Basic Stats

Bill Last Updated:

26 March, 2020

Contents

Preface: Motivation **5**

1 443 **7**
1.1 Some basic concepts 7
1.2 Discrete Random Variables 8
1.3 Continuous Random Variables 8
1.4 Large Sample Theory 9

2 556 **13**
2.1 Statistics and Sampling Distributions 13
2.2 Point Estimation 16
2.3 Sufficient and completeness 28

3 530_533 **37**
3.1 Definition of the general linear model 37
3.2 Simple Linear Regression 38
3.3 Full rank, less than full rank 42
3.4 Assumptions, checking assumptions 44
3.5 Bootstrapping used in linear models 45
3.6 Generalized linear models 47

4 512 **49**
4.1 Basics 49
4.2 Completely randomized designs 49
4.3 Randomized complete block designs 50
4.4 Incomplete block designs 50
4.5 Row-column design (Latin square designs) 51
4.6 Split-plot designs 51

5 Logit and Probit **55**
5.1 Logit 55
5.2 Probit 57

6 Normal distribution **61**

6.1 Basics . . . 61
6.2 Confidence intervals for normal distributions . . . 61
6.3 Percentile . . . 62

7 MLE **65**
7.1 Basic idea of MLE . . . 65
7.2 Coin flip example, probit, and logit . . . 66
7.3 Further on logit . . . 67
7.4 References . . . 69

8 Score, Gradient and Jacobian **71**
8.1 Score . . . 71
8.2 Fisher scoring . . . 72
8.3 Gradient and Jacobian . . . 72
8.4 Hessian and Fisher Information . . . 73

9 Canonical link function **75**

10 Ordinary Least Squares (OLS) **77**
10.1 Taylor series . . . 79
10.2 References . . . 79

11 Cholesky decomposition **81**
11.1 Example 1 . . . 81
11.2 Example 2 . . . 82
11.3 Example 3 . . . 83

Preface: Motivation

All the notes I have done here are about basic stats. While I have tried my best, probably there are still some typos and errors. Please feel free to let me know in case you find one. Thank you!

Chapter 1

443

1.1 Some basic concepts

1.1.1 Random variable

Any numerical realization of a random experiment results in a random variable.

In layman terms, it is a number that has a certain probability attached to it.

1.1.2 Permutation

An ordered arrangement of a set of objects is known as a permutation.

e.g., The number of permutations of n distinguishable objects is $n!$. e.g., The number of permuations of n distinct objects taken r at a time is

$$ {}_nP_r = \frac{n!}{(n-r)!} $$

1.1.3 Combinations

If the order of the objects is not important, then one may simply be interested in the number of combinations.

$$ \binom{n}{r} = \frac{n!}{r!(n-r)!} $$

1.1.4 Partitioning

The number of ways of partitioning a set of n objects into k cells with r_1 objects into the first cell, r_2 in the second cell, and so forth is

$$\frac{n!}{r_1!r_2!...r_k!}$$

1.2 Discrete Random Variables

1.2.1 Binomial

$$X \sim BIN(n, p)$$

$$\binom{n}{x} P^x (1-P)^{n-x}$$

mean: np

variance: npq

(Note that, Bernoulli is written as $BIN(1, p)$)

1.2.2 Poisson

$$X \sim POI(\mu)$$

$$\frac{e^{-\mu}\mu^x}{x!}$$

mean: μ

variance: μ

1.3 Continuous Random Variables

1.3.1 Uniform

$$X \sim UNIF(a, b)$$

$$\frac{1}{b-a}$$

Mean: $\frac{a+b}{2}$

Variance: $\frac{(b-a)^2}{12}$

1.3.2 Exponential

$$X \sim EXP(\theta)$$

$$\frac{1}{\theta} e^{-x/\theta}$$

Mean: θ

Variance: θ^2

1.3.3 Normal

$$X \sim N(\mu, \sigma^2)$$

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Mean: μ

Variance" σ^2

1.4 Large Sample Theory

1.4.1 Convergence in distribution

https://en.wikipedia.org/wiki/Law_of_large_numbers

$$\bar{X} \to \mu \quad (n \to \infty)$$

$$Var(\bar{X}) = Var(\frac{1}{n}(X_1 + ... + X_n)) = \frac{1}{n^2} Var(X_1 + ... + X_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

1.4.2 Weak law

There are two different versions of the Law of Large Numbers: Strong law of large numbers and Weak law of large numbers.

The weak law of large numbers: The sample average converges **in probability** towards the expected value.

$$\bar{X_n} \xrightarrow{p} \mu \quad (n \to \infty)$$

This, for any positive number ϵ

$$lim_{n\to\infty} Pr(|\bar{X_n} - \mu| > \epsilon) = 0$$

1.4.3 Strong law

The sequence of sample mean $\bar{X}_n$ converges to μ **almost surely**.

$$\bar{X_n} \xrightarrow{a.s.} \mu \quad (n \to \infty)$$

This is,

$$Pr(lim_{n\to\infty}\bar{X_n} = \mu) = 1$$

1.4.4 Central limit theorem

If $X_1, ..., X_n$ is a random sample from a distribution with mean μ and variance $\sigma^2 < \infty$, then the limiting distribution of

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$$

is the standard normal, $Z_n \xrightarrow{d} Z \sim N(0,1)$ as $n \to \infty$.

1.4.4.1 Bernoulli law of large number

$\hat{p_n}$ converges stochasticalltly to p as n approchaes infinity. For example, if a coin is tossed repeatedly, and $A = \{H\}$, then the successive relative frequencies of A correspond to a sequence of random variables that will converge stochastically to $p = 1/2$.

1.4.4.2 Normal approximation to Binomial

$$Z_n = \frac{Y_n - np}{\sqrt{npq}} \xrightarrow{d} Z \sim N(0,1)$$

Example: The probability that a basketball player hits a shot is $p = 0.5$. If he takes 20 shorts, what is the probability that he hits at least 9?

$$\begin{aligned} P[Y_{20} \geq 9] &= 1 - P[Y_{20} < 8] \\ &= 1 - \sum_{y-0}^{8} \binom{20}{y} 0.5^y 0.5^{20-y} \\ &= 0.7483 \end{aligned}$$

A normal approximation is

$$\begin{aligned} P[Y_{20} \geq 9] &= 1 - P[Y_{20} < 8] \\ &= 1 - \Phi(\frac{8-10}{\sqrt{5}}) \\ &= 0.8133 \end{aligned}$$

1.4.4.3 Normal approximation to Poisson

$$\begin{aligned} P[10 \leq Y_{20} \leq 30] &= P[Y_{20} \leq 30] - P[Y_{20} \leq 10] \\ &= \Phi[\frac{30.5-20}{\sqrt{20}}] - \Phi[\frac{9.5-20}{\sqrt{20}}] \\ &= 0.981 \end{aligned}$$

1.4.5 Poisson approximation to binomial

We know that the mean for binomial is

$$\mu = np \rightarrow p = \frac{\mu}{n}$$

The moment generating function for Binomial is

$$M_n(t) = (1 - p + pe^t)^n = (1 + \frac{\mu(e^t - 1)}{n})^n$$

$$lim_{n\rightarrow\infty} M_n(t) = e^{\mu(e^t - 1)}$$

Note that the MGF for Poisson is as follows.

$$POI(\lambda) : e^{\lambda(e^t - 1)}$$

Thus,

$$Y_n \rightarrow Y \sim POI(\mu)$$

Chapter 2

556

2.1 Statistics and Sampling Distributions

2.1.1 Statistics

2.1.1.1 Definition of Statistic

P.264

A fucntion of observable random variables, $T = t(X_1, ...X_n)$, which does not depend on any unknown parameters is called statistic.

For example, let $X_1, ..., X_n$ represent a random sample from a population with *pdf* $f(x)$. The sample mean provides an example of a statistic with the function

$$t(x_1, ..., x_n) = (x_1 + ... + x_n)/n$$

This statistic usually is denoted by

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

When a random sample is observed, the value of $\bar{X}$, computed from the data, usually is denoted by lower case $\bar{x}$.

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

2.1.1.2 Sample and parameters

P.265

If $X_1, ..., X_n$ denotes a random sample from $f(x)$ with $E(X) = \mu$ and $var(X) = \sigma^2$, then

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

Example:

A random sample of size n from a Bernoulli distribution $X_i \sim BIN(1, p)$. We know Bernoulli has $\mu = p$ and $\sigma^2 = pq$. In this case, the sample mean is

$$\bar{X} = Y/n = \hat{p}$$

Thus,

$$E(\hat{p}) = p$$

$$Var(\hat{p}) = \frac{pq}{n}$$

Thus, sample mean is the unbiased estimate for the population mean. However, the variance of the mean is not equal to population variance. That lead to definition of sample variance.

P.266

Sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$E(S^2) = \sigma^2$$

(Question: does it mean that sample varaince is an unbiased estimator of population variance?)

2.1.2 $\chi^2, t, F, beta$

2.1.2.1 χ^2

P.271

If $X_1, ..., X_n$ denotes a random sample from $N(\mu, \sigma^2)$, then

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim x^2(n)$$

$$\frac{n(\bar{X} - \mu)^2}{\sigma^2} \sim x^2(1)$$

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$$

(Thus, we can see this is a bit weired, as the numerator is $\bar{X}$ is from the sample, whereas σ^2 is from the population. Thus, assume we know σ^2?)

Thus, we can

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

(You can compare $\bar{X}$ with μ, we can see the only difference is that the χ^2 has one less degree of freedom because we use this degree of freedom to calculate the mean.)

For the mean and variance of χ^2:

Assume that

$$X \sim x^2(v)$$

$$mean : v$$

$$variance : 2v$$

2.1.2.2 t

Definition

$$t(k) = \frac{N(0,1)}{\sqrt{\frac{\chi^2(k)}{k}}}$$

Property 1

t distribution is symmetrical.

Given that t distribution is symmetrical, we can get

$$H(-c) = 1 - H(c)$$

Property 2

t distribution has heavier tails than the normal.

My note: t distribution only has a parameter of k, which is determined by the χ^2's degree of freedom. Of course, χ^2 also only has one parameter, namely the degree of freedom.

2.1.2.3 F

If $V_1 \sim \chi^2(v_1)$ and $V_2 \sim \chi^2(v_2)$ are independent, then the random variable

$$\frac{V_1/v_1}{V_2/v_2} \sim F(v_1, v_2)$$

2.1.2.4 Beta

If $X \sim F(v_1, v_2)$

$$Y = \frac{(v_1/v_2)X}{1 + (v_1/v_2)X} \sim Beta(\alpha, \beta)$$

2.1.3 Large-sample approximations

P.280

If $Y_v \sim x^2(x)$, then

$$Z_v = \frac{Y_v - v}{\sqrt{2v}} \xrightarrow{d} Z \sim N(0, 1)$$

(The proof is based on CLT. In addition, as disscused above, we can get chi-square from normal distributions.)

2.2 Point Estimation

2.2.1 Method of moments estimators

2.2.1.1 Definition about moments (chapter 2)

P.73

THe **kth moment about the origin** of a random variable X is

$$\mu_k^{'} = E(X^k)$$

and the **kth moment about the mean** is

$$\mu_k = E[X - E(X)]^k = E(X - \mu)^k$$

Thus,$k = E(X^k)$ may be considered as the kth moment of X or the first moment of X^k.

The first moment about the mean is zero,

$$\mu_1 = E[X - E(X)] = E(X) - E(X) = 0$$

The second moment about the mean is the variance,

$$\mu_2 = E[(X - \mu)^2] = \sigma^2$$

Note that the definition of variance:

P.73

$$Var(X) = E[(X - \mu)^2]$$

(**Note: Based on the information above, it seems important to understand the difference between the moment of the origin and the moment of the mean.**)

2.2.1.2 Definition

Based on the last chapater (i.e., Chapter 8), sample mean $\bar{X}$ is an estimator of the population mean μ. A more general approach, which produced estimators known as the **method of moments estimators(MMEs)** , can be developed.

If $X_1, ..., X_n$ is a random sample from $f(x; \theta_1, ..., \theta_k)$, the first k sample moments are given by

$$M_j^{'} = \frac{\sum_{i=1}^{n} X_i^j}{n}$$

where,

$$j = 1, 2, ...k$$

Example 1:

P.291

Consider a random sample from a distribution with two unknown parameters, the mean μ and the variance σ^2. We know from earlier considerations that $\mu = \mu_1^{'}$ and $\sigma^2 = E(X^2) - \mu^2 = \mu_2^{'} - (\mu_1^{'})^2$.

Thus,

$$\hat{\sigma}^2 = \mu_2^{'} - (\mu_1^{'})^2 = \frac{\sum_{i=1}^{n} X_i^2}{n} - \bar{X}^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n}$$

(Thus, we can see that the MME estimation of σ^2 is not the same as the definition of sample variance S^2. $\hat{\sigma}^2 = \frac{n-1}{n} S^2$. So, the MME estimator is not an unbiased one?)

Example 2:

P.292

If a sample is from a Gamma distribution $X_i \sim GAM(\theta, k)$, and we want to estimate the θ and k.

We know that for Gamma distribution, the mean is $k\theta$, and the variance is $k\theta^2$.

We also know that $\mu_1^{'} = \mu = k\theta$ and $\mu_2^{'} = \sigma^2 + \mu^2 = k\theta^2 + k^2\theta^2 = k\theta^2(1+k)$.

Thus, we can get

$$\bar{X} = k\theta$$

$$\sum \frac{X_i^2}{n} = k\theta^2(1+k)$$

Thus, we can get

$$\hat{\theta} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n\bar{X}} = \frac{(n-1)/nS^2}{\bar{X}}$$

$$\hat{k} = \frac{\bar{X}}{\hat{\theta}}$$

(**Note: To sum up, (1) we know how to calculate moments based on sample. And, (2) we know that connection between moment and parameter of mean and variance. Thus, combining (1) and (2), we can get the estimators for the parameters.**)

2.2.1.3 Property

The joint MGF of $(X_1, ..., X_n)$ is defined as $M(t_1, ..., t_n) = E(e^{\sum_{i=1}^{n} t_i X_i})$

When $X_1, ..., X_n$ are independent if and only if

$$M(t_1, ..., t_n) = \prod_{i=1}^{n} M_{X_i}(t_i)$$

where $M_{X_i}(t_i)$ is the MGF of X_i

2.2.1.4 Well-known MGF

(1) Bernoulli with success probability p: $1 - p + pe^t$

(2) Binomial $Bin(n, p)$: $(1 - p + pe^t)^n$

(3) Poisson $POI(\lambda)$: $e^{\lambda(e^t - 1)}$

(4) Normal $N(\mu, \sigma^2)$: $e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

(5) Gamma $GAM(\theta, k)$: $(1 - \theta t)^{-k}$

Two special cases:

(6) Chi-square $\chi^2(v) = GAM(2, \frac{v}{2})$: $(1 - 2t)^{-\frac{v}{2}}$

(7) Exponential $EXP(\theta) = GAM(\theta, 1)$: $(1 - \theta t)^{-1}$

2.2.2 least squares estimators

2.2.3 likelihood function and maximum likelihood estimators

2.2.3.1 Likelihood function

P.293

The joint density function of n random variables $X_1, ...X_n$ evaluated at $x_1, ...x_n$, say $f(x_1, ..., x_n; \theta)$, is refererred to as a *likelihood function.*

2.2.3.2 Maximum likelihood estimators

P.294

Let $L(\theta) = f(x_1, ...x_n; \theta), \theta \in \Omega$, be the joint pdf of $X_1, ..., X_n$. For a given set of observations, $(x_1, ...x_n)$, a value $\hat{\theta}$ in Ω at which $L(\theta)$ is a maximum is called a *maximum likelihood estimate (MLE)* of θ. That is $\hat{\theta}$ is a value of θ that satisfies

$$f(x_1, ...x_n; \theta) = MAX_{\theta \in \Omega} f(x_1, ..., x_n; \theta)$$

2.2.4 Invariance property of MLEs

P.296

If $\hat{\theta}$ is the MLE of θ and if $u(\theta)$ is a function of θ, then $u(\hat{\theta})$ is an MLE of $u(\theta)$.

Example 1:

We know that the *pdf* of exponential distribution ($X \sim EXP(\theta)$) is as follows:

$$\frac{1}{\theta}e^{-\frac{X}{\theta}}$$

Thus, its likelihood function is as follows

$$L(\theta) = \frac{1}{\theta^n}e^{-\frac{\sum X_i}{\theta}}$$

Thus, log-likelihood is as follows.

$$lnL(\theta) = -nln(\theta) - \frac{\sum X_i}{\theta}$$

Thus,

$$\frac{d}{d\theta}lnL(\theta) = -n\frac{1}{\theta} + \frac{\sum X_i}{\theta^2}$$

Thus, we can get the MLE for θ is $\hat{\theta} = \bar{x}$.

If we want to estimate $\tau(\theta) = P(X \geq 1)$:

$$\tau(\theta) = P(X \geq 1) = \int_1^\infty \frac{1}{\theta}e^{-\frac{X}{\theta}}dx = -\int_1^\infty e^{-\frac{X}{\theta}}d(-\frac{x}{\theta}) = -[e^{-\frac{X}{\theta}}]_1^\infty = -[0-e^{-\frac{1}{\theta}}] = e^{-\frac{1}{\theta}}$$

Thus, based on the invariance property, we know that the MLE for $\tau(\theta)$ is as follows.

$$e^{-\frac{1}{\bar{x}}}$$

Example 2: MLE vs. MME

P.296

Assume a random sample from a two-parameter exponential distribution, $X_i \sim EXP(1, \eta)$. Thus, the *pdf* is $e^{-(x-\eta)}$. Thus, the likelihood function is

$$L(\eta) = e^{-\sum(x_i-\eta)}$$

Thus, the log likelihood,

$$lnL(\eta) = -\sum(x_i - \eta) = n\eta - n\bar{X}$$

Thus, we know that as η increases, the log likelihood increases accordingly. Thus, we want to find the maximum η. Note that a two-parameter exponential distribution has the support of $x_i \geq \eta$. Thus, all η are smaller than any X_i. Thus, we can get the ML estimator is the first order statistic

$$\hat{\eta} = X_{1:n}$$

Note that the estimators above is based on ML. What would be the answer if using MME?

We know that for a two-parameter exponential distribution, its mean is $\mu = 1+\eta$. And, we know that based on MME, $\mu = \bar{X}$. Thus, we can get the following,

$$\hat{\eta} = \bar{X} - 1$$

Conclusion

We can see that ML and MME have didferent estimators for the same η for a two-parameter exponential distribution.

2.2.5 Unbiased estimators

An estimator T is said to be an unbiased estimator of $\tau(\theta)$ if

$$E(T) = \tau(\theta)$$

for all $\theta \in \Omega$. Otherwise, we said that T is biased stimator of $\tau(\theta)$.

For instance, if we want to estimate a percentile, say the 95th percentile of $N(\mu, 9)$. Note that the percentiles that we know are about standardized noraml (i.e., $N(0, 1)$). Thus, we need to have some calculation to get the non-standard one.

$$\frac{X_{95\ percentile} - \mu}{\sigma} = 1.645$$

Thus, we can get

$$X_{95\ percentile} = 1.645 \times \sigma + \mu$$

We know that $\bar{X}$ is the unbiased estimate for μ. Thus, we can get

$$X_{95\ percentile} = 1.645 \times \sigma + \mu = 4.94 + \mu$$

We know that

$$E(T) = E(\bar{X} + 4.94) = \mu + 4.94$$

Thus, $T = \bar{X} + 4.94$ is the unbiased estimator of $\tau(\mu) = \mu + 4.94$.

2.2.6 Unbiased estimators vs. Invariance property of MLEs

Do not apply "Invariance property of MLEs" to the Unbiased estimators.

(**Note that:** You can apply the Invariance property to "Unbiased estimators" when it is a linear combination. In that case, $E(a\theta + b) = aE(\theta) + b$. Thus, if you find a T that is a unbiased estimator for θ, it should be unbiased estimator for $a\theta + b$ as well. Thus, note that $\frac{1}{\theta}$ is not a linear combination of θ, thus $\frac{1}{\theta}$ has a very different estimator, compared to θ.)

P.303

For example, consider a random sample of size n from an exponential distribution, $X_i \sim EXP(\theta)$, with parameter θ. We know that, $\bar{X}$ is unbiased for θ (which is the mean of an exponential distribution).

If we want to estimate $\tau(\theta) = \frac{1}{\theta}$, then by the invariance property of MLE is $T_1 = \frac{1}{\bar{X}}$.

However, T_1 is a biased estimators of $\frac{1}{\theta}$. Specifically,

We know that if $X \sim Gam(\theta, k)$, then $Y = \frac{2X}{\theta} \sim x^2(2k)$. We know that exponential distributions are a specicial case of Gamma distribution, $EXP(\theta) = Gam(\theta, 1)$, thus we get the following,

$$Y = \frac{2n\bar{X}}{\theta} = \frac{2n}{\theta}\frac{\sum_{i=1}^n X_i}{n} = \frac{\sum_{i=1}^n 2X_i}{\theta} \sim \sum_{i=1}^n x^2(2 \cdot 1) = \sum_{i=1}^n x^2(2) = x^2(2n)$$

We further know that if $Y \sim x^2(v)$, $E(Y^r) = 2^r \frac{\Gamma(v/2+r)}{\Gamma(v/2)}$. Thus, we know that,

$$E(Y^{-1}) = 2^{-1}\frac{\Gamma(2n/2-1)}{\Gamma(2n/2)} = \frac{1}{2} \cdot \frac{1}{n-1}$$

Thus,

$$E(Y^{-1}) = E(\frac{\theta}{2n\bar{X}}) = \frac{\theta}{2n}E(\frac{1}{\bar{X}}) = \frac{1}{2} \cdot \frac{1}{n-1}$$

Thus.

$$E(\frac{1}{\bar{X}}) = \frac{1}{n-1}\frac{n}{\theta} = \frac{n}{n-1}\frac{1}{\theta}$$

Thus,

$$E(\frac{n-1}{n}\frac{1}{\bar{X}}) = \frac{1}{\theta}$$

Conclusion:

$\frac{1}{\bar{X}}$ is not the unbiased estimator for $\frac{1}{\theta}$. However, we can adjust it to $\frac{n-1}{n}\frac{1}{\bar{X}}$, which is the unbiased estimator for $\frac{1}{\theta}$. When the sample size is big enough, we know that $\frac{n-1}{n}$ will be close to 1.

2.2.7 UMVUE and Cramer-Rao lower bound

2.2.7.1 UMVUE

Let $X_1, X_2, ..., X_n$ be a random sample of size n from $f(x;\theta)$. An estimator T^* of $\tau(\theta)$ is called a *uniformly minimum variance ubiased estimator* (UMVUE) of $\tau(\theta)$ if

1. T^* is unbiased for $\tau(\theta)$.
2. For any other unbiased estimator T of $\tau(\theta)$, $Var(T^*) \leq Var(T)$ for all $\theta \in \Omega$.

2.2.7.2 Cramer-Rao lower bound

If T is an unbiased estimator of $\tau(\theta)$, then the Cramer-Rao lower bound (CRLB), based on a random sample, is

$$Var(T) = \frac{[\tau^{'}(\theta)]^2}{nE[\frac{\partial}{\partial\theta}lnf(X;\theta)]^2}$$

Example:

Consider a random sample from an exponential distribution, $X_i \sim Exp(\theta)$. Because

$$f(x;\theta) = \frac{1}{\theta}e^{-\frac{x}{\theta}}$$

Thus,

$$ln(f(x;\theta)) = -\frac{x}{\theta} - ln\theta$$

$$\frac{\partial}{\partial\theta} ln(f(X;\theta)) = \frac{X}{\theta^2} - \frac{1}{\theta} = \frac{X-\theta}{\theta^2}$$

Thus,

$$E[\frac{\partial}{\partial\theta} ln(f(X;\theta))]^2 = E[\frac{(X-\theta)^2}{\theta^4}] = \frac{1}{\theta^4} E(X-\theta)^2 = \frac{1}{\theta^4} Var(X) = \frac{\theta^2}{\theta^4} = \frac{1}{\theta^2}$$

Thus, the CRLB for $\tau(\theta) = \theta$ is as follows.

$$Var(T) \geq \frac{[\tau^{'}(\theta)]^2}{nE[\frac{\partial}{\partial\theta} lnf(X;\theta)]^2} = \frac{[\frac{\partial}{\partial\theta}\theta]^2}{n\frac{1}{\theta^2}} = \frac{1^2}{\frac{n}{\theta^2}} = \frac{\theta^2}{n}$$

We know that the variance for the sample mean of the exponential distribution:

$$Var(\bar{X}) = \frac{Var(X)}{n} = \frac{\theta^2}{n}$$

In addition, we also know that sample mean $\bar{X}$ is the unbiased estimator for population mean.

Thus,

$\bar{X}$ is the UMVUE of θ.

2.2.8 Best linear unbiased estimation (BLUE or MVLUE)

2.2.9 Consistency, asymptotic unbiasedness

Simple consistency:

P.311

Let $\{T_n\}$ be a sequence of estimators of $\tau(\theta)$. These estimators are said to be **consistent** estimators of $\tau(\theta)$ if for every $\varepsilon > 0$,

$$lim_{n\to\infty} P[|T_n - \tau(\theta)| < \varepsilon] = 1$$

for every $\theta \in \Omega$.

In the terminology of Chapter 7, T_n converges **stochastically** to $\tau(\theta)$, $T_n \xrightarrow{P} \tau(\theta)$ as $n \to \infty$. Sometimes this also is referred to as **simple** consistency.

One interpretation of consistency is that for large sample size the estimator tends to be more concentrated about $\tau(\theta)$, and by making n sufficiently large T_n can be made as concentrated as desired.

MSE consistency:

If $\{T_n\}$ is a sequence of estimator of $\tau(\theta)$, then they are called **mean squared error consistent** if

$$lim_{n\to\infty} E[T_n - \tau(\theta)]^2 = 0$$

for every $\theta \in \Omega$.

Asymptotic Unbiased

A sequence $\{T_n\}$ is said to be **asymptotically unbiased** for $\tau(\theta)$ if

$$lim_{n\to\infty} E(T_n) = \tau(\theta)$$

for every $\theta \in \Omega$.

Example:

P.313

For a sample from $X_i \sim EXP(\theta)$, we know that $T_n = 1/\bar{X}$ is an MLE estimator for $\tau(\theta) = 1/\theta$. Howwever, T_n is not a unbiased estimator for $\tau(\theta)$, as

$$E(T_n) = \frac{n}{n-1} \cdot \frac{1}{\theta}$$

We also know that

$$Var(T_n) = \frac{(\frac{n}{n-1})^2}{(n-2)\theta^2}$$

Thus, while T_n is not unbiased, it is asymptotically unbiased and MSE consistent for $\tau(\theta) = \frac{1}{\theta}$.

Note that, MSE consistency is a stronger property than simple consistency. Thus, if a sequenc $\{T_n\}$ is mean squared error consistent, it is also simply consistent.

2.2.10 Efficiency

P.308

The relative efficiency of an **unbiased** estimator T of $\tau(\theta)$ to another **unbiased** estimator T^* of $\tau(\theta)$ is given by

$$re(T, T^*) = \frac{Var(T^*)}{Var(T)}$$

An unbiased estimator T^* of $\tau(\theta)$ is said to be efficient if $re(T, T^*) \leq 1$ for all unbiased estimators T of $\tau(\theta)$, and all $\theta \in \Omega$. The efficiency of an unbiased estimator T of $\tau(\theta)$ is given by

$$e(T) = re(T, T^*)$$

if T^* is an efficient estimator of $\tau(\theta)$.

Thus, **an effecient estimator is just a UMVUE.**

Example:

P.233 Example 7.2.2

P.303 Example 9.3.2

P.309 Example 9.3.7

Let $X_1, X_2, ..., X_n$ sample from $X_i \sim EXP(\theta)$. We know that

$$X_{1:n} = EXP(\theta/n)$$

Thus,

$$E(nX_{1:n}) = nE(X_{1:n}) = n\frac{\theta}{n} = \theta$$

Thus,

$$nX_{1:n}$$

is the unbiased estimator for θ.

Thus,

$$re(T, T^*) = \frac{Var(T^*)}{Var(T)} = \frac{Var(\bar{X})}{Var(nX_{1:n})} = \frac{\theta^2/n}{n^2 Var(X_{1:n})} = \frac{\theta^2/n}{n^2(\theta/n)^2} = \frac{\theta^2/n}{\theta^2} = \frac{1}{n}$$

Thus, $T^* = \bar{X}$ is a more efficient estimator for θ than $T = nX_{1:n}$.

2.2.11 Asymptotic efficiency

Let $\{T_n\}$ and $\{T_n^*\}$ be the two asymptotically unbiased sequences of estimators for $\tau(\theta)$. The **asymptotic relative efficiency** of $T_n$relative to T_n^* is given by

$$are(T_n, T_n^*) = lim_{n\to\infty}\frac{Var(T_n^*)}{Var(T_n)}$$

The sequence $\{T_n^*\}$ is said to asymptotically efficient if $are\{T_n, T_n^*\} \leq 1$ for all other asymptotically unbiased sequences $\{T_n\}$, and all $\theta \in \Omega$.

2.2.12 Asymptotic properties of MLEs

P.316

If certain regularity conditions are satisfied, then the solutions, $\hat{\theta}$, of the MLE have the following properties:

(1) $\hat{\theta_n}$ exists and is unique.

(2) $\hat{\theta_n}$ is a consistent estimator of θ.

(3) $\hat{\theta_n}$ is asymptotically normal with asymptotic mean θ and variance

$$\frac{1}{2}E[\frac{\partial}{\partial\theta}ln\ f(X;\theta)]^2$$

(4) $\hat{\theta}$ is asymptotically efficient.

Note that the asymptotic efficiency of $\hat{\theta}$ follows from the fact that the asymptotic variance is the same as the **CRLB** for unbiased estimators of θ. Thus, for large n, approximately

$$\hat{\theta_n} \sim N(\theta, CRLB)$$

Example

P.317

From Example 9.2.7, we know the MLE of the mean θ of an exponential distribution is the sample mean, $\hat{\theta_n} = \bar{X}$.

We can infer the same asymptotic properties either from the discussion above or from the Central Limit Theorem. That is, $\hat{\theta_n}$ is asymptotically normal with asymptotic mean θ and variance θ^2/n. From example 9.3.4, we know that $CRLB = \theta^2/n$.

Thus,

$$\sqrt{n}\frac{\hat{\theta_n} - \theta}{\theta} \sim N(0, 1)$$

2.3 Sufficient and completeness

2.3.1 Sufficiency and minimal sufficiency

P.335

Sufficient statistic:

A statistic S will be consiered a "sufficient" statistic for a parameter θ if the conditional distribution of any other statistic T given the value of S does not involve θ.

Jointly sufficient statistics:

Let $X = (X_1, ..., X_n)$ have joint pdf $f(x, \theta)$, and let $S = (S_1, ..., S_k)$ be a k-dimensional statistic.

Then, $S_1, ..., S_k$ is a set of **jointly sufficient statistics** for θ if for any other vector of statistics, T, the conditional pdf of T given $S = s$, denoted by $f_{T|s}(t)$, does not depend on θ.

In the one-dimension case, we simply say that S is a sufficient statistic for θ.

If k unknown parameters are present in the model, then quite often there will exist a set of k sufficient statistics. In some cases, the number of sufficient statistics will exceed the number of parameters.

Minimal sufficient:

A set of statistics is called a **minimal sufficient** set if the members of the set are jointly sufficient for the parameters and if they are a function of every other set of jointly sufficient statistics.

In other words, once the value of a sufficient statistic is known, the observed value of any other statistic does not contain any further inforamtion about the parameter.

Example:

P.339

Assume a random sample from an exponential distribution, $X_i \sim EXP(\theta)$. It follows that

$$f(x_1, ..., x_n; \theta) = \frac{1}{\theta^n} e^{-\frac{\sum X_i}{\theta}}$$

which suggests checking the stastic $S = \sum X_i$. We know that $S \sim GAM(\theta, n)$, thus

$$f_S(s; \theta) = \frac{1}{\theta^n \Gamma(n)} s^{n-1} e^{-\frac{s}{n}}$$

If $s = \sum x_i$, then,

$$\frac{f(x_1, ..., x_n; \theta)}{f_S(s; \theta)} = \frac{\Gamma(n)}{s^{n-1}}$$

which is free of θ, and thus by definition S is sufficient for θ.

2.3.2 Neyman factorization theorem, minimal sufficiency of MLEs

Factorization criterion:

If $X_1, ..., X_n$ have joint pdf $f(x_1, ..., x_n; \theta)$, and if $S = (S_1, ..., S_n)$, then $S_1, ..., S_k$ are joinly sufficient for θ if and only if

$$f(x_1, ...x_n; \theta) = g(s; \theta)h(x_1, ..., x_n)$$

where $g(s; \theta)$ does not depend on $x_1, ..., x_n$, except through s, and $h(x_1, ..., x_n)$ does not involve θ.

Example 1:

P.340

$X_i \sim Bin(1, \theta)$. We can use factorization criterion to check its sufficient statistic.

$$f(x_1, ..., x_n; \theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} = \theta^s(1-\theta)^{n-s} = g(s; \theta)h(x_1, ..., x_n)$$

where $s = \sum X_i$ and $h(x_1, ...x_n) = 1$. Thus, s is sufficient for θ.

Example 2:

It is important to specify the regions of zero probability. The following shows that care must be exercised in this matter.

P.340

$X_i \sim UNIF(0, \theta)$, where θ is unknown. We get the joint *pdf* of $X_1, ..., X_n$ is

$$f(x_1, ...x_n; \theta) = \frac{1}{\theta^n}$$

where,

$$0 < x_i < \theta$$

It is easier to specify this *pdf* in terms of the minimum, $x_{1:n}$, and maximum, $x_{n:n}$, of $x_1, ..., x_n$. In particular,

$$0 < x_{1:n} \;\; x_{n:n} < \theta$$

Thus, $x_{n:n}$ is sufficient for θ.

Example 3:

P.341

Consider a random sample from a normal distribution, $X_i \sim N(\mu, \sigma^2)$.

We know the *pdf* for normal is

$$f = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Thus, the joint *pdf* is as follows.

$$\begin{aligned} f(x_1, ..., x_n; \mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum(x_i^2+\mu^2-2x_i\mu)} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}(\sum x_i^2+n\mu^2-2\mu\sum x_i)} \end{aligned}$$

Thus, $S_1 = \sum X_i$ and $S_2 = \sum X_i^2$ are jointly sufficient for $\theta = \mu, \sigma^2$.

Note that,

P.341

based on *P.298* , Example 9.2.10, ML estimation results in

$$\hat{\mu} = \bar{x}$$

$$\hat{\sigma^2} = \frac{\sum(x_i - \bar{x})^2}{n}$$

Further, note that

$$\hat{\mu} = \bar{x} = \frac{S_1}{n}$$

$$\begin{aligned}
\hat{\sigma^2} &= \frac{\sum(x_i - \bar{x})^2}{n} = \frac{\sum(x_i^2 + \bar{x}^2 - 2x_i\bar{x})}{n} \\
&= \frac{\sum x_i^2 + n\bar{x}^2 - 2\bar{x}\sum x_i}{n} \\
&= \frac{\sum x_i^2 + n\bar{x}^2 - 2n\bar{x}^2}{n} \\
&= \frac{\sum x_i^2 - n\bar{x}^2}{n} \\
&= \frac{\sum x_i^2}{n} - \bar{x}^2 \\
&= \frac{\sum x_i^2}{n} - (\frac{\sum x_i}{n})^2 \\
&= \frac{S_2}{n} - (\frac{S_1}{n})^2
\end{aligned}$$

Thus, we can see that $\hat{\mu}$ and $\hat{\sigma^2}$ correspond to a one-to-one stransformation of S_1 and S_2, thus $\hat{\mu}$ and $\hat{\sigma^2}$ also are jontly sufficient for μ and σ^2.

2.3.3 Rao-Blackwell theorem

Let $X_1, ..., X_n$ have joint *pdf* $f(x_1, ..., x_n; \theta)$, and let $S = (S_1, ..., S_k)$ be a vector of jointly sufficient statistics for θ. If T is any unbiased estimator of $\tau(\theta)$, and if $T^* = E(T|S)$, then

(1) T^* is an unbiased estimator of $\tau(\theta)$,

(2) T^* is a function of S, and

(3) $Var(T^*) \leq Var(T)$ for every θ, and $Var(T^*) < Var(T)$ for some θ unless $T^* = T$ with probability 1.

Discussion:

P.345

"It is clear from the Rao-Blackwell theorem that if we are searching for an unbiased estimator with small variance, we may as well restrict attention to function of sufficient statistics.

If any ubiased estimator exists, then there will be one that is a function of sufficient statistics, namely $E(T|S)$, which also is unbiased and has variance at least as small or smaller. In particular, we still are interested in knowing how to find a UMVUE for a parameter, and the above theorem narrows our problem down somewhat.

For instance, consider a one-parameter model $f(x; \theta)$, and assume that a single sufficient statistic, S, exists. We know that we must consider only unbiased functions of S in searching for a UMVUE. In some cases it may be possible to

show that only one function of S is unbiased, and in that case we would know that it is a UMVUE.

The concept of "completeness" is helpful in determinning unique unbiased estimators, and this concept is defined in the next section."

2.3.4 completeness

Completeness:

P.345

A family of density functions $\{f_T(t,\theta);\theta\in\Omega\}$, is called **complete** if $E[u(T)]=0$ for all $\theta\in\Omega$ implies $u(T)=0$ with probability 1 for all $\theta\in\Omega$.

P.345

"This sometimes is expressed by saying that there are no nontrivial unbiased estimators of zero. In particular, it means that two different functions of T cannot have the same expected value."

For example, if $E[u_1(T)]=\tau(\theta)$ and $E[u_2(T)]=\tau(\theta)$, which implies $u_1(T)-u_2(T)=0$, or $u_1(T)=u_2(T)$ with probablity 1, if the family of density fucntions is complete. That is, any unbiased estimator is unique in this case.

We mainly are interested in knowing that the family of density functions of a sufficient statistic is complete, because in that case an unbiased function of the sufficient statistic will be unique, and it must be a UMVUE by the Rao-Blackwell theorem.

2.3.5 Lehmann-Scheffe completeness theorem

Let $X_1,...,X_n$ have joint *pdf* $f(x_1,...,x_n;\theta)$ and let S be a vector of jointly complete sufficient statistics for θ. If $T^*=t^*(S)$ is a statistic that is unbiased for $\tau(\theta)$ and a function of S, then T^* is a UMVUE of $tau(\theta)$.

Example:

P.346

2.3.6 Exponential class, complete sufficient statistics

A density fucntion is said to be a member of the **regular exponential class** if it can be expressed in the form

$$f(x;\theta)=c(\theta)h(x)e^{\sum_{j=1}^{k}q_j(\theta)t_j(x)}$$

where,

$$x\in A$$

And zero otherwise, where $\theta = (\theta_1, ..., \theta_k)$ is a vector of k unknown parameters, if the parameter space has the form

$$\Omega = \{\theta | a_i \leq \theta_i \leq b_i, i = 1, ...k\}$$

(note that $a_i = -\infty$ and $b_i = \infty$ are permissible values), and if it satisfies regularity conditions 1, 2, and 3a or 3b given by:

(1) The set $A = \{x : f(x; \theta) > 0\}$ does not depend on θ.

(2) The functions $q_j(\theta)$ are nontrivial, functionally independent, continuous fucntion of the θ_i.

3a. For a continuous random variable, the derivatives $t_k^{'}(x)$ are linearly independent continous functions of x over A.

3b. For a discrete random variable, the $t_k(x)$ are nontrivial functions of x on A, and none is a linear function of the others.

Example

P.348

Consider Bernoulli distribution $X \sim Bin(1, p)$. It follows that

$$\begin{aligned} f(x; p) &= p^x (1-p)^{1-x} \\ &= P^x (1-p)(1-p)^{-x} \\ &= (1-p)(1-p)^{-x} P^x \\ &= (1-p) e^{ln(1-p)^{-x} P^x} \\ &= (1-p) e^{x ln(\frac{p}{1-p})} \end{aligned}$$

Compared to the definition of **Exponential Class** defined above, we can get the following.

$$c(\theta) = (1-p)$$

$$h(x) = 1$$

$$q_1(\theta) = ln(\frac{p}{1-p})$$

$$t_1(x) = x$$

Theorem 10.4.2

P.348

If $X_1, ..., X_n$ is a random sample from a member of the regular exponential class $REC(q_1, ..., q_k)$, then the statistics

$$S_1 = \sum_{i=1}^{n} t_1(X_i), ..., S_k = \sum_{i=1}^{n} t_k(X_i)$$

are a minimal set of complete sufficient statistics for $\theta_1, ...\theta_k$.

Example 1:

P.348

Again, consider Bernoulli distribution $X \sim Bin(1, p)$. For a random sample of size n, $t(x_i) = x_i$ and thus based on **Theorem 10.4.2 (see above)** $S = \sum_{i=1}^{n} X_i$ is a complete sufficient statistic for p.

How about we want to find UMVUE for $Var(X) = p(1-p)$**?**

We might try $\bar{X}(1-\bar{X})$.

$$\begin{aligned} E[\bar{X}(1-\bar{X})] &= E(\bar{X}) - E(\bar{X}^2) \\ &= p - (p^2 + var(\bar{X})) \\ &= p - p^2 - \frac{p(1-p)}{n} \\ &= p(1-p)(1-\frac{1}{n}) \end{aligned}$$

Thus,

$$E[\frac{1}{1-\frac{1}{n}}\bar{X}(1-\bar{X})] = p(1-p)$$

Thus, $\frac{1}{1-\frac{1}{n}}\bar{X}(1-\bar{X})$ is the UMVUE of the $p(1-p)$.

but **why?**

Example 2:

P.349

If $X \sim N(\mu, \sigma^2)$, then

$$\begin{aligned} f(x; \mu, \sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x^2+\mu^2-2x\mu)} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} + \frac{x\mu}{\sigma^2}} \end{aligned}$$

Thus, $S_1 = \sum X_i^2$ and $S_2 = \sum X_i$ are jointly complete and sufficient statistics of μ and σ^2. **Refer to the normal example in the section of "Neyman factorization theorem, minimal sufficiency of MLEs"**

Theorem 10.4.3

P.349

If a CRLB estimator T exists for $\tau(\theta)$, then a single sufficient statistic exists, and T is a function of the sufficient statistic. Conversely, if a single sufficient statistic exists and teh CRLB exists, then a CRLB estimator exists for some $\tau(\theta)$.

Chapter 3

530_533

https://www.ssc.wisc.edu/sscc/pubs/RegressionDiagnostics.html

http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/

3.1 Definition of the general linear model

$$Y = X\beta + \varepsilon$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & ... & x_{1m} \\ 1 & x_{21} & x_{22} & x_{23} & ... & x_{2m} \\ ... & & & & & \\ 1 & x_{n1} & x_{n2} & x_{n3} & ... & x_{nm} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{bmatrix}$$

Where,

Y: Response vector

X: Design matrix

β: parameter vector

ε: error vector

If ε follows a mutivariate normal distribution then we will be under the General Linear Model (GLM) framework.

$$\varepsilon \sim N(0, \sigma^2 I_{x \times n})$$

If X is continuous we have regression.

If X is categorical we have ANOVA.

If X is a mix of both, we have ANCOVA.

3.2 Simple Linear Regression

Simple linear regression is a linear regression model with a single explanatory variable. In addition, we typically assume that this is under the GLM framework and thus we also assume that the residuals follow normal distribution.

3.2.1 Least squares and propoerties of the regression parameters

3.2.1.1 Basic idea

Of note The following method can always calculate the vector of β, not related to any specific methods.

$$Y_{n\times 1} = X_{n\times m}\beta_{m\times 1}$$

$\rightarrow$

$$[X^T]_{m\times n}Y_{n\times 1} = [X^T]_{m\times n}X_{n\times m}\beta_{m\times 1}$$

$\rightarrow$

$$[X^TX]^{-1}_{m\times m}[X^TY]_{m\times 1} = \beta_{m\times 1}$$

$\rightarrow$

$$\beta_{m\times 1} = [(X^TX)^{-1}\cdot(X^TY)]_{m\times 1}$$

3.2.1.2 Least Squares

Assume the following model:

$$Y = X\beta + \varepsilon$$

When β only has a dimension of 2×1, we can write it as follows.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon$$

Thus,

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

We can calculate the partial derivates as follows.

$$\frac{\partial}{\partial \beta_0} Q \to \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i] = 0$$

$$\frac{\partial}{\partial \beta_1} Q \to \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i] X_i = 0$$

Combining the two pieces of information, we can get

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$b_0 = \frac{1}{n} (\sum Y_i - b_1 \sum X_i) = \bar{Y} - b_1 \bar{X}$$

Properties of Least Squares Estimators (Gauss- Markov theorem):

P.18

Under the conditions of regression model shown above, the least squares estimator b_0 and b_1 are unbiased and have minimum variance among all unbiased linear estimators.

That is,

$$E(b_0) = \beta_0$$

$$E(b_1) = \beta_1$$

Point Estimation of Mean Response:

Given sample estimators b_0 and b_1 in the regression function

$$E\{Y\} = \beta_0 + \beta_1 X$$

We estimate the regression function as follows:

$$\hat{Y} = b_0 + b_1 X$$

We call a value of the response variable a response and $E\{Y\}$ the *mean response.*

Propoerties of fitted regression line

P.22

The resitual:

$$e_i = Y_i - \hat{Y}_i$$

The sum of the residuals is zero

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

3.2.1.3 Point estimator of σ^2

Note that, there are two S^2 below, and they have different formulus, even though the only difference is the degree of freedom and the logic of calculating df is the same across these two cases (i.e., **Single Population vs. Regression Model**).

1. Single Pupulation

Sum of squares:

P.25

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

The sum of squares is then divided by the degrees of freedom associated with it.

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

S^2 is an unbiased estimator of the variance σ^2. The sample variance S^2 is often called a mean square, because a sum of squares has been divided by the appropriate number of degrees of freedom.

$$E(S^2) = \sigma^2$$

2. Regression model

Sum of Squares Error (SSE) (or, Sum of Squares Residual):

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

Mean Square Error (MSE) (or, Mean Square Residual):

$$S^2 = MSE = \frac{SSE}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

MSE is an unbiased estimator of σ^2 for regression model.

$$E(MSE = S^2) = \sigma^2$$

3.2.2 MLE

P.30

Note that the **Normal Error Regression Model** is as follows.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon$$

(This is a special case of the definition provided earlier, i.e., $Y = X\beta + \varepsilon$)

For this model, each Y_i observation is normally distributed with mean $\beta_0 + \beta_1 x_i$ and a standard devision σ.

Givern Y_i follows normal distribution, we can use *pdf* of normal distributions and MLE to estimate parameters.

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} exp[-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2] \end{aligned}$$

Since the variance $\sigma^2$is usually unknown, the likelihood fucntion is a function of three parameters, β_0, β_1, and σ^2.

Thus, we can calculate β_0, β_1, and σ^2 analytically, and the results of β_0 and β_1 are the same as least squares estimators (see above).

The variance σ^2 is as follows.

$$\hat{\sigma}^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n}$$

As noted in the least square estimation, we know that $\hat{\sigma}^2$ is biased. Thus, we know that:

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

3.3 Full rank, less than full rank

http://www.biostat.jhsph.edu/~iruczins/teaching/140.751/notes/ch7.pdf

$$Y_{n\times 1} = X_{n\times m}\beta_{m\times 1}$$

If the rank r of $X_{n\times m}$ is smaller than m, i.e., $r < m$, there is not a unique solution $\hat{\beta}$. We have three ways to find a solution $\hat{\beta}$ and the orthogonal projection $\hat{Y}$:

1. Reducing the model to the full rank.

Let X_1 consist of r linear independent columns from X and let X_2 consistn of the remaining columns. Then, $X_2 = X_1F$ because the columns of X_2 are linearly dependent on the columns of X_1.

$$X = (X_1, X_2) = (X_1, X_1F) = X_1[I_{r\times r}, F]$$

This is a special case of the factorization $X = KL$, where rank $(K_{n\times r}) = r$ and rank $(L_{r\times p}) = r$.

$$E[Y] = x\beta = KL\beta = k\alpha$$

Since K has full rank, the Least Squares Estimate of α is $\hat{\alpha} = (K^TK)^{-1}\cdot K^TY$.

The orthogonal project,

$$\hat{Y} = K\cdot\alpha = K\cdot(K^TK)^{-1}\cdot K^TY = X_1\cdot(X_1^TX_1)^{-1}\cdot X_1^TY$$

$$\begin{bmatrix} Y_{11} \\ \dots \\ Y_{1n_1} \\ Y_{21} \\ \dots \\ Y_{2n_2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ \dots & \dots & \dots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \dots & \dots & \dots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \dots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \dots \\ \varepsilon_{2n_2} \end{bmatrix}$$

Let X_1 consist of the first 2 columns of X, then

$$X = X_1 \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix}$$

Thus,

$$X_1 = K = \begin{bmatrix} 1 & 1 \\ ... & ... \\ 1 & 1 \\ 1 & 0 \\ ... & ... \\ 1 & 0 \end{bmatrix}$$

$$\hat{\alpha} = (K^T K)^{-1} \cdot K^T Y = \begin{bmatrix} n & n_1 \\ n_1 & n_1 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y_{1j} + \sum y_{2j} \\ \sum y_{1j} \end{bmatrix} = \begin{bmatrix} \bar{Y_2} \\ \bar{Y_1} - \bar{Y_2} \end{bmatrix}$$

$$\hat{Y} = X_1 \hat{\alpha} = \begin{bmatrix} 1 & 1 \\ ... & ... \\ 1 & 1 \\ 1 & 0 \\ ... & ... \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \bar{Y_2} \\ \bar{Y_1} - \bar{Y_2} \end{bmatrix} = \begin{bmatrix} \bar{Y_1} \\ ... \\ \bar{Y_1} \\ \bar{Y_2} \\ ... \\ \bar{Y_2} \end{bmatrix}$$

2. Find the generalized inverse $(X^T X)^{-1}$.

As noted above, when $X^T X$ has a full rank, we can directly calculate the inverse of $X^T X$. That is,

$$\beta_{m \times 1} = [(X^T X)^{-1} \cdot (X^T Y)]_{m \times 1}$$

We can just find **some columns** within X that are independent, and then calculate the inverse of it.

This is because if a matrix W with a rank r and can be partitioned as follows.

$$W = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

Assume A has rank r, then

$$W^{-1} = \begin{bmatrix} A^{-1} & 0 \\ 0 & 0 \end{bmatrix}$$

Thus, let $X = (X_1, X_2)$, where X_1 consists of r linearly independent columns from X. Then a generalized inverse of $X^T X$ is

$$[X^T X]^{-1} = \begin{bmatrix} [X_1^T X_1]^{-1} & 0 \\ 0 & 0 \end{bmatrix}$$

Thus, all other steps are similar to the full rank case.

3. Impose identifiifiability constraints

(Not very sure this one.)

3.4 Assumptions, checking assumptions

The following are some of the possible violations of assumptions:

3.4.1 Regression function is nonlinear___

Plot explanatory and response variables to see whether their relationship is linear, it is not, suggesting that a linear regression function is not appropriate.

Stat Methods:

F-test for lack of fit.

3.4.2 Non-constant variance___

Plots of the residuals against the predictor variable (i.e., X) or against the fitted values (i.e., $\hat{Y}$), to study whether the linear model is appropriate and whether the **variance of the error terms** is constant.

Stat Methods:

(1) Modified Levene's test (Brown-Forsythe test):

To check and see if the variability for Y for the smaller $X's$ are different than the vairability of Y for the larger $X's$.

Thus, we can break up the data into two groups based $X's$ values. Then, we test to see if the vairability of these two groups significantly differ.

Note that Modefied Levene's test does not depend on normality of the error terms. That is, this test is robust against serious departures from normality.

(2) Breusch-Pagan test:

Want to see if there is any relationship between σ_i and the X_i's. To do this, we can fit a regression of $log(\sigma_i)$ on X_i's:

$$log(\sigma_i) = b_0 + b_1 X_i$$

3.4.3 Error terms not independent___

We want a random pattern. Any type of pattern indicates time orderred problems indicates that the model is not appropriate.

Stat Methods:

Durbin Watson Test Test for presence of autocorrelation amongst observations.

3.4.4 Possibility of outliers___

Scatter plot the dependent variable and independent variables (each X_i is plotted seprately.)

3.4.5 Non normal distribution of error___

Use histogram to check whether the resitual follows normal distribution.

Stat Methods:

P.115

Correlation between ordered residuals and their expected values under normality.

P.70 512 note

You can use Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling in SAS to assess the normal errors.

3.4.6 Omission of some of the important predictors___

Plots should be made betweeen variables omitted from the model and the dependent variable. Such omitted variables may have important effects on the response.

3.5 Bootstrapping used in linear models

P.458

Intro

For standard fitted regression models, methods described earlier chapters are available for evaluating the precision of estimated regression coefficients, fitted values, and predictions of new observations.

However, in many nonstandard situations, such as when nonconstant error variance are estimated by iteratively reweighted least sqaure or when robust regression estimation is used, standard methods for evaluating the precision may not be available or may only approximately applicable when the sample size is large.

Bootstrapping was developed by Efron to provide estimates of the precision of sample estimates for these complex cases.

Conceptual

Suppose that we have fitted a regression model (simple or multiple) by some procedure and get the coefficient b_1. We now wish to evaluate the precision of this esimtate by bootsrap method.

(1) In essensce, the bootstrap method call for the selection from the observed sample data of a random sample of size n with replacement.

(2) Next, the bootstrap method calculates the estimated regression coefficient from the boostrap sample, using the same fitting procedure as employed for the original fitting. This leads to the first bootstrap estimate b_1^*.

This procedure repeated a large number of times, each time a bootstrap sample of size n is selected with replacement from the original sample and the estimated coefficient is obtained.

(3) The estimated standard deviation of all the bootstrap estimates b_1^* denoted by $S^*\{b_1^*\}$, is an estimate of the variability of the sampling distribution of b_1 and therefore a measure of the prevision of b_1.

Some Math

Bootstrapping sampling for regression can be done in two basic ways.

(1) Whne the regression fucntion being fitted is a good model for the data, the error terms have constant variance, and the predictor variables can be regarded as fixed, fixed X sampling is appropriate.

Here, the resitudals e_i from the original fitting are regarded as the sample data to be sampled with replacement. After bootstrap sample of the resituals of size n have beeb obtained, denoted by $e_1^*, ..., e_n^*$. The bootstrap sample residuals are added to the fitted values from the original fitting to obtain nuew bootstrap Y values, denoted by $Y_1^*, ..., Y_n^*$:

$$Y_i^* = \hat{Y}_i + e_i^*$$

These bootstrap Y^* values are then regressed on the original X variables by the same procedure used initially to obtain the bootstrap estimate b_1^*.

(2) When there is some doubt about the adequacy of the regressio nmodel, or the rror variances are not constant, or the predictor variables can not be regarded as fixed, random X sampling is appropriate.

For simple regression, the pair X and Y data in the oringal sample are considered to be the data to be sampled with replacedment. This this second procedure samples cases with replacement $n times, yelding a bootstrap sample of n pairs of (X^*, Y^*) values. This bootstrap sample is then used for obtaining the bootstrap estimate b_1, as with fixed X sampling.

Bootstrap confidence intervals

P.460

3.6 Generalized linear models

3.6.1 Definition, similarities and differences from general linear models

Generalized Linear Model (GLiM) loosens this assumption that ε follows a mutivariate normal distribution, and allows for a variety of other distributions from the exponential family for the residuals.

Of note, the GLM is a special case of the GLiM in which the distribution of the residuals follow a conditionally normal distribution.

3.6.2 Advantage and disadvantages

3.6.3 logistic and poisson regression

Logistic regression

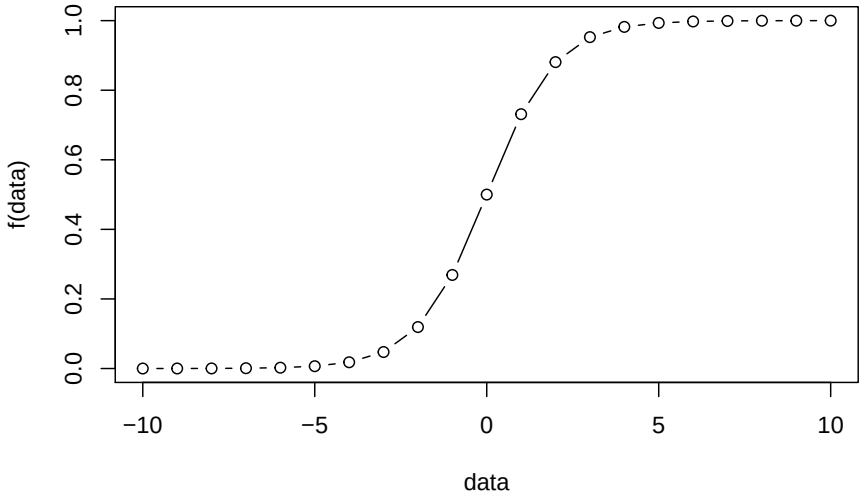
Logistic regression is a special case of GLiM with binomial distribution on the $Y's$ and the link function

The basic idea of logistic regression:

$$p(y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+...+\beta_n x_n)}} = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_n x_n}}$$

Thus, $\beta_0 + \beta_1 x_1 + ... + \beta_n x_n$ can be from $-\infty$ to $+\infty$, and $p(y=1)$ will be always within the range of $(0, 1)$.

```
f<-function(x){exp(x)/(1+exp(x))}
data<-seq(-10,10,1)
plot(data,f(data),type = "b")
```



We can also write the function into another format as follows:

$$log\frac{p(y=1)}{1-p(y=1)} = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$$

Thus, we know that the regression coeficients of β_i actually change the "log-odds" of the event. Of course, note that the magnitude of β_i is dependent upon the units of x_i.

Chapter 4

512

4.1 Basics

4.1.1 Random vs. Fixed

A factor is **random** if its levels represent a random sample from a population consisting of a large number of possible levels.

A factor is **fixed** if its levels are selected by a non-random process or if its levels consist of the entire population of possible levels.

4.1.2 Crossed vs. Nested

A factor, A, is said to be **crossed** with respect to a second factor, say B, if each level of factor A is exactly the same for each level of factor B, and each level of factor B is exactly the same for each level of factor A.

Otherwise, the factor is said to be **nested**. A factor, say B, is **nested** within another factor, say A, if the levels of factor B are **not** the same across all levels of factor A.

4.2 Completely randomized designs

Completely random design (CRD):

All experimental units are **randomly assigned** to all treatment combinations and therefore the design is a CRD design. That is, across all experimental units, they are assumed to be homogeneous before getting the assigned treatments assigned.

This is **Completely Randomized Design**. That is, we have a **one-way treatment structure** (Variety) in a **completely randomized design struc-**

Variety 1	farm 7	farm 5	farm 4
Variety 2	farm 3	farm 9	farm 6
Variety 3	farm 8	farm 1	farm 2

Figure 4.1: CRD

Farm 1	variety 2	variety 3	variety 1
Farm 2	variety 3	variety 1	variety 2
Farm 3	variety 3	variety 2	variety 1

Figure 4.2: RCBD

ture. Note that the experimental unit (Farm) is **nested** within treatment (Variety).

4.3 Randomized complete block designs

Randomized complete block design (RCBD):

Suppose there are **t-treatments**, with each to be observed **b times**, so that $b \times t$ experimental units are necessary. An RCBD arranges the $b \times t$ experimental units into b homogeneous groups (blocks) with each containing t base units.

The t treatments are randomly assigned to the t units within each block so that each treatment appears once with a block.

This is a **Randomized Complete Block Design**. That is, we have a one-way treatment structure (Variety) in a randomized complete block design structure.

We call the **farms** blocks or **blocking factor**. Note that block and treatment (Variety) are crossed.

4.4 Incomplete block designs

512 P. 293

These designs are similar to the RCBD except that each block (e.g., each farm) contains fewer than t units and thus, only a portion of the t treatments are applied to the units within each block. (e.g., not all the farms will get all the varieties of wheat).

	Fertility		
	High	Medium	Low
Farm 1	variety 1	variety 2	variety 3
Farm 2	variety 3	variety 1	variety 2
Farm 3	variety 2	variety 3	variety 1

Figure 4.3: LSD

4.5 Row-column design (Latin square designs)

(**My own short summary:** Latin Square Designs basically are a double-block design.)

This design is a form of double blocking. Suppose there are t treatments of interest and t2 experimental units that are arranged in t rows which intersect with t columns. The experiment units of the rows form a homogeneous structure (block), as do the columns. As with the RCBD, each row (block) receives all t treatments, as with each column (block).

Example:

Randomly select 3 farms and divide each farm into 3 subunits with all three Varieties to be assigned to each farm.

Suppose however, that these farms have fertility gradients (high, medium and low) which may influence the response (yield). The subunits in each farm might be arranged in order of their fertility level. The treatments are assigned so that each farm (row) receives all three Varieties, but so does each fertility (column) level.

This arrangement might appear as in the following table.

This is a **Latin Square Design**. That is, we have blocked both on **rows (farms)** and **columns (fertility)** such that each treatment occurs once, and only once, within a row or column (e.g., a block).

Again, treatments are arranged in a **one-way treatment structure**. Note that the rows and columns are crossed factors and also crossed with the treatment (Variety).

4.6 Split-plot designs

(**My own short summary:** The split-plot design is a bit similar to Latin Square Designs, but not exactly the same.)

Farm, Variety 1		Farm, Variety 2		Farm, Variety 3
1, V_1		5, V_2		9, V_3
2, V_1		6, V_2		10, V_3
3, V_1		7, V_2		11, V_3
4, V_1		8, V_2		12, V_3

Figure 4.4: RCD2

Source of Variation	Degrees of Freedom	Sum of Squares
Variety	3 - 1 = 2	$SS_{Variety}$
Farm(Variety) = Whole Plot Error	3(4 - 1) = 9	$SS_{Farm(Variety)}$

Figure 4.5: RCD2_ANOVA

Consider a completely randomized design with a one-way treatment structure. In this case, suppose the treatment structure consists of 3 variteis of wheat (V_1, V_2, V_2), each planted on 4 randomly selected farms.

The ANOVA table for this experiment would appear as follows. As shown, **farms** are nested within **Variety**. As we can see, we do not need to model specifically the **variety** within each farm.

Whole Plot and Treatment Structure

However, now suppose that the researcher is also interested in the effect of two different fertilizers (F_1, F_2) on yield. The completely randomized design can be modified by ___splitting___each farm in half. In this experiment, there are two different sizes of experimental units: the large units, which are the farms and the small units, which are the Farm Halves (i.e., fertility levels).

Sub-plot Design and Treatment Structure

_Split plot vs split block design___

See 512 P.363

Farm, Variety 1		Farm, Variety 2		Farm, Variety 3	
1, V_1, F_1	1, V_1, F_2	5, V_2, F_1	5, V_2, F_2	9, V_3, F_2	9, V_3, F_1
2, V_1, F_2	2, V_1, F_1	6, V_2, F_1	6, V_2, F_2	10, V_3, F_1	10, V_3, F_2
3, V_1, F_1	3, V_1, F_2	7, V_2, F_2	7, V_2, F_1	11, V_3, F_1	11, V_3, F_2
4, V_1, F_1	4, V_1, F_2	8, V_2, F_2	8, V_2, F_1	12, V_3, F_1	12, V_3, F_2

Figure 4.6: splitting

Fertilizer	2 - 1 = 1	$SS_{Fertilizer}$
Fertilizer*Variety	(3 -1)(2 - 1) = 2	$SS_{Fertilizer*Variety}$
Subplot Error = Fertilizer*Farm(Variety)	3(4 - 1)(2 - 1) = 9	$SS_{Sub\text{-}plot\ Error}$

Figure 4.7: splitting_ANOVA

Chapter 5

Logit and Probit

5.1 Logit

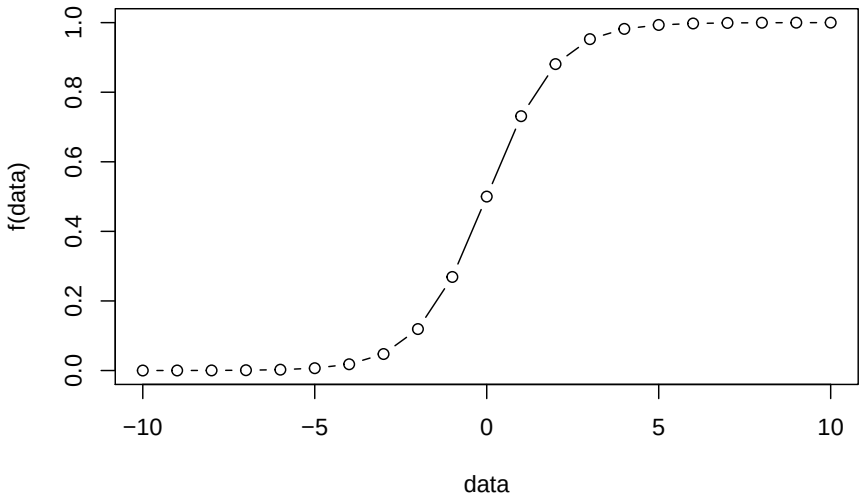
$$f(x) = log(\frac{p(y=1)}{1-p(y=1)})$$

The basic idea of logistic regression:

$$p(y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+...+\beta_n x_n)}} = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_n x_n}}$$

Thus, $\beta_0 + \beta_1 x_1 + ... + \beta_n x_n$ can be from $-\infty$ to $+\infty$, and $p(y=1)$ will be always within the range of $(0, 1)$.

```
f<-function(x){exp(x)/(1+exp(x))}
data<-seq(-10,10,1)
plot(data,f(data),type = "b")
```



We can also write the function into another format as follows:

$$log\frac{p(y=1)}{1-p(y=1)} = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$$

Thus, we know that the regression coeficients of β_i actually change the "log-odds" of the event. Of course, note that the magnitude of β_i is dependent upon the units of x_i.

The following is an example testing whether that home teams are more likely to win in NFL games. The results show that the odd of winning is the same for both home and away teams.

```
mydata = read.csv(url('https://raw.githubusercontent.com/nfl-football-ops/Big-Data-Bow
mydata$result_new<-ifelse(mydata$HomeScore>mydata$VisitorScore,1,0)
summary(mydata$result_new)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4945  1.0000  1.0000
```

```
mylogit1 = glm(result_new~1, family=binomial, data=mydata)
summary(mylogit1)
```

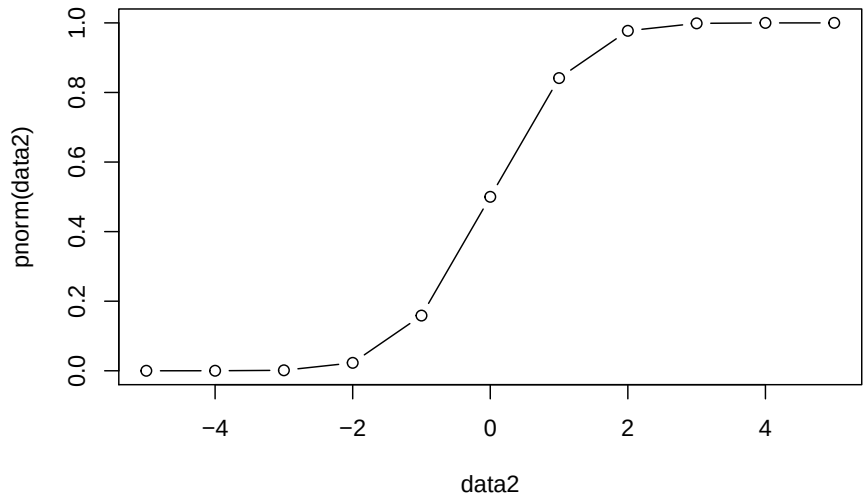
```
##
## Call:
## glm(formula = result_new ~ 1, family = binomial, data = mydata)
##
## Deviance Residuals:
```

```
##    Min      1Q  Median      3Q     Max
## -1.168  -1.168  -1.168   1.187   1.187
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.02198    0.20967  -0.105    0.917
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 126.14  on 90  degrees of freedom
## Residual deviance: 126.14  on 90  degrees of freedom
## AIC: 128.14
##
## Number of Fisher Scoring iterations: 3
```

5.2 Probit

As noted above, logit $f(x) = log(\frac{p(y=1)}{1-p(y=1)})$ provides the resulting range of $(0, 1)$. Another way to provide the same rage is through the cdf of normal distribution.The following R code is used to illusrate this process.

```
data2<-seq(-5,5,1)
plot(data2,pnorm(data2),type = "b")
```



Thus, the cdf of normal distribution can be used to indicate the probability of $p(y = 1)$.

$$\Phi(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n) = p(y = 1)$$

Similar to logit model, we can also write the inverse function of the cdf to get the function that can be from $-\infty$ to $+\infty$.

$$\beta_0 + \beta_1 x_1 + ... + \beta_n x_n = \Phi^{-1}(p(y = 1))$$

Thus, for example, if $X\beta$ = -2, based on $\Phi(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n) = p(y = 1)$ we can get that the $p(y = 1) = 0.023$.

In contrast, if $X\beta = 3$, the $p(y = 1) = 0.999$.

```
pnorm(-2)
```

```
## [1] 0.02275013
```

```
pnorm(3)
```

```
## [1] 0.9986501
```

Let's assume that there is a latent variable called Y^* such that

$$Y^* = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2)$$

You could think of Y^* as a kind of "proxy" between $X\beta + \epsilon$ and the observed $Y(1 or 0)$. Thus, we can get the following. Note that, it does not have to be zero, and can be any constant.

$$Y^* = \begin{cases} 0 & if\ y_i^* \leq 0 \\ 1 & if\ y_i^* > 0 \end{cases}$$

Thus,

$$y_i^* > 0 \Rightarrow \beta^{'} X_i + \epsilon_i > 0 \Rightarrow \epsilon_i > -\beta^{'} X_i$$

Thus, we can write it as follows. Note that $\frac{\epsilon_i}{\sigma} \sim N(0, 1)$

$$p(y = 1|x_i) = p(y_i^* > 0|x_i) = p(\epsilon_i > -\beta^{'} X_i) = p(\frac{\epsilon_i}{\sigma} > \frac{-\beta^{'} X_i}{\sigma}) = \Phi(\frac{\beta^{'} X_i}{\sigma})$$

We thus can get:

$$p(y = 0|x_i) = 1 - \Phi(\frac{\beta^{'} X_i}{\sigma})$$

For $p(y=1|x_i) = \Phi(\frac{\beta^{'} X_i}{\sigma})$, we can not really estimate both β and σ as they are in a ratio. We can assume $\sigma = 1$, then $\epsilon \sim N(0,1)$. We know y_i and x_i since we observe them. Thus, we can write it as follows.

$$p(y=1|x_i) = \Phi(\beta^{'} X_i)$$

Chapter 6

Normal distribution

6.1 Basics

μ and σ determine the center and spread of the distribution.

The empirical rule holds for all normal distributions:

(1) 68% of the area under the curve lies between $(\mu - \sigma, \mu + \sigma)$.

(2) 95% of the area under the curve lies between $(\mu - 2\sigma, \mu + 2\sigma)$.

(3) 99.7% of the area under the curve lies between $(\mu - 3\sigma, \mu + 3\sigma)$.

6.2 Confidence intervals for normal distributions

$$\bar{X} \pm Z\frac{\sigma}{\sqrt{n}}$$

where,

$\bar{X}$ is the mean

Z is the Z value (see the table below)

σ is the standard deviation

n is the number of observations

(We can see the connection between this formula and information shown in the *Basics* section.)

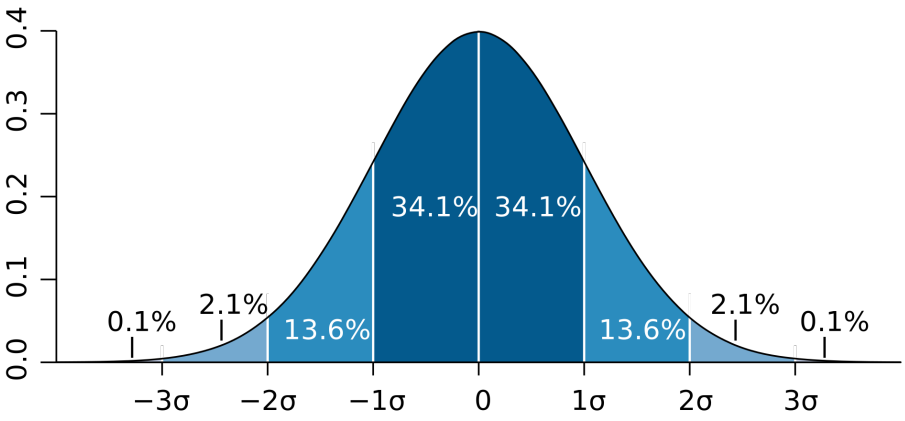


Figure 6.1: Normal

$$\begin{bmatrix} Confidence\ Levels & Z \\ 80 & 1.282 \\ 85 & 1.440 \\ 90 & 1.645 \\ 95 & 1.960 \\ 99 & 2.576 \\ 99.5 & 2.807 \\ 99.9 & 3.291 \end{bmatrix}$$

6.3 Percentile

A percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations falls.

For example, the 20th percentile is the value (or score) below which 20% of the observations may be found.

For normal distribution,

-3 σ is the 0.13th percentile (i.e., $\frac{100-99.7}{2} = 0.15$);

-2 σ is the 2.28th percentile ((i.e., $\frac{100-95}{2} = 2.50$));

-1σ is the 15.87th percentile (i.e., $\frac{100-68}{2} = 16$);

0 σ is 50th percentile.

+2 σ is the 97.72nd percentile (i.e., $100 - \frac{100-95}{2} = 100 - 2.5 = 97.50$);

+3 σ is the 99.87th percentile (i.e., $100 - \frac{100-99.70}{2} = 100 - 0.15 = 99.85$).

This is related to the 68-95-99.7 rule or the three-sigma rule.

(Note that, it is *related*, not *direct* 68-95-99.7 rule, which is about symmetric situations. See the figure above)

$$\begin{bmatrix} Percentile & Z \\ 90 & 1.282 \\ - & 1.440 \\ 95 & 1.645 \\ - & 1.960 \\ - & 2.576 \\ - & 2.807 \\ 99.9 & 3.000 \end{bmatrix}$$

Chapter 7

MLE

7.1 Basic idea of MLE

Suppose that we flip a coin, $y_i = 0$ for tails and $y_i = 1$ for heads. If we get p heads from n trials, we can get the proportion of heads is p/n, which is the sample mean. If we do not do any further calculation, this is our best guess.

Suppose that the true proablity is ρ, then we can get:

$$\mathbf{L}(y_i) = \begin{cases} \rho & y_i = 1 \\ 1 - \rho & y_i = 0 \end{cases}$$

Thus, we can also write it as follows.

$$\mathbf{L}(y_i) = \rho^{y_i}(1-\rho)^{1-y_i}$$

Thus, we can get:

$$\prod \mathbf{L}(y_i|\rho) = \rho^{\sum y_i}(1-\rho)^{\sum(1-y_i)}$$

Further, we can get a log-transformed format.

$$log(\prod \mathbf{L}(y_i|\rho)) = \sum y_i log\rho + \sum(1-y_i)log(1-\rho)$$

To maximize the log-function above, we can calculate the derivative with respect to ρ.

$$\frac{\partial log(\prod \mathbf{L}(y_i|\rho))}{\partial \rho} = \sum y_i \frac{1}{\rho} - \sum(1-y_i)\frac{1}{1-\rho}$$

Set the derivative to zero and solve for ρ, we can get

$$\sum y_i \frac{1}{\rho} - \sum (1 - y_i) \frac{1}{1 - \rho} = 0$$
$$\Rightarrow (1 - \rho) \sum y_i - \rho \sum (1 - y_i) = 0$$
$$\Rightarrow \sum y_i - \rho \sum y_i - n\rho + \rho \sum y_i = 0$$
$$\Rightarrow \sum y_i - n\rho = 0$$
$$\Rightarrow \rho = \frac{\sum y_i}{n} = \frac{p}{n}$$

Thus, we can see that the ρ maximizing the likelihood function is equal to the sample mean.

7.2 Coin flip example, probit, and logit

In the example above, we are not really trying to estimate a lot of regression coefficients. What we are doing actually is to calculate the sample mean, or intercept in the regresion sense. What does it mean? Let's use some data to explain it.

Suppose that we flip a coin 20 times and observe 8 heads. We can use the R's glm function to esimate the ρ. If the result is consistent with what we did above, we should observe that the *cdf* of the esimate of β_0 (i.e., intercept) should be equal to $8/20 = 0.4$.

```
coins<-c(rep(1,times=8),rep(0,times=12))
table(coins)
```

```
## coins
##  0  1
## 12  8
```

```
coins<-as.data.frame(coins)
```

7.2.1 Probit

```
probitresults <- glm(coins ~ 1, family = binomial(link = "probit"), data = coins)
probitresults
```

```
##
## Call:  glm(formula = coins ~ 1, family = binomial(link = "probit"),
##     data = coins)
##
## Coefficients:
## (Intercept)
##     -0.2533
```

```
##
## Degrees of Freedom: 19 Total (i.e. Null);  19 Residual
## Null Deviance:       26.92
## Residual Deviance: 26.92     AIC: 28.92
```

```
pnorm(probitresults$coefficients)
```

```
## (Intercept)
##         0.4
```

As we can see the intercept is -0.2533, and thus $\Phi(-0.2533471) = 0.4$

7.2.2 Logit

We can also use logit link to calculate the intercept as well. Recall that

$$p(y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_n x_n)}} = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_n x_n}}$$

Thus,

$$p(y=1) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$$

```
logitresults <- glm(coins ~ 1, family = binomial(link = "logit"), data = coins)
logitresults$coefficients
```

```
## (Intercept)
##  -0.4054651
```

```
exp(logitresults$coefficients)/(1+exp(logitresults$coefficients))
```

```
## (Intercept)
##         0.4
```

Note that, the defaul link for the binomial in the glm function in logit.

7.3 Further on logit

The probablity of $y = 1$ is as follows:

$$p = p(y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_n x_n)}} = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_n x_n}}$$

Thus, the likelihood function is as follows:

$$L = \prod p^{y_i}(1-p)^{1-y_i} = \prod \left(\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_n x_n)}}\right)^{y_i} \left(\frac{1}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_n x_n}}\right)^{1-y_i}$$

$$= \prod (1 + e^{-(\beta_0+\beta_1 x_1+...+\beta_n x_n)})^{-y_i} (1 + e^{\beta_0+\beta_1 x_1+...+\beta_n x_n})^{-(1-y_i)}$$

Thus, the log-likelihood is as follows:

$$logL = \sum (-y_i \cdot log(1+e^{-(\beta_0+\beta_1 x_1+...+\beta_n x_n)}) - (1-y_i) \cdot log(1+e^{\beta_0+\beta_1 x_1+...+\beta_n x_n}))$$

Typically, optimisers minimize a function, so we use negative log-likelihood as minimising that is equivalent to maximising the log-likelihood or the likelihood itself.

```r
#Source of R code: https://www.r-bloggers.com/logistic-regression/

mle.logreg = function(fmla, data)
{
  # Define the negative log likelihood function
  logl <- function(theta,x,y){
    y <- y
    x <- as.matrix(x)
    beta <- theta[1:ncol(x)]

    # Use the log-likelihood of the Bernouilli distribution, where p is
    # defined as the logistic transformation of a linear combination
    # of predictors, according to logit(p)=(x%*%beta)
    loglik <- sum(-y*log(1 + exp(-(x%*%beta))) - (1-y)*log(1 + exp(x%*%beta)))
    return(-loglik)
  }

  # Prepare the data
  outcome = rownames(attr(terms(fmla),"factors"))[1]
  dfrTmp = model.frame(data)
  x = as.matrix(model.matrix(fmla, data=dfrTmp))
  y = as.numeric(as.matrix(data[,match(outcome,colnames(data))]))

  # Define initial values for the parameters
  theta.start = rep(0,(dim(x)[2]))
  names(theta.start) = colnames(x)

  # Calculate the maximum likelihood
  mle = optim(theta.start,logl,x=x,y=y, method = 'BFGS', hessian=T)
  out = list(beta=mle$par,vcov=solve(mle$hessian),ll=2*mle$value)
}
```

```r
mydata = read.csv(url('https://stats.idre.ucla.edu/stat/data/binary.csv'))
mylogit1 = glm(admit~gre+gpa+as.factor(rank), family=binomial, data=mydata)
```

```
mydata$rank = factor(mydata$rank) #Treat rank as a categorical variable
fmla = as.formula("admit~gre+gpa+rank") #Create model formula
mylogit2 = mle.logreg(fmla, mydata) #Estimate coefficients


 print(cbind(coef(mylogit1), mylogit2$beta))
```

```
##                            [,1]         [,2]
## (Intercept)      -3.989979073 -3.772676422
## gre               0.002264426  0.001375522
## gpa               0.804037549  0.898201239
## as.factor(rank)2 -0.675442928 -0.675543009
## as.factor(rank)3 -1.340203916 -1.356554831
## as.factor(rank)4 -1.551463677 -1.563396035
```

7.4 References

http://www.columbia.edu/~so33/SusDev/Lecture_9.pdf

Chapter 8

Score, Gradient and Jacobian

8.1 Score

The score is the gradient (the vector of partial derivatives) of $logL(\theta)$, with respect to an m-dimensional parameter vector θ.

$$S(\theta) = \frac{\partial \ell}{\partial \theta}$$

Typically, they use ∇ to denote the partical derivative.

$$\nabla \ell$$

Such differentiation will generate a $m \times 1$ row vector, which indicates the sensitivity of the likelihood.

Quote from Steffen Lauritzen's slides: "Generally the solution to this equation must be calculated by iterative methods. One of the most common methods is the Newton–Raphson method and this is based on successive approximations to the solution, using Taylor's theorem to approximate the equation."

For instance, using logit link, we can get the first derivative of log likelihood logistic regression as follows. We can not really find β easily to make the equation to be 0.

$$\begin{aligned}\frac{\partial \ell}{\partial \beta} &= \sum_{i=1}^{n} x_i^T [y_i - \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}] \\ &= \sum_{i=1}^{n} x_i^T [y_i - \hat{y}_i]\end{aligned}$$

8.2 Fisher scoring

[I will come back to this later.]

https://www2.stat.duke.edu/courses/Fall00/sta216/handouts/diagnostics.pdf

https://stats.stackexchange.com/questions/176351/implement-fisher-scoring-for-linear-regression

8.3 Gradient and Jacobian

Remarks: This part discusses gradient in a more general sense.

When $f(x)$ is only in a single dimension space:

$\mathbb{R}^n \to \mathbb{R}$

$$\nabla f(x) = [\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_n}]$$

When $f(x)$ is only in a m-dimension space (i.e., Jacobian): $\mathbb{R}^n \to \mathbb{R}^{\geq}$

$$Jac(f) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \frac{\partial f_1}{\partial x_3} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_2}{\partial x_3} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \cdots & & & & \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \frac{\partial f_n}{\partial x_3} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

For instance,

$\mathbb{R}^n \to \mathbb{R}$:

$$f(x, y) = x^2 + 2y$$

$$\nabla f(x, y) = [\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}] = [2x, 2]$$

$\mathbb{R}^n \to \mathbb{R}^{\geq}$

$$f(x, y) = (x^2 + 2y, x^3)$$

$$Jac(f) = \begin{bmatrix} 2x & 2 \\ 2x^2 & 0 \end{bmatrix}$$

8.4 Hessian and Fisher Information

Hessian matrix or Hessian is a square matrix of second-order partial derivatives of a scalar-valued function, or scalar field.

$\mathbb{R}^n \to \mathbb{R}$

$$Hessian = \nabla^2(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_1 \partial x_3} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \frac{\partial^2 f}{\partial x_2 \partial x_3} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \frac{\partial^2 f}{\partial x_3 \partial x_1} & \frac{\partial^2 f}{\partial x_3 \partial x_2} & \frac{\partial^2 f}{\partial x_3^2} & \cdots & \frac{\partial^2 f}{\partial x_3 \partial x_n} \\ \cdots & & & & \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \frac{\partial^2 f}{\partial x_n \partial x_3} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

As a special case, in the context of logit:

Suppose that the log likelihood function is $\ell(\theta)$. θ is a m demension vector.

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \cdots \\ \theta_m \end{bmatrix}$$

$$Hessian = \nabla^2(\ell) = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_3} & \cdots & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_m} \\ \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_2^2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_3} & \cdots & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_m} \\ \frac{\partial^2 \ell}{\partial \theta_3 \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_3 \theta_2} & \frac{\partial^2 \ell}{\partial \theta_3^2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_3 \partial \theta_m} \\ \cdots & & & & \\ \frac{\partial^2 \ell}{\partial \theta_m \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_m \theta_2} & \frac{\partial^2 \ell}{\partial \theta_m \partial \theta_3} & \cdots & \frac{\partial^2 \ell}{\partial \theta_m \partial \theta_m} \end{bmatrix}$$

"In statistics, the observed information, or observed Fisher information, is the negative of the second derivative (the Hessian matrix) of the"log-likelihood" (the logarithm of the likelihood function). It is a sample-based version of the Fisher information." (Direct quote from Wikipedia.)

Thus, the observed information matrix:

$$-Hessian = -\nabla^2(\ell)$$

Expected (Fisher) information matrix:

$$E[-\nabla^2(\ell)]$$

Chapter 9

Canonical link function

Inspired by a Stack Exchange post, I created the following figure:

$$\frac{Paramter}{\theta} \longrightarrow \gamma^{'}(\theta) = \mu \longrightarrow \frac{Mean}{\mu} \longrightarrow g(\mu) = \eta \longrightarrow \frac{Linearpredictor}{\eta}$$

For the case of n time Bernoulli (i.e., Binomial), its canonical link function is logit. Specifically,

$$\frac{Paramter}{\theta = \beta^T x_i} \longrightarrow \gamma^{'}(\theta) = \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} \longrightarrow \frac{Mean}{\mu = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}} \longrightarrow g(\mu) = log\frac{\frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}}{1 - \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}} \longrightarrow \frac{Linearpredictor}{\eta = \beta^T x_i}$$

Thus, we can see that,

$$\theta \equiv \eta$$

The link function $g(\mu)$ relates the linear predictor $\eta = \beta^T x_i$ to the mean μ.

Remarks:

(1) Parameter is $\theta = \beta^T x_i$ (Not μ!).

(2) $\mu = p(y = 1) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$ (Not logit!).

(3) Link function (i.e., $g(\mu)$) = logit = logarithm of odds = log $\frac{Event-Happened}{Event-Not-Happened}$.

(4) $g(\mu) = log\frac{\mu}{1-\mu} = \beta^T x_i$. Thus, link function = linear predictor = log odds!

(5) Quote from the Stack Exchange post "Newton Method and Fisher scoring for finding the ML estimator coincide, these links simplify the derivation of the MLE."

(Recall, we know that μ or $p(y = 1)$ is the mean function. Recall that, n trails of coin flips, and get p heads. Thus $\mu = \frac{p}{n}$.)

Chapter 10

Ordinary Least Squares (OLS)

Suppose we have n observation, and m variables.

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & ... & x_{1m} \\ x_{21} & x_{22} & x_{23} & ... & x_{2m} \\ ... & & & & \\ x_{n1} & x_{n2} & x_{n3} & ... & x_{nm} \end{bmatrix}$$

Thus, we can write it as the following n equations.

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_m x_{1m} \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_m x_{2m} \\ y_3 &= \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + ... + \beta_m x_{3m} \\ &... \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_m x_{nm} \end{aligned}$$

We can combine all the n equations as the following one:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_m x_{im} (i \in [1, n])$$

We can further rewrite it as a matrix format as follows.

$$y = X\beta$$

Where,

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ ... \\ y_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & ... & x_{1m} \\ 1 & x_{21} & x_{22} & x_{23} & ... & x_{2m} \\ ... & & & & & \\ 1 & x_{n1} & x_{n2} & x_{n3} & ... & x_{nm} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ ... \\ \beta_m \end{bmatrix}$$

Since later we need the inverse of X, we need to make it into a square matrix.

$$X^T y = X^T X \hat{\beta} \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y$$

We can use R to implement this calculation. As we can see, there is no need to do any iterations at all, but rather just pure matrix calculation.

```
X<-matrix(rnorm(1000),ncol=2) # we define a 2 column matrix, with 500 rows
X<-cbind(1,X) # add a 1 constant
beta_true<-c(2,1,2) # True regression coefficients
beta_true<-as.matrix(beta_true)
y=X%*%beta_true+rnorm(500)

transposed_X<-t(X)
beta_hat<-solve(transposed_X%*%X)%*%transposed_X%*%y
beta_hat
```

```
##           [,1]
## [1,] 2.0261400
## [2,] 0.9868326
## [3,] 1.9929680
```

Side Notes The function of as.matrix will automatically make c(2,1,2) become the dimension of 3×1, you do not need to transpose the β.

10.1 Taylor series

$$\begin{aligned} f(x)|_a &= f(a) + \frac{f^{'}(a)}{1!}(x-a) + \frac{f^{'}(a)}{2!}(x-a)^2 + \frac{f^{''}(a)}{3!}(x-a)^3 + ... \\ &= \sum_{n=0}^{\infty} \frac{f^n(a)}{n!}(x-a)^n \end{aligned}$$

For example:

$$\begin{aligned} e^x|_{a=0} &= e^a + \frac{e^a}{1!}(x-a) + \frac{e^a}{2!}(x-a)^2 + ... + \frac{e^a}{n!}(x-a)^n \\ &= 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + ... + \frac{1}{n!}x^n \end{aligned}$$

if $x = 2$

$e^2 = 7.389056$

$e^2 \approx 1 + \frac{1}{1!}x = 1 + \frac{1}{1!}2 = 3$

$e^2 \approx 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 = 1 + \frac{1}{1!}2 + \frac{1}{2!}2 = 5$ ...

$e^2 \approx 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \frac{1}{3!}x^2 + \frac{1}{4!}x^2 + \frac{1}{5!}x^2 = 7.2666...$

10.2 References

1. Steffen Lauritzen's slides:

http://www.stats.ox.ac.uk/~steffen/teaching/bs2HT9/scoring.pdf

2. The Stack Exchange post:

https://stats.stackexchange.com/questions/40876/what-is-the-difference-between-a-link-function-and-a-canonical-li

3. Wilipedia for OLS

https://en.wikipedia.org/wiki/Ordinary_least_squares

4. Gradient and Jacobian

https://math.stackexchange.com/questions/1519367/difference-between-gradient-and-jacobian

https://www.youtube.com/watch?v=3xVMVT-2_t4

https://math.stackexchange.com/questions/661195/what-is-the-difference-between-the-gradient-and-the-directiona

5. Hessian

https://en.wikipedia.org/wiki/Hessian_matrix

6. Observed information

https://en.wikipedia.org/wiki/Observed_information

7. Fisher information

https://people.missouristate.edu/songfengzheng/Teaching/MTH541/Lecture%20notes/Fisher_info.pdf

8. Link function

https://en.wikipedia.org/wiki/Generalized_linear_model#Link_function

https://stats.stackexchange.com/questions/40876/what-is-the-difference-between-a-link-function-and-a

Chapter 11

Cholesky decomposition

11.1 Example 1

Use Cholesky decomposition to generate 1,000 trivariate normal deviates $X_1, ..., x_{1000}$ with mean $\mu = (-2, 4, 3)$ and covariance matrix

$$X = \begin{bmatrix} 2 & -1 & 0.5 \\ -1 & 4 & 1 \\ 0.5 & 1 & 5 \end{bmatrix}$$

```
Nsim = 10
means = c(-2,4,3)
N_columns = 3

# Generating random standard normal distribution numbers
Generated_numbers = matrix(rnorm(N_columns * Nsim), nrow = N_columns)

# The provided covariance matrix
cov_matrix = rbind(c(2, -1, 0.5), c(-1, 4, 1), c(0.5, 1, 5))

# Cholesky decomposition
Cholesky_decom_results = chol(cov_matrix)

# Data is transformed using the Cholesky decomposition
adjusted_data = t(Generated_numbers) %*% Cholesky_decom_results

Final_data = t(t(adjusted_data) + means)
```

```
# calculating column means
colMeans(Final_data)
```

```
## [1] -1.965932  3.609447  3.428036
```

```
# calculating column variances
apply(Final_data,2,var)
```

```
## [1] 1.253737 6.077783 4.886638
```

```
# calculating covariance matrix
cov(Final_data)
```

```
##            [,1]       [,2]      [,3]
## [1,]  1.2537367 -0.1369531 -0.526759
## [2,] -0.1369531  6.0777827  3.989427
## [3,] -0.5267590  3.9894273  4.886638
```

11.2 Example 2

AR(1) Covariance Matrix with Correlatiion Rho and Variance SigmaSq. Note that, there is only one individual or participant in this data simulation.

```
n = 10;
SigmaSq = 5;
Rho = 0.8;

V = matrix(rep(n*n,0),n,n);

for (i in 1:n)
{
  for (j in i:n)
  {
    V[i,j]=SigmaSq*Rho^(j-i)
    V[j,i]=V[i,j]
  }
}

set.seed(123)
random_normal<-rnorm(n,2,1)
#chol(V) %*% random_normal
#colSums (chol(V))
b2<-t(as.matrix(random_normal))%*%chol(V)

pi = exp(b2)/(1 + exp(b2));

y<-ifelse(pi>runif(1),1,0)
```

```
y
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    1    1    1    1    1    1    1    1    1     1
```

```
# The code above basically completes the generating job!
# The code below is to check

b = b2[2:n]
c = b2[1:(n-1)]
cor(b,c)
```

```
## [1] 0.8967058
```

```
sd(as.vector(b2))
```

```
## [1] 3.535119
```

```
# note that, you can not use var, as the mean is not zero, but rather it is 2
var(as.vector(b2))
```

```
## [1] 12.49707
```

```
#Not sure why the means are not the same ?
mean(as.vector(b2))
```

```
## [1] 10.01925
```

```
mean(random_normal)
```

```
## [1] 2.074626
```

11.3 Example 3

The following code very similar to the code shown above. However, it had only one observation. To illustrate the situation where there are more than one individual (or, participant), I did the code below.

```
n =25;   #the number of time points
m= 15;   # the number of participants or individuals, whichever ways you would like to think
SigmaSq = 5;
Rho = 0.8;

filling_numbers<-rep(n*n,0)
V = matrix(filling_numbers,n,n);

for (i in 1:n)
{
  for (j in i:n)
```

```
  {
    V[i,j]=SigmaSq*Rho^(j-i)
    V[j,i]=V[i,j]
  }
}

set.seed(2345)
random_normal<-matrix(rnorm(m*n),nrow = m)
#chol(V) %*% random_normal
#colSums (chol(V))
b2<-random_normal%*%chol(V)

pi = exp(b2)/(1 + exp(b2));

random_unfirom<-matrix(runif(m*n),nrow = m)

y<-ifelse(pi>random_unfirom,1,0)
y
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
##  [1,]    0    0    0    0    0    0    1    1    1     1     1     0     1
##  [2,]    0    1    1    1    1    1    0    1    0     0     0     0     0
##  [3,]    1    0    0    0    0    1    0    1    0     1     1     0     0
##  [4,]    1    0    0    0    0    0    0    0    1     0     0     0     0
##  [5,]    0    0    0    0    1    0    0    0    0     0     0     0     1
##  [6,]    0    0    1    1    0    0    0    1    0     0     0     0     0
##  [7,]    0    0    0    0    0    1    0    0    1     0     0     0     1
##  [8,]    1    1    1    1    0    0    0    0    0     1     1     0     0
##  [9,]    1    1    1    1    0    1    1    1    0     1     0     0     0
## [10,]    1    0    1    1    1    1    1    1    1     1     0     0     0
## [11,]    1    1    1    1    1    1    1    1    1     0     1     0     0
## [12,]    1    1    1    1    1    1    0    1    1     1     1     1     0
## [13,]    0    1    0    0    0    0    0    0    1     1     1     1     0
## [14,]    1    1    0    1    1    1    1    1    1     0     0     0     0
## [15,]    1    1    0    1    1    1    1    0    0     1     0     0     1
##       [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24]
##  [1,]     1     1     1     1     0     1     0     0     1     0     0
##  [2,]     0     1     1     1     1     1     0     0     0     1     0
##  [3,]     0     1     0     0     1     1     1     1     0     0     1
##  [4,]     1     0     0     0     0     0     1     0     1     0     0
##  [5,]     0     0     1     1     1     0     1     1     1     0     1
##  [6,]     1     1     0     0     1     0     0     1     0     1     1
##  [7,]     0     0     0     0     0     1     0     0     1     0     0
##  [8,]     0     0     0     0     0     0     0     0     0     1     1
##  [9,]     0     1     1     1     0     0     1     0     0     0     0
```

```
## [10,]     1     1     1     1     1     1     1     1     1     1     1
## [11,]     0     0     0     1     0     0     0     0     1     1     1
## [12,]     0     0     0     0     1     1     1     0     0     0     1
## [13,]     0     1     0     0     0     1     0     1     1     1     1
## [14,]     0     0     0     0     1     0     1     0     0     0     0
## [15,]     0     0     0     0     0     0     1     0     1     0     0
##       [,25]
##  [1,]     0
##  [2,]     1
##  [3,]     1
##  [4,]     0
##  [5,]     1
##  [6,]     1
##  [7,]     0
##  [8,]     1
##  [9,]     0
## [10,]     1
## [11,]     0
## [12,]     0
## [13,]     1
## [14,]     0
## [15,]     1
```

```
# The code above basically completes the generating job! The code below is to check

# The following calcuates variance
# calculate variance of each column
mean(apply(b2, 2, var))
```

```
## [1] 4.330903
```

```
# calculate variance of each row
mean(apply(b2, 1, var))
```

```
## [1] 3.568107
```

```
# The whole table
var(as.vector(b2))
```

```
## [1] 4.299165
```

```
# The following code calculates the correlation
b = b2[,2:n]
c = b2[,1:(n-1)]

collected_cor<-rep(0,m-1) #creating an empty vector to collect correlation.
for (i in 1:(m-1))
{collected_cor[i]<-cor(b[i,],c[i,])}
collected_cor
```

```
##  [1] 0.8473037 0.7065013 0.6376223 0.5481540 0.7851062 0.6576329 0.4844481
##  [8] 0.6950847 0.6731673 0.6409116 0.7966547 0.7184030 0.8001861 0.7913736
```

```
mean(collected_cor)
```

```
## [1] 0.6987535
```

```
mean(y)
```

```
## [1] 0.456
```

```
log(mean(y)/(1-mean(y)))
```

```
## [1] -0.1764564
```

```
# It will always get a value close to zero, since we set the mean to be zero when simu
```