Ch0. Preliminaries

1. Probability models
2. Multivariate Normal distribution
3. Regression and diagnostics

Basic statistics

- Population: target of interest. Hence, knowing (describing) the population would be the ultimate goal of study. How can we describe the population? First, you need to understand that there is a randomness in the population. In statistics, such randomness is characterized by probability. That is, before it is realized, we cannot know the outcome of a unit (subject).
- Probability explains randomness in terms of numbers between 0 to 1. Two ways of interpreting (defining) probability
  - Long-run relative frequency: See the proportion of time observing event A out of many repetitions

$$P(A) = \lim_{n \to \infty} \frac{num(A)}{n}.$$

  - Axioms of probability Probability is a non-negative, countably additive and add up to 1 set function onto $[0, 1]$.
- Then, what is the role of the statistics?

Basic statistics

- In principle, we need to know the whole sample to calculate the probability. But, it takes too much time and money to examine the population, so it is not a feasible method. Thus, in practice, we estimate probability through only a part of population called sample.
- Statistical inference: guess how a population looks like from the sample data.
- Naive estimation procedure is to calculate the relative frequency out of N sample data to estimate probability.

$$\widehat{P}(A) = \frac{num(A)}{N} \approx P(A).$$

This is called the empirical estimation of probability and the baseline of method of moment estimation.

Basic statistics

- As you well know, however, estimating probability by relative frequency has a major practical problem: What if the sample space is continuous (uncountably many possible values)?
  - May need more sample than discrete case, but even for sufficiently large number of sample, bandwidth (bin) selection is critical.
- Histogram is the graphical way of representing relative frequency of data. For continuous data, we count the number of data fall into (predetermined) class interval. Here is an example of it.

Call center service time

- The service time in call centers is the time that an agent spends to serve individual.
- As you know, it costs a lot of money to maintain call center, hence it is critical to know how this service time looks like.
- See the movie and describe its feature.

- The bottom line is that it is very hard to determine the number of bins in practice!

Parametric modelling

- We have seen that estimating probability from the sample data using relative frequency is not easy in practice. This is called the non-parametric method of estimating probability.
- Statisticians can do better by assuming certain structure on the probability distribution. This is called the parametric modeling of probability distribution.
- For example, we can assume Binomial distribution, Poisson distribution, Normal distribution, Exponential distribution, etc for underlying distribution.

Random Variable

- Axioms of probability $(\Omega, \mathcal{F}, \mathcal{P})$: $P(A)$ is a set function onto $[0, 1]$ satisfying

$$i) P(A) \geq 0$$
$$ii) P(\Omega) = 1$$
$$iii) P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i), \quad A_i\text{'s are disjoint}$$

- Random variable X is a (measurable) function from a sample space Ω to $\mathbb{R}$ (real line). That is, assign some number for each elementary outcome in Ω.
- Induced probability by random variable X is called the probability distribution,

$$F_X(x) := P_X((-\infty, x]) = \mathcal{P}\left(X^{-1}(-\infty, x]\right)$$

Properties of distribution function

- Distribution uniquely determines the random variable X. If $F_X(x) = F_Y(y)$ for all $x \in \mathbb{R}$, then $X \stackrel{d}{=} Y$.
- F is non-decreasing, $F(X) \leq F(y)$ if $x \leq y$.
- F is right-continuous $F(y) \downarrow F(x)$ as $y \downarrow x$.
- $F(x) \to 1$ as $x \to \infty$, $F(y) \to 0$ as $y \to -\infty$.
- If random variable X has a density function, then $f(x) = F'(x)$.
- Expectation is given by

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) dF(x)$$

Extension to n-variables

Consider we are performing n experiments at the same time and want to assign probability on it.

- First embedding $\Omega_1, \ldots, \Omega_n$ into a single large sample space Ω

  $$\Omega = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_n = \{(\omega_1, \ldots, \omega_n) | \omega_1 \in \Omega_1, \ldots, \omega_n \in \Omega_n\}$$

  and assign probability.
- For random variables, we can consider a column vector of random variable $\mathbf{X} = (X_1, \ldots, X_n)'$ defined on $\mathbb{R}^n$ and induced probability distribution

  $$P_{\mathbf{X}}(\mathbf{x}) := F_{\mathbf{X}}(x_1, \ldots, x_n) = P_X(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

  $$= \mathcal{P}(\mathbf{X}^{-1}((-\infty, x_1] \times (-\infty, x_2] \ldots \times (-\infty, x_n]))$$

Mean and Covariance

- Mean and covariance of $\mathbf{X}$ is given by

$$\mu_{\mathbf{X}} = (EX_1, \ldots, EX_n)'$$

$$\Sigma_{XX} = \mathrm{Cov}(\mathbf{X}, \mathbf{X}) = E(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))'$$

$$= \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \ldots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \ldots & \mathrm{Cov}(X_n, X_n) \end{pmatrix}$$

- Σ_{XX} is a covariance matrix of $\mathbf{X}$ and it is symmetric and non-negative definite

$$a'\Sigma_{XX}a \geq 0 \quad \forall a = (a_1, \ldots, a_n)'$$

- Since symmetric matrix is always diagonalizable,

$$\Sigma_{XX} = P\Lambda P', \quad PP' = P'P = I$$

Multivariate Normal distribution

Definition (MVN)

A random vector $\mathbf{X} = (X_1, \ldots, X_n)'$ *follows a multivariate normal distribution with mean* $\mu = (\mu_1, \ldots, \mu_n)'$ *and covariance matrix* $\Sigma = Cov(\mathbf{X}, \mathbf{X})$ *if one of the following holds:*

$i)$ $f(x) = |2\pi\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)$

$ii)$ $M_X(t) := E(e^{t'X}) = E(e^{t_1X_1+\ldots t_nX_n}) = \exp\left(t'\mu + \frac{1}{2}t'\Sigma t\right)$

$iii)$ *For any* $n \times 1$ *vector* $\mathbf{a}$,

$\mathbf{a}'X$ *follows a univariate normal distribution.*

Least squares estimation - OLS

- Simple linear regression model:

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \sim i.i.d.\ \mathcal{N}(0, \sigma^2)$$

  OLS (ordinary least squares) minimizes

$$S(\theta_0, \theta_1) = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

- We find θ_0 and θ_1 from normal equation:

$$\frac{\partial S(\theta_0, \theta_1)}{\partial \theta_0} =$$
$$\frac{\partial S(\theta_0, \theta_1)}{\partial \theta_1} =$$

Least squares estimation - OLS

- Estimation of σ^2: Observe that

$$\sigma^2 = E(\epsilon_i^2) \approx \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2$$

from the method of moment estimation. However, ϵ_i^2 is not observable, so need to be estimated.

$$\hat{\epsilon}_i = y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i$$

$$\hat{\sigma}^2 \approx \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 =: \frac{SSE}{n}.$$

It is known that it is better to divide by $(n-2)$ (why?)

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

- It implies that

$$\hat{y}_i | x_i \approx \mathcal{N}(\hat{\theta}_0 + \hat{\theta}_1 x_i, \hat{\sigma}^2)$$

Least squares estimation - OLS using matrix

Just stack observations in a column gives

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \theta_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \theta_1 \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$= \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

In a matrix notation, hence we have

$$\boxed{\mathbf{Y} = \mathbf{X}\theta + \boldsymbol{\epsilon}}$$

Least squares estimation - OLS using matrix

- $$\hat{\theta}^{OLS} = \underset{\theta}{\operatorname{argmin}}(\mathbf{Y} - \mathbf{X}\theta)'(\mathbf{Y} - \mathbf{X}\theta)$$

  $$= \underset{\theta}{\operatorname{argmin}}(\mathbf{Y}'\mathbf{Y} - \theta'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\theta + \theta'\mathbf{X}'\mathbf{X}\theta)$$

  Solving

  $$\frac{\partial}{\partial\theta}(\mathbf{Y} - \mathbf{X}\theta)'(\mathbf{Y} - \mathbf{X}\theta) = 0$$

  gives

  $$\hat{\theta}^{OLS} =$$

- Geometrically, it is a projection

  $$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$$

  $$\mathbf{X} \perp (\mathbf{Y} - \hat{\mathbf{Y}}) \iff \mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\theta}) = 0$$

Least squares estimation - OLS properties

- $E(\hat{\theta}^{OLS}) = \theta$
- $\text{Var}(\hat{\theta}^{OLS}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
- Gauss-Markov Theorem: $\hat{\theta}$ is Best Linear Unbiased Estimator (BLUE). Amongst all linear unbiased estimators, $\hat{\theta}$ has the smallest variance (hence the smallest MSE amongst all linear unbiased estimators)

$$\text{Var}(\theta^*) \geq \text{Var}(\hat{\theta})$$

for any θ^* such that $E(\theta^*) = \theta$.

- Sum of squares decomposition

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$
$$SST = SSE + SSR$$

(Why subtract $\bar{y}$ for SST?)

Least squares estimation - OLS properties

- $SSE/\sigma^2 \sim \chi^2(n-p-1)$. Thus, $E(SSE/\sigma^2) = n-p-1$ implies that

$$\hat{\sigma}^2 := \frac{SSE}{n-p-1}$$

  is an unbiased estimator (hence better than divide by n).
- $SSR/\sigma^2 \sim \chi^2(p)$.
- $MSR/MSE := (SSR/p)/(SSE/(n-p-1) \sim F(p, n-p-1)$.
- $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
- For testing whether θ_i is necessary in the regression model

$$H_0 : \theta_i = 0 \quad \text{vs} \quad H_1 : \theta_i \neq 0$$

  use

$$\frac{\hat{\theta}_i - 0}{\hat{\sigma}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{ii}}} \sim t(n-p-1)$$

Regression: diagnostics

- F-test to see whether regression variables are needed, that is test for $H_0 : \theta_1 = \ldots = \theta_p = 0$
- R^2 for the goodness-of-fit.
- t-test for each regression variable; $H_0 : \theta_i = 0$.
- Check residuals. Intuitively, if $\hat{\theta}_i \approx \theta_i$, then

$$e_i := y_i - \hat{\theta}_0 - \ldots - \hat{\theta}_p x_{pi} \approx y_i - \theta_0 - \ldots - \theta_p x_{pi} = \epsilon_i$$

This can be checked from the normal equation. Geometrically, it means that the best solution is the orthogonal projection, hence

$$e_i \perp \hat{y}_i \quad e_i \perp x_i$$

For example, if there are any trend, we need to consider adding more explanatory variable.

Regression: diagnostics

- Since we have assumed that $\epsilon_i \sim i.i.d.\ \mathcal{N}(0, \sigma^2)$, residuals e_i should look $i.i.d.$ normal distribution with mean 0 and variance $\hat{\sigma}^2$.
- We will check three things i) independence between residuals, ii) identically distributed (e.g. constant mean and homogeneous variance) and iii) normally distributed.
- Plot (studentized) residuals over index i, then see whether they are i.i.d. That is, it should randomly scattered with zero mean and unit variance.

Regression: diagnostics

- Normality can be checked by drawing QQ-plot. QQ plots are a way to evaluate whether a sample of observations can be modeled by a family of distributions. It compares the empirical quantiles with theoretical quantiles of a target distribution of interest.
  1. Sort data $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.
  2. Then $(i - .5)/n^{th}$ sample quantile is $x_{(i)}$.
  3. The corresponding population quantile is $F^{-1}((i - .5)/n)$
  4. The QQ plot is obtained by plotting

$$\left(F^{-1}\left(\frac{i - .5}{n} \right), x_{(i)} \right)$$

  5. If the sample can by modeled by that distribution, it will form a straight line.
- You can apply other methods learned in regression analysis, for example, checking leverage, outliers, multicollinearity etc.