)** model is the simplest:

$$x_t = \epsilon_t + \theta \epsilon_{t-1}$$

  - MA is a *finite-memory* structure.
  - $\theta = 0$: pure noise
  - Small $|\theta|$: weak carryover of the previous shock.
  - Large $|\theta|$: strong short-term shock propagation.

ARMA Model

- AR and MA capture different kinds of temporal structure.
  - AR: dependence on past observed values
  - MA: dependence on recent noise terms
- AR is good for smooth continuation (i.e., inertia).
- MA is good for short-lived transient effects.

- **ARMA(p,q)** describes many real-world sequences that show both long persistence and short-lived effects, and is more flexible than AR or MA alone.

$$x_t = \underbrace{\sum_{k=1}^{p} \phi_k x_{t-k}}_{AR(p)} + \epsilon_t + \underbrace{\sum_{k=1}^{q} \theta_k \epsilon_{t-k}}_{MA(q)}$$

Order Selection in ARMA

- Choosing appropriate orders p and q is important in ARMA modeling.
- Effect of model order:
  - Larger p: looks further back into past values.
  - Larger q: retains the effects of shocks for more timesteps.
  - Smaller p, q: simpler but may miss important temporal structure.
- Model complexity trade-off:
  - **Underfitting:** p and q are too small - model cannot explain the data well.
  - **Overfitting:** p and q are too large - model fits noise rather than the actual trend.

ARIMA Model

- ARMA usually works best when **X** is approximately stationary.
  - Stationary sequence: constant mean and variance over time.
  - However, many real-world sequences are not stationary.

- **ARIMA** model instead applies ARMA to the differenced sequence (Δ).

$$y_t = \Delta x_t = x_t - x_{t-1}$$

  - ARIMA = Autoregressive + Integrated (differencing) + Moving Average
  - The "integrated" part refers to reversing differencing after prediction.
  - After modeling the differenced sequence, we can reconstruct the original sequence by cumulative summation, given an initial value x_0.

$$x_t = x_0 + \sum_{k=1}^{t} y_k$$

Conclusion

- AR(I)MA model is useful but we need some useful techniques such as filtering, smoothing and so on.
- We can use Deep Learning Model (ex: CNN, RNN) with using time series.
- These techniques are useful when analyzing image, voice, etc.

Machine Learning

- In Machine Learning, regression model is used for supervised tasks.
- Supervised tasks are data situations where learning the functional relationship between inputs (features) and output (target) is useful.
- The two most basic tasks are regression and classification, depending on whether the target is numerical or categorical.

Machine Learning: Classification Basics

- In regression, the target variable y is continuous ($y \in \mathbb{R}$).
- In **classification**, the target variable y is discrete.
  - Binary classification: $y \in \{0, 1\}$ or $\{-1, 1\}$
  - Multiclass classification: $y \in \{1, 2, \ldots, K\}$

- **Goal**: Given a feature vector $\mathbf{x} \in \mathbb{R}^D$, predict the class label y.
- We want to find a mapping function (or decision rule) $f : \mathbb{R}^D \to \{1, \ldots, K\}$.

- From a probabilistic perspective, we want to estimate the conditional probability $P(y \mid \mathbf{x})$.

Two Approaches to Classification

- **Discriminative Models**
  - Directly estimate the conditional probability $P(y \mid \mathbf{x})$.
  - Learn the decision boundary between classes.
  - *Examples*: Logistic Regression, Support Vector Machines (SVM), Neural Networks.

- **Generative Models**
  - Estimate the class-conditional density $P(\mathbf{x} \mid y)$ and the prior $P(y)$.
  - Use Bayes' rule to compute the posterior $P(y \mid \mathbf{x})$:

$$P(y \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid y)P(y)}{P(\mathbf{x})}$$

  - Can generate new data points from the learned distribution.
  - *Examples*: Naive Bayes, Gaussian Discriminant Analysis (GDA).

Logistic Regression in ML: Cross-Entropy Loss

- Logistic regression is a **discriminative** model.
- It models the probability of class 1 as:

$$\hat{y} = P(y = 1 \mid \mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^\top \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x})}$$

- In ML, we estimate parameters $\mathbf{w}$ by minimizing a **Loss Function**.
- **Negative Log-Likelihood (NLL) / Cross-Entropy Loss**:

$$\mathcal{L}(\mathbf{w}) = -\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

  - If $y_i = 1$, we want $\hat{y}_i \approx 1 \implies$ loss ≈ 0.
  - If $y_i = 0$, we want $\hat{y}_i \approx 0 \implies$ loss ≈ 0.

Optimization: Gradient Descent

- Unlike Linear Regression, Logistic Regression has no closed-form solution for $\mathbf{w}$ (due to the non-linear sigmoid function).
- We must use iterative optimization algorithms like **Gradient Descent**.
- **Update Rule**:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla \mathcal{L}(\mathbf{w}^{(t)})$$

  - $\eta > 0$ is the **learning rate** (step size).
  - $\nabla \mathcal{L}(\mathbf{w})$ is the gradient of the loss function with respect to $\mathbf{w}$.

- The gradient for Logistic Regression with Cross-Entropy loss is elegantly simple:

$$\nabla \mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} (\hat{y}_i - y_i)\mathbf{x}_i$$

References

- 서병태. *회귀분석입문* [Lecture notes]. Sungkyunkwan University (Spring 2025).
- 김찬민. *Nonparametric Statistics* [Lecture notes]. Sungkyunkwan University (Fall 2025).
- 김재직. *Statistical Simulation* [Lecture notes]. Sungkyunkwan University (Spring 2025).
- 이재용, 이기재. *베이즈 데이터 분석*. 방송통신대학교출판문화원 (2022).
- 이재용. *통계이론1* [Lecture notes]. Seoul National University (Spring 2026).
- 심규홍. *시계열데이터처리개론* [Lecture notes]. Sungkyunkwan University (Spring 2026).
- Tamer Abuhmed. *Fundamentals of Machine Learning* [Lecture notes]. Sungkyunkwan University (Spring 2026).

Thank you!