REGRESSION & TIME SERIES MODELS

BUDDHANANDA BANERJEE

Contents

Date: Updated on February 13, 2024,
Weekly classes:MON(11:00-11:55) , TUE(08:00-08:55) , TUE(09:00-09:55) .

1. Estimation

Let $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ be the observed/ realized values of a set of i.i.d. random variables $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ where$X_i \overset{iid}{\sim} f_\theta$ for some $\theta \in \Theta$. Here a family of distributions is denoted by

$$\mathcal{F} = \{f(x|\theta)|\theta \in \Theta\} \quad \text{or} \quad \{F(x|\theta)|\theta \in \Theta\}$$

Parametric Estimation: In a parametric inference problem it is assumed that the family of the distribution is known but the particular value of the parameter is unknown. We estimate the value of the parameter θ as a function of the observations $\mathbf{x}$. The ultimate goal is to approximate the p.d.ff_θ or F_θ through the estimation of θ itself. Parametric estimation has two aspects, namely, (a) **Point estimation** and (b) **Interval estimation** .[We will learn it after Testing]

In point estimation we will learn

(a) Definition of an estimator

(b) Good properties of an estimator

(c) Methods of estimation (MME and MLE)

Statistic: A statistic is a function of random variables and it is free from any unknown parameter. Being a (measurable) function,$T(\mathbf{X})$ say , of random variables it is also a random variable.

Estimator: If the statistic $T(\mathbf{X})$ is used to estimate a parametric function $g(\theta)$ then $T(X)$ is said to be {an estimator of $g(\theta)$. And a realized value of it for $\mathbf{X} = \mathbf{x}$ i.e. $T(\mathbf{x})$ is know as **an estimate** of θ. We often abuse the notation as $g(\hat{\theta}) = T(\mathbf{x})$ and $g(\hat{\theta}) = T(\mathbf{X})$ which are understood from the context.

Definition 1.1. Unbiased estimator: An estimator $T(\mathbf{X})$ is said to be an unbiased estimator of a parametric function $g(\theta)$ if $E(T(\mathbf{X}) - g(\theta)) = 0 \ \forall \ \theta \in \Theta$.

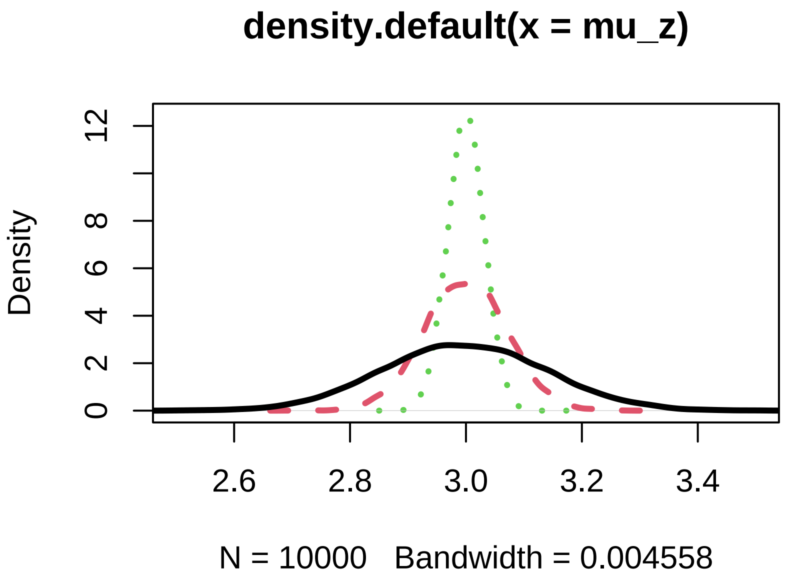
Remark 1. It does not require$T(\mathbf{x}) = g(\theta)$ to be hold or it may hold with probability zero.

Bias: The bias of an estimator $T(\mathbf{X})$ while estimating a parametric function $g(\theta)$ is $B_{g(\theta)}(T(\mathbf{X})) = E(T(\mathbf{X}) - g(\theta)) \ \forall \ \theta \in \Theta$.

Definition 1.2. Consistent estimator: An estimator T_n is said to be consistent estimator $g(\theta)$ if $T_n \xrightarrow{P} g(\theta)$ i.e.

$$\lim_{n\to\infty} P(|T_n - g(\theta)| < \epsilon) = 1 \ \forall \ \theta \in \Theta, \epsilon > 0$$

```
mu<-3
sigma<-1
nx<-50 # Sample size
ny<-200 # Sample size
nz<-1000 # Sample size
itrn<-10000
mu_x<-array(0,dim=c(itrn))
mu_y<-array(0,dim=c(itrn))
mu_z<-array(0,dim=c(itrn))
for(i in 1 : itrn){
  x<-rnorm(nx, mu, sigma)
  y<-rnorm(ny, mu, sigma)
  z<-rnorm(nz, mu, sigma)
  mu_x[i]<-mean(x)
  mu_y[i]<-mean(y)
  mu_z[i]<-mean(z)
}
par(mfrow=c(1,1))
plot(density(mu_z), col=3, lwd=3,lty=3 , xlim=c(mu-0.5,mu+0.5))
lines(density(mu_y), col=2, lwd=3,lty=2 )
lines(density(mu_x), col=1, lwd=3,lty=1 )
```



Definition 1.3. Mean squared error (MSE): The MSE of an estimator $T(\mathbf{X})$ while estimating a parametric function $g(\theta)$ is

$$MSE_{g(\theta)}(T(\mathbf{X})) = E\big[(T(\mathbf{X}) - g(\theta))^2\big] \ \forall\ \theta \in \Theta.$$

Exercise 1. Show that $MSE_{g(\theta)}(T(\mathbf{X})) = Var(T(\mathbf{X})) + B^2_{g(\theta)}(T(\mathbf{X}))$

Exercise 2. If $MSE_{g(\theta)}(T_n(\mathbf{X})) \downarrow 0$ as $n \uparrow \infty$ then show that $(T_n(\mathbf{X}))$ is a consistent estimator.

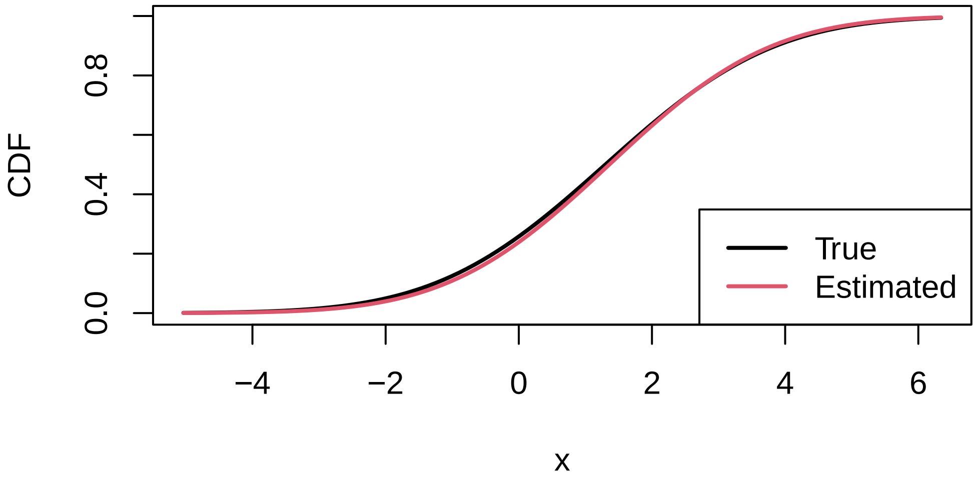
Exercise 3. Let $(X_1, X_2, \cdots, X_n)$ be i.i.d random variables with $E(X) = \mu$ and $Var(X) = \sigma^2$. and define $T_n(\mathbf{X}) = \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $S_1^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ and $S_2^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. Show that
(a) $T_n(\mathbf{X})$ is an unbiased estimator of μ.
(b) S_1^2 is an unbiased estimator of σ^2

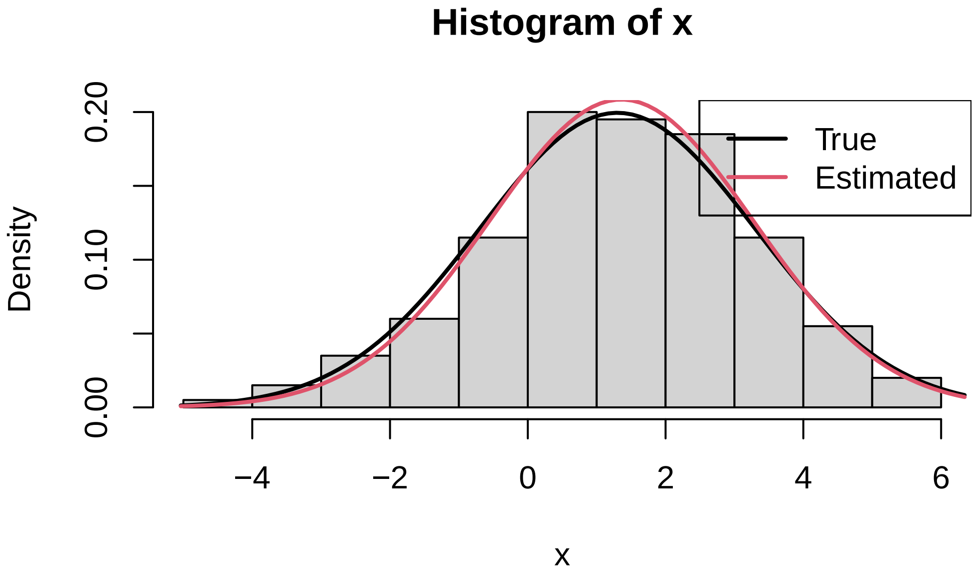
Exercise 4. Let $(X_1, X_2, \cdots, X_n) \overset{iid}{\sim} N(\mu, \sigma^2)$. Show that $MSE(S_2^2) < MSE(S_1^2)$. Note: Unbiased estimator need not have minimum MSE.

Method of Moment for Estimation (MME): Consider $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ be the observed/ realized values of a set of i.i.d. random variables $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ where $X_i \overset{iid}{\sim} f_\theta$ for some $\theta \in \Theta$. Then
Step 1: Computer theoretical moments from the p.d.f.
Step 2: Computer empirical moments from the data.
Step 3: Construct k equations if you have k unknown parameters.
Step 4: Solve the equations for the parameters.

```
# Distribution : Normal
mu<-1.3 # mean
s<- 2   # sigma
n<- 200 # sample size
x<- rnorm(n,mean = mu,sd = s) # data
xmin<- min(x) # min of data
xmax<-max(x)  # max data
l<- seq(xmin-0.5, xmax+0.5, length=100)
######### Estimation ##########
muh<-mean(x)
sh<-sd(x)
###############################
cat("True mean=", mu, "estimated mean=", muh,"\n")
cat("True sigma=", s, "estimated sigma=", sh,"\n")
###############################

plot(pnorm(q = l,mean = mu,sd = s)~l, type = 'l', col=1, lwd=2, ylab = "CDF", xlab
     ='x')
lines(pnorm(q = l,mean = muh,sd = sh)~l, type = 'l', col=2, lwd=2)
#lines(ecdf(x),col=3, lty=2)
legend("bottomright",legend =  c("True", "Estimated"), col = c(1,2), lwd = c(2,2))



hist(x,probability = T, xlab
     ='x')
lines(dnorm(x = l,mean = mu,sd = s)~l, type = 'l', col=1, lwd=2, ylab = "PDF")
lines(dnorm(x=l,mean = muh,sd = sh)~l, type = 'l', col=2, lwd=2)
legend("topright",legend =  c("True", "Estimated"), col = c(1,2), lwd = c(2,2))
```

```
## True mean= 1.3 estimated mean= 1.358959
## True sigma= 2 estimated sigma= 1.913706
```





Remark 2. We can not use MME to estimate the parameters of $C(\mu, \sigma)$, because the moments does not exists for Cauchy distribution.

Maximum Likelihood Estimator: Consider $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ be the observed/ realized values of a set of i.i.d. random variables $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ where $X_i \overset{iid}{\sim} f_\theta$ for some $\theta \in \Theta$. Then the joint p.d.f. of $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ is a function of $\mathbf{x}$ when the parameter value is fixed i.e.

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

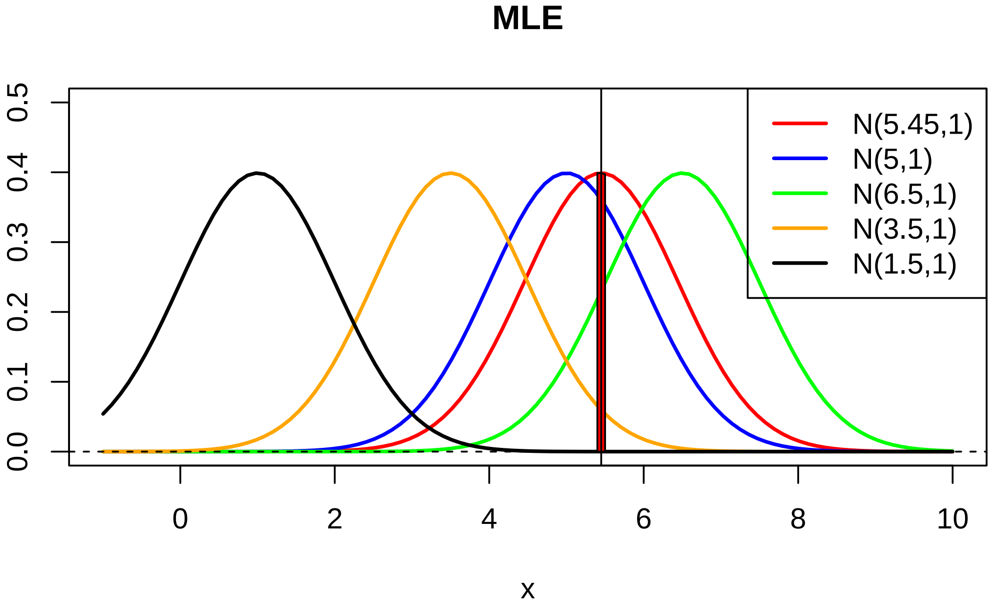
and the likelihood of a function of parameter for a given set of data $\mathbf{X} = \mathbf{x}$ i.e.

$$\ell(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i, \theta).$$

Hence the maximum likelihood estimator of θ is

$$\hat{\theta}_{mle} = \arg\max_{\theta\in\Theta} \ell(\theta|\mathbf{x}) = \arg\max_{\theta\in\Theta} \log \ell(\theta|\mathbf{x})$$

NOTE: Finding the maxima through differentiation is possible **only of** ℓ is a smoothly differentiable function w.r.t θ. Otherwise it has to be maximized by some other methods. **Differentiation is not the only way of finding maxima or minima.**



Exercise 5. Let $(X_1, X_2, \cdots, X_n) \overset{iid}{\sim} N(\mu, \sigma^2)$.

(a) Find the MME and MLE of μ and σ^2. Are they same ?

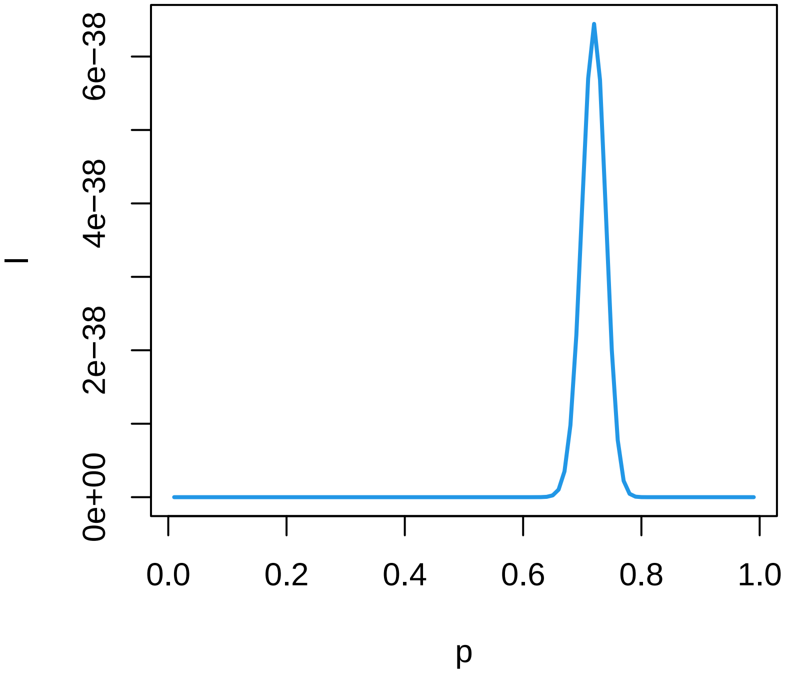
(b) Are they unbiased estimators?

Exercise 6. Let $(X_1, X_2, \cdots, X_n) \overset{iid}{\sim} Gamma(\alpha, \lambda)$.
(a) Find the MME of (α, λ)?
(b) Find MLE of (α, λ) by by an iterative method of solution.
NOTE: You may use the MME as an initial value of iteration to obtain the MLE.

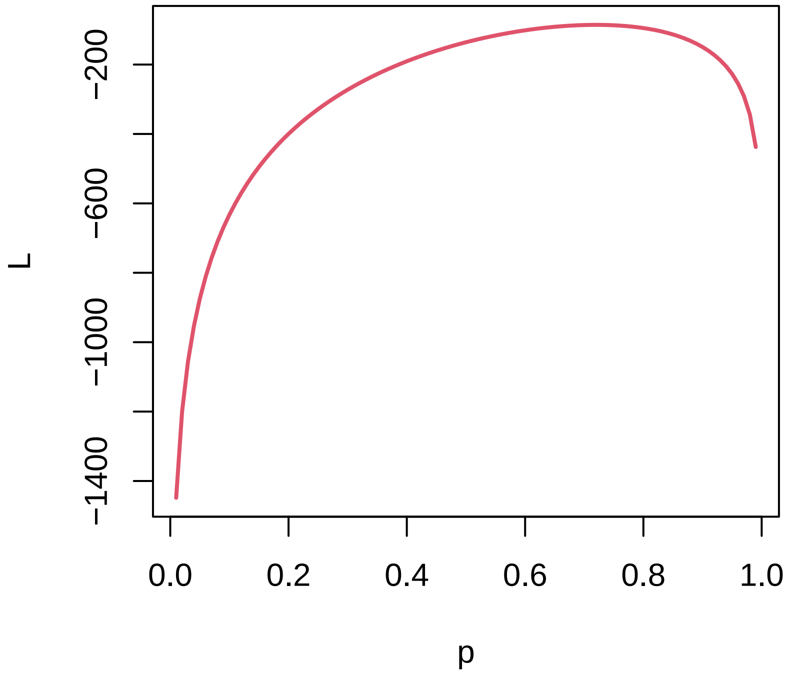
```
# MLE of binomal parameter
set.seed(12)
n<-10  # size of binomial
x<- sort(rbinom (50, n, 0.7))  # sample  given
print(x)
```

```
##  [1]  4  5  5  5  5  6  6  6  6  6  6  6  6  6  6  6  7  7  7  7  7  7  7  7  7
## [26]  7  7  7  8  8  8  8  8  8  8  8  8  8  8  8  8  8  9  9  9  9  9  9 10 10
```

```
# MLE finding
p<-seq(0.01,0.99,by = 0.01)
l<-array(0,dim=c(length(p)))
for (i in 1 : length(p)){
  l[i]<-prod(dbinom(x,n,p[i]))  # product of likelihood
}

plot(l~p, type='l', col=4, lwd=2)
```



```
mle1<-p[which(l==max(l))]
print(mle1)
## [1] 0.72
L<-array(0,dim=c(length(p)))
for (i in 1 : length(p)){
  L[i]<-sum(log(dbinom(x,n,p[i])))  #sum of log likelihood
}

plot(L~p,type='l', col=2, lwd=2)
```



```
mle2<-p[which(L==max(L))]
print(mle2)
## [1] 0.72
```

Properties of MLE:

(a) MLE need not be unique.

(b) MLE need not be an unbiased estimator.

(c) MLE is always a consistent estimator.

(d) MLE is asymptotically normally distributed up to some location and scale when some regularity condition satisfied like

(1) Range of the random variable is free from parameter.

(2) Likelihood is smoothly differentiable for up to 3rd order and corresponding expectations exists.

Definition 1.4. Interval Estimation: Consider a pair of statistic $(L(\mathbf{X}), U(\mathbf{X}))$ such that for a parameter θ ,

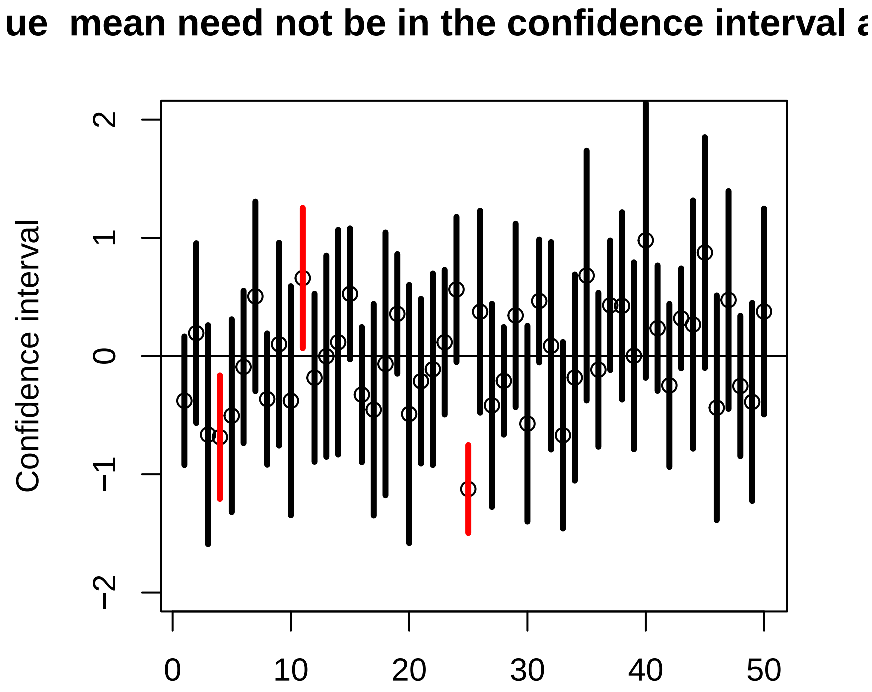
$$P_\theta(\theta \in [L(\mathbf{X}), U(\mathbf{X})]) = 1 - \alpha$$

Then a $100(1-\alpha)\%$ confidence interval of θ is considered to be $[L(\mathbf{X}), U(\mathbf{X})]$.

Example 1. If $X_{1,}X_{2,} \ldots, X_n$ are i.i.d random variables with $N(\mu, \sigma^2)$ distribution with known value of σ^2. Then a $100(1-\alpha)\%$ CI of μ is

$$\left[L(\mathbf{X}) = \overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, U(\mathbf{X}) = \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right]$$

```
set.seed(10)
N <- 50
n <- 8 # sample size
v <- matrix(c(0,0),nrow=2)
for (i in 1:N) {
  x <- rnorm(n)
  v <- cbind(v, t.test(x)$conf.int)
}
v <- v[,2:(N+1)]
plot(apply(v,2,mean), ylim=c(-2,2), ylab='Confidence interval', xlab='')
abline(0,0)
c <- apply(v,2,min)>0 | apply(v,2,max)<0
segments(1:N,v[1,],1:N,v[2,], col=c(par('fg'),'red')[c+1], lwd=3)
title(main="True  mean need not be in the confidence interval always")
```



Example 2. If $X_1, X_2, \dots, X_n$ are i.i.d random variables with $N(\mu, \sigma^2)$ distribution . Then a $100(1-\alpha)\%$ CI of μ is

$$\left[L(\mathbf{X}) = \overline{X} - \frac{\hat{\sigma_u}}{\sqrt{n}}\tau_{\alpha/2,n-1}, U(\mathbf{X}) = \overline{X} + \frac{\hat{\sigma_u}}{\sqrt{n}}\tau_{\alpha/2,n-1}\right]$$

$\hat{\sigma}_u^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ is an unbiased estimator of unknown variance and a $100(1-\alpha)\%$ CI of σ^2 is

$$\left[L(\mathbf{X}) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\chi^2_{\alpha/2,(n-1)}}, U(\mathbf{X}) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\chi^2_{1-\alpha/2,(n-1)}}\right]$$

Text:: Rice, J. A. (2006). Mathematical statistics and data analysis. Chapter : 8.1,8.2,8.3,8.4, 8.5 with examples

2. Testing of Hypothesis

Hypothesis: A hypothesis in parametric inference is a statement about the population parameter. It has two categories. A **null hypothesis** (H_0) specifies a subset Θ_0 in the parameter space Θ. If Θ_a is a singleton set then it called a **simple null**, otherwise a **composite null.** On the other hand an **alternative hypothesis** (H_1) specifies another subset $\Theta_a \subset \Theta$ which is disjoint to Θ_0.

Test Rule: A test rule is a statistical procedure, based on the distribution of the test statistic, which will reject the null hypothesis in favour of the alternative hypothesis.

Rejection Region or Critical region: A rejection Region or critical region is a subset $C \subset \mathbb{R}^n$ such that $\mathbf{X} \in C \Leftrightarrow T(\mathbf{X})$ will reject the null hypothesis.

Level-α test: For any $\alpha \in (0,1)$, a test is said to be level-α test if

$$\sup_{\theta\in\Theta_0} P_\theta(\mathbf{X} \in C) \leq \alpha.$$

Size-α test: For any $\alpha \in (0,1)$, a test is said to be size-α test if

$$\sup_{\theta\in\Theta_0} P_\theta(\mathbf{X} \in C) = \alpha.$$

Power-function: A power function is a function

$$P_\theta(\mathbf{X} \in C) : \Theta_a \to [0,1]$$

Remark 3. More than one tests with same level can be compared in terms of power functions. A test procedure with more power than the other with same level can be considered a better test.

Definition 2.1. Type-I error: The event $\mathbf{X} \in C$ when $\theta \in \Theta_0$ is known as Type-I error. **Type-II error:** The event $\mathbf{X} \in C^c$ when $\theta \in \Theta_a$ is known as Type-II error. Power is 1-P(Type-II error).

How to perform a test ??
Step1: Estimate the parameter for which the testing to be done.
Step2: Estimate the unknown parameters if any.
Step3: Construct the test statistic and obtain its value.
Step4: Obtain the exact or asymptotic distribution of the test statistic under the null hypothesis.
Step5: Depending on the alternative hypothesis (H_1) and level (α) decide the cut-off value or rejection condition.
Step6: Compare the observed value of test statistic (from Step 4) and the cut off value (from Step 5) to conclude the test. You may use **p-value** also.

Exercise 7. Let $X_1, ..., X_n \sim N(\mu,\sigma^2)$ Perform a test at size 0.05 for
(a)$H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$. when σ^2 is known
(b)$H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$. when σ^2 is unknown

(a)$H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 \neq \sigma_0^2$ when μ is unknown

Exercise 8. Let $X_1, ..., X_n \sim N(\mu_1, \sigma^2)$ (iid) and $Y_1, ..., Y_m \sim N(\mu_2, \sigma^2)$ (iid) are independent. Perform a test at size 0.05 for $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$.

Exercise 9. Let $X_1, ..., X_n \sim N(\mu_1, \sigma_1^2)$ (iid) and $Y_1, ..., Y_m \sim N(\mu_2, \sigma_2^2)$ (iid) are independent. Perform a test at size 0.05 for $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$.

```
library("TeachingDemos")
n<-10
mu_true<-10.5
sd_true<-1.2
x<-rnorm(10,mu_true,sd_true) # generate data
#########################
print(x)
##  [1] 10.404652 11.918102 13.123373 10.987410  9.613969  8.152216  8.159945
##  [8]  9.370803 11.937344  9.750913
cat("Unbiased estimate of mean =",mean(x), "\n")
## Unbiased estimate of mean = 10.34187
cat("Unbiased estimate of variance =",var(x), "\n")
## Unbiased estimate of variance = 2.729432
alpha<-0.05
## (a)H_0: mu = 10  vs H_1: mu not equal to 10  when sigma^2 = (1.2)^2 is known
za<-z.test(x,mu = 10,stdev =  sd_true ,alternative =c("two.sided"),conf.level = (1-alpha))
print(za)
##
##  One Sample z-test
##
## data:  x
## z = 0.90091, n = 10.00000, Std. Dev. = 1.20000, Std. Dev. of the sample
## mean = 0.37947, p-value = 0.3676
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   9.598119 11.085627
## sample estimates:
## mean of x
##  10.34187
##(b)H_0: mu = 10  vs H_1: mu not equal to 10  when sigma^2  is unknown
```

```
 ta<-t.test(x, mu = 10,alternative =c("two.sided"),conf.level = (1-alpha))
print(ta)
##
##  One Sample t-test
##
## data:  x
## t = 0.65438, df = 9, p-value = 0.5292
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   9.160032 11.523713
## sample estimates:
## mean of x
##  10.34187
##(c)H_0: sigma^2 = 1  vs H_0: sigma^2 neq 1  when mu is unknown
 va<-sigma.test(x, sigma = 1,alternative = "two.sided", conf.level = (1-alpha))
print((va))
##
##  One sample Chi-squared test for variance
##
## data:  x
## X-squared = 24.565, df = 9, p-value = 0.006984
## alternative hypothesis: true variance is not equal to 1
## 95 percent confidence interval:
##  1.291341 9.096795
## sample estimates:
## var of x
## 2.729432
```

3. What is a Regression Problem?

• In a broader sense the main purpose regression is prediction. In other words, it an attempt to access beyond than that has been already observed. For example

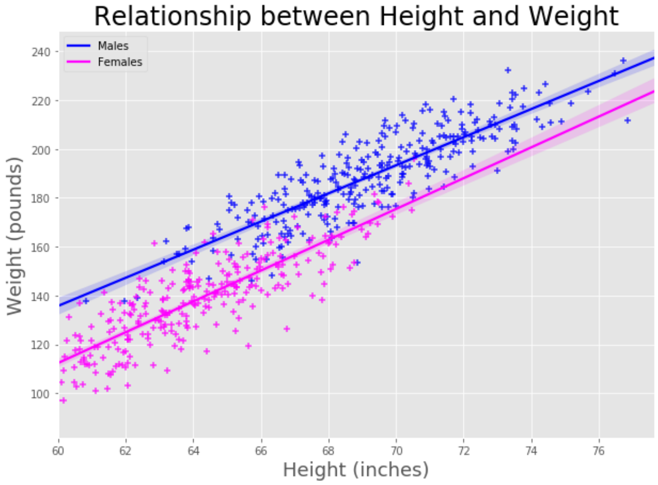
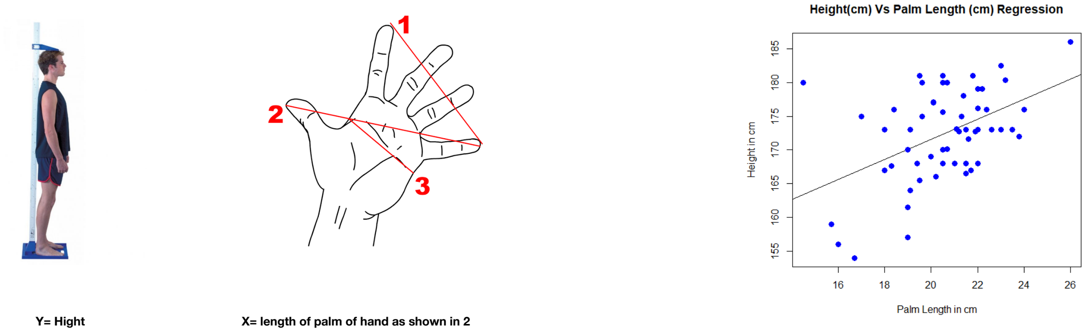


Figure 3.1. Weight vs Height



Y= Hight

X= length of palm of hand as shown in 2

$\hat{\beta}_0$ = 141.8916 $\hat{\beta}_1$ = 1.4833

Figure 3.2. Height vs Palm length

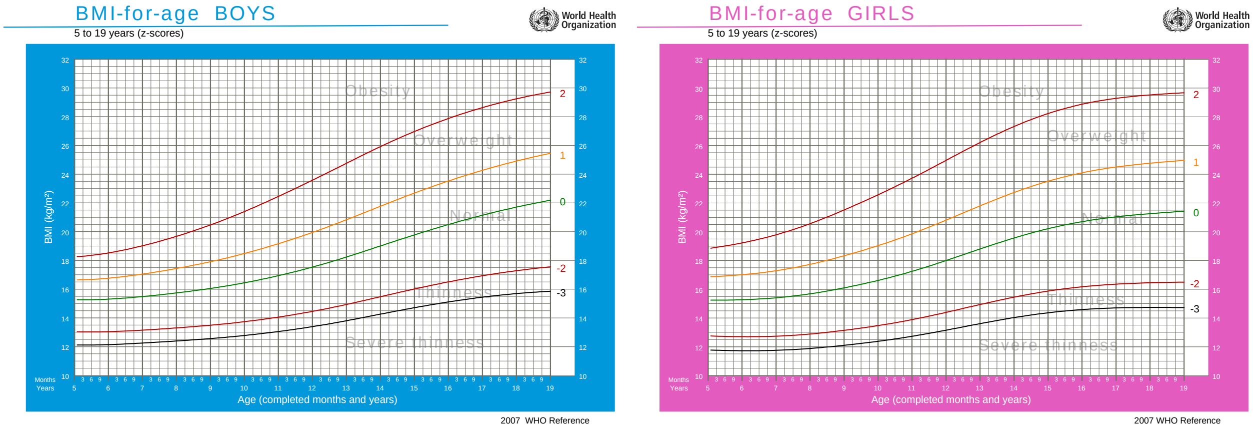


Figure 3.3. Body mass index vs Obisity

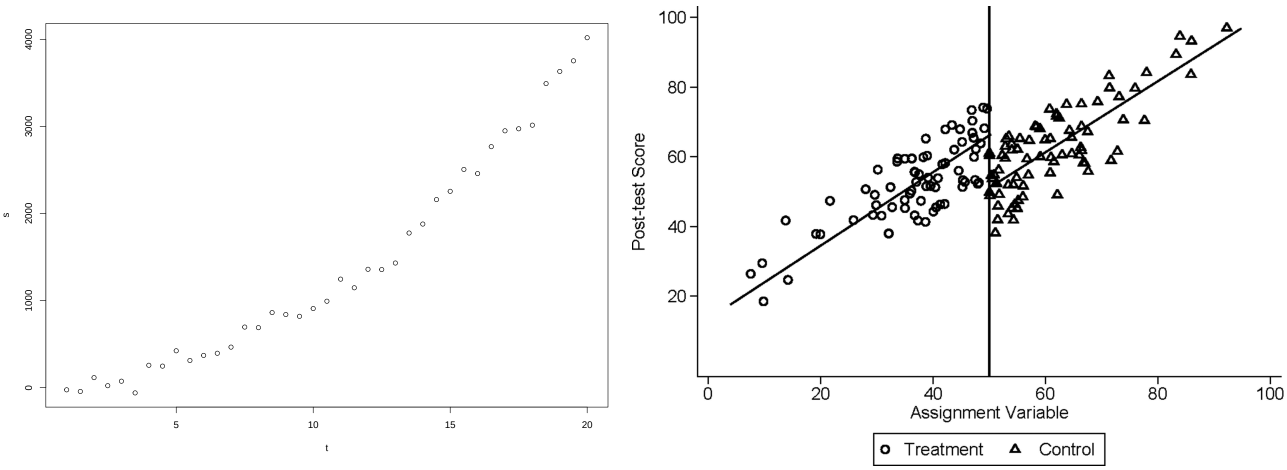


Figure 3.4. Distance measure and Treatment comparison

Definition 3.1. Let $(Y, \mathbf{X})$ be a random vector. The conditional expectation of Y given $\mathbf{X} = \mathbf{x}$, is known as the regression of Y on $\mathbf{X}$. It can be denoted as

$$\hat{y} = g(\mathbf{x}, \beta) = E(Y|\mathbf{X} = \mathbf{x})$$

▷ $g(\mathbf{x}, \beta)$ can be a line, curve, plane, surface etc. or may be unknown

▷ $\mathbf{x}$ can be stochastic or non-stochastic

▷ Y is always stochastic or a random viable

Some associated statistical problems with regression:

- ▷ Prediction of an interval for the response variable : Interval estimation
- ▷ Inportance cheking for a perticular regressior: Testing of hypotheis
- ▷ Estimability of a certain combination of regressors : Admissibility
- ▷ Efficacy of a certain combination of regressor : Contrast
- ▷ Necessity of building regression model : ANOVA
- ▷ Categorical prediction : Classification

Some associated mathematical problems with regression:

- ▷ What will be a considerable notion of error ?
- ▷ How the error space will interact will the prediction space ?
- ▷ Which subspace of regressors will be optimal for regression ?
- ▷ How to reduce the bias and variance of regression coefficients ?
- ▷ If we incorporate more regressor variables, how significantly the error can be reduced ?

We are not yet ready to answer these important questions!!!!

Results from Linear Algebra and Multivariate Analysis can help us.

4. Mathematics on $\mathbb{R}^n$

Vector Space: A vector space $\mathbf{V}$ over a real numbers $\mathbb{R}$ is a collection of vectors such that
(1) $+ : \mathbf{V} \times \mathbf{V} \to \mathbf{V}$ [closed under vector addition]
(2) $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$, for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{V}$ [associative]
(3) There exists $\mathbf{0} \in \mathbf{V}$ such that $\mathbf{0} + \mathbf{x} = \mathbf{x} + \mathbf{0} = \mathbf{x}$ for all $\mathbf{x} \in \mathbf{V}$ [identity element exists]
(4) There exists $-\mathbf{x} \in \mathbf{V}$ for each $\mathbf{x}$ such that $(-\mathbf{x}) + \mathbf{x} = \mathbf{x} + (-\mathbf{x}) = \mathbf{0}$ [inverse exists]
(5) $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ [commutative]
(6) $a \cdot (b \cdot \mathbf{x}) = (ab) \cdot \mathbf{x}$ for all $a, b \in \mathbb{R}$ and $\mathbf{x} \in \mathbf{V}$
(7) $1 \cdot \mathbf{x} = \mathbf{x}$ for all $\mathbf{x} \in \mathbf{V}$
(8) $(a + b) \cdot \mathbf{x} = (a \cdot \mathbf{x}) + (b \cdot \mathbf{x})$ for all $a, b \in \mathbb{R}$ and $\mathbf{x} \in \mathbf{V}$
(9) $a \cdot (\mathbf{x} + \mathbf{y}) = a \cdot (\mathbf{x}) + a \cdot (\mathbf{y})$

Example 3. (a) $\mathbb{R}$, (b)$\mathbb{R}^n$, (c) $\mathbb{C}^n$, (d) $\mathbb{P}_n$: all polynomials with degree less or equal to n.

Subspace: If $\mathbf{S} \subseteq \mathbf{V}$ is a vector space then $\mathbf{S}$ is a subspace of $\mathbf{V}$.

Remark 4. **How to check S is a subspace of V?**
(1) Whether $\mathbf{0} \in \mathbf{S}$?
(2) Whether $\mathbf{x} + a \cdot \mathbf{y} \in \mathbf{S}$? for all $\mathbf{x}, \mathbf{y} \in \mathbf{S}$ and $a \in \mathbb{R}$.

Example 4. (1) All lines passing through $(0, 0)$ in $\mathbb{R}^2$.
(2) All planes passing through origin in $\mathbb{R}^n$.
(3) $\mathbb{P}_5$ in $\mathbb{P}_7$

Linearly independent vectors: A set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, v_k\} \in \mathbf{V}$ are said to be linearly independent iff $\sum_{i=1}^k c_i \mathbf{v}_i = \mathbf{0} \implies c_1 = c_2 = \cdots = c_n = 0$. On the other hand if $\sum_{i=1}^k c_i \mathbf{v}_i = \mathbf{0}$ holds for some non zero $c_i \in \mathbb{R}$ the the vectors are called linearly dependent.

Span: The span of a set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_k\} \in \mathbf{V}$ is the collection

$$Sp\{\mathbf{v}_1, \mathbf{v}_2, \cdots, v_k\} = \left\{ \sum_{i=1}^{k} c_i \mathbf{v}_i | c_i \in \mathbb{R} \right\}$$

which is the all possible linear combinations of $\{\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_k\}$.

Basis & dimension: If $\{\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_k\}$ are linearly independent then it is a basis of $Sp\{\mathbf{v}_1, \mathbf{v}_2, \cdots, v_k\}$, and the dimension of $Sp\{\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_k\}$ is the number of linearly independent elements in $\{\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_k\}$.

Orthogonal Vectors: Two vectors $\mathbf{u}, \mathbf{v} \in \mathbf{V}$ are said to be orthogonal if $\mathbf{u}^T\mathbf{v} = \sum_i u_i v_i = 0$

Remark 5. (1) Basis is not unique.
(2) Elements of a basis are need not be orthogonal to each other.
(3) Linear independence need not imply orthogonality.
(4) Orthogonality implies independence.
(5) Orthogonal vectors with unit length are known as orthonormal vectors.

Orthogonal complement: If $\mathbf{S} \subseteq \mathbf{V}$ is a subspace then the orthogonal complement of $\mathbf{S}$ denoted by $\mathbf{S}^{\perp}$ is a collection

$$\mathbf{S}^{\perp} = \{\mathbf{v} | \mathbf{v} \in \mathbf{V}, \mathbf{u}^T\mathbf{v} = 0, \forall \mathbf{u} \in S\}$$

and $dim(\mathbf{S}^{\perp}) = dim(\mathbf{V}) - dim(\mathbf{S})$.

Example 5. (a) $Sp\{e_1 = (1, 0, 0, 0), e_3 = (0, 0, 1, 0)\} \perp Sp\{e_2 = (0, 1, 0, 0), e_4 = (0, 0, 0, 1)\}$
(b) $Sp\{v_1 = (1, 1, 0, 0), v_2 = (0, 1, 1, 0), v_3 = (1, 0, 1, 0), \} \perp Sp\{v_4 = (0, 0, 0, 1)\}$

Definition 4.1. If $\mathbf{S} \subseteq \mathbf{V}$ then the projection matrix of subspace $\mathbf{S}$ is P_s satisfying
(a) $P_s\mathbf{v} = \mathbf{v}$ if $\mathbf{v} \in \mathbf{S}$
(b) $P_s\mathbf{v} \in \mathbf{S}$ for all $\mathbf{v} \in \mathbf{V}$
A projection matrix P_s is an orthogonal projection matrix of subspace $\mathbf{S} \subseteq \mathbf{V}$ if $(\mathbf{I} - P_s)$ is a projection matrix of $\mathbf{S}^{\perp} \subseteq \mathbf{V}$ too.

Theorem 4.2. A projection matrix is an idempotent matrix. **[Prove it]**

Theorem 4.3. An idempotent matrix has eigen values 0 and 1. **[Prove it]**

Theorem 4.4. If $\{\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_k\}$ is an orthonormal basis of the subspace $\mathbf{S} \subseteq \mathbf{V}$ then the orthogonal projection matrix of $\mathbf{S}$ is $P_s = \sum_{i=1}^{k} v_i v_i^T$

Column Space: The column space of a matrix $A = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n]$ with columns $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n$ is

$$C(A) = Sp\{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n\} = \{A\mathbf{x} | \ \mathbf{x} \in \mathbb{R}^n.\}$$

Hence, row-space of A denoted by $\mathcal{R}(A) = C(A^T)$.

Properties:

(1) $\mathcal{C}(A : B) = \mathcal{C}(A) + \mathcal{C}(B)$

(2) $\mathcal{C}(AB) \subseteq \mathcal{C}(A)$

(3) $dim(\mathcal{C}(A)) = Rank(A)$

(4) $\mathcal{C}(AA^T) = \mathcal{C}(A) \implies Rank(AA^T) = Rank(A)$ **[Prove it]**

Definition 4.5. A square matrix $\mathbf{A} = ((A_{ij}))_{n\times n}$ is said to be

(a) **positive definite** if $\mathbf{x^T A x} > 0$ for all $\mathbf{x} \neq \mathbf{0} \in \mathbb{R}^n$.

(b) **positive semi-definite (p.s.d)** if $\mathbf{x^T A x} \geq 0$ for all $\mathbf{x} \neq \mathbf{0} \in \mathbb{R}^n$. [Also called non-negative definite (n.n.d.)]

Properties:

(a) If $\mathbf{A}$ is p.d. then $|\mathbf{A}| > 0$.

(b) If $\mathbf{A}$ is p.s.d. then $|\mathbf{A}| \geq 0$.

Generalized Inverse: A matrix G is said to be a generalize inverse of a matrix A if $AGA = A$. Usually G is denoted by A^-.

Properties:

(1)If A is $m \times n$ then A^- is $n \times m$.

(2) A^- is not unique.

(3) For any matrix A the projection matrix to $\mathcal{C}(A)$ is AA^- and **the orthogonal projection matrix to $\mathcal{C}(A)$ is** $A(A^TA)^-A^T$.

Invers of Block matrix : If a matrix is partitioned into four blocks, it can be inverted blockwise as follows:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \\ -\left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{C}\mathbf{A}^{-1} & \left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \end{bmatrix},$$

Vector differentiation:

- ▷ $\frac{\partial \mathbf{x}^\top \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^\top \left(\mathbf{A} + \mathbf{A}^\top\right)$
- ▷ $\frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^\top \mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}^\top$

Suggested Reading :

(1) Introduction to Linear Regression Analysis, 5th Edition: Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining : APPENDIX C.2.1,C.2.2.

(2) Applied regression analysis: a research tool.: John O. Rawlings Sastry G. Pantula David A. Dickey: 2 INTRODUCTION TO MATRICES

5. Statistics on $\mathbb{R}^n$

Let $\mathbf{X} = (X_1, X_2, \cdots, X_n)^T$ be a random vector with finite expectation for each of the component the we define expectation of a random vector as

$$E(\mathbf{X}) = (E(X_1), E(X_2), \cdots, E(X_n))^T.$$

Similarly if $\mathbf{Y} = ((Y_{ij}))_{m\times n}$ is a random matrix with finite expectation for each of the component the we define expectation of a random matrix as $E(\mathbf{Y}) = ((E(Y_{ij})))_{m\times n}$.

Definition 5.1. Dispersion matrix: The dispersion matrix or the variance-covariance matrix is

$$D(\mathbf{X}) = ((Cov(X_i, X_j)))_{n\times n} = E[(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^T] = Cov(\mathbf{X}, \mathbf{X}) = \Sigma$$

NOTE: (1) $Cov(\mathbf{U}_p, \mathbf{V}_q) = ((Cov(U_i, V_j)))_{p\times q}$
(2) $E(\mathbf{X} + \mathbf{b}) = E(\mathbf{X}) + \mathbf{b}$
(3) $D(\mathbf{X} + \mathbf{b}) = D(\mathbf{X})$
(4)$Cov(\mathbf{X} + \mathbf{b}, \mathbf{Y} + \mathbf{c}) = Cov(\mathbf{X}, \mathbf{Y})$

Important Results: Let $\mathbf{X}$ be a random vector with n-components such that $E(\mathbf{X}) = \mu$ and $D(\mathbf{X}) = \Sigma$ then

(1) $E(l^T\mathbf{X}) = l^T\mu$, where $l \in \mathbb{R}^n$ is a constant vector
(2) $D(l^T\mathbf{X}) = l^T\Sigma l$
(3) $E(\mathbf{A}\mathbf{X}) = \mathbf{A}\mu$, where $\mathbf{A} \in \mathbb{R}^{p\times n}$ is a constant matrix
(4) $D(\mathbf{A}\mathbf{X}) = \mathbf{A}\Sigma\mathbf{A}^T$ and $Cov(\mathbf{A}\mathbf{X}, \mathbf{B}\mathbf{X}) = \mathbf{A}\Sigma\mathbf{B}^T$
(5) If $Cov(\mathbf{U}_p, \mathbf{V}_q) = \Gamma$ then $Cov(\mathbf{A}\mathbf{U}, \mathbf{B}\mathbf{V}) = \mathbf{A}\Gamma\mathbf{B}^T$

Exercise 10. Prove (3). It will imply (1).

Exercise 11. Prove (5). It will imply (2) and (4).

Exercise 12. Show that $D(\mathbf{X})$ is a p.s.d. matrix.

Theorem 5.2. Let $\mathbf{X}$ be a random vector with n-components such that $E(\mathbf{X}) = \mu$ and $D(\mathbf{X}) = \Sigma$ then $P((\mathbf{X} - \mu) \in \mathcal{C}(\Sigma)) = 1$.

Exercise 13. Prove the above theorem.

Exercise 14. Let $\mathbf{X}$ be a random vector with n-components such that $E(\mathbf{X}) = \mu$ and $D(\mathbf{X}) = \Sigma$. Show that $E(\mathbf{X}^T A\mathbf{X}) = trace(\Sigma A) + \mu^T A\mu$

Definition 5.3. Multivariate Normal: A random vector $\mathbf{X}$ is said to follow multivariate normal $N(\mu, \Sigma)$ if it has a density

$$f(\mathbf{x}) = \frac{\exp\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\}}{(\sqrt{2\pi})^n\sqrt{|\Sigma|}}$$

for some $\mu \in \mathbf{R}^n$ and p.s.d. Σ

NOTE: If $\mathbf{X} \sim N(\mu, \Sigma)$ then $A\mathbf{X} \sim N(A\mu, A\Sigma A^T)$

Exercise 15. (X, Y) follow Bivariate normal $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ if (X, Y) has joint density function

$$f(x,y) = \frac{e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2+\left(\frac{y-\mu_y}{\sigma_y}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]}}{2\pi\sigma_x, \sigma_y\sqrt{1-\rho^2}}$$

Exercise 16. If (X, Y) follow Bivariate normal $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ the show that $Y|X = x$ follows $N(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x), (1-\rho^2)\sigma_y^2)$

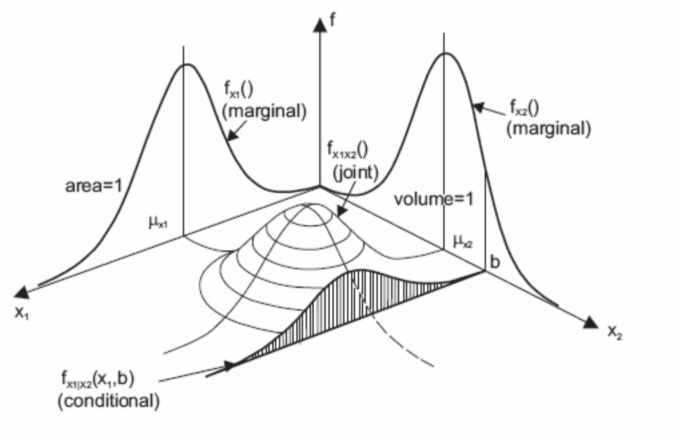


FIGURE 5.1. Conditional Normal

Exercise 17. Perform a large sample test for $H_0 : \rho = 0$ Vs $H_1 : \rho \neq 0$.

Definition 5.4. Chi-Squared distribution: If $\mathbf{X} \sim N(\mu, \mathbf{I}_n)$ then $\mathbf{X}^T\mathbf{X}$ is said to follow Chi-squared distribution with degrees of freedom (d.f.) n and non-centrality parameter (n.c.p) $\mu^T\mu$.

Exercise 18. If $\mathbf{X} \sim N(\mu, \mathbf{I}_n)$, show that $E(\mathbf{X}^T\mathbf{X}) = n + \mu^T\mu$

Theorem 5.5. If $\mathbf{X} \sim N(\mu, \mathbf{I}_n)$ then $\mathbf{X}^T A\mathbf{X}$ has Chi-squared distribution iff A is idempotent. Moreover $\mathbf{X}^T A\mathbf{X} \sim \chi^2_{df=Rank(A), ncp=\mu^T A\mu}$ [Proof is out of syllabus]

Corollary 1. *If A_1 and A_2 are symmetric and idempotent matrices such that $Q = A_1 - A_2$ be a p.s.d. matrix then $\mathbf{X}^T Q\mathbf{X}$ and $\mathbf{X}^T A_2\mathbf{X}$ are independently distributed.*

Corollary 2. *If A is symmetric and $CA = \mathbf{0}$ then $\mathbf{X}^T A\mathbf{X}$ and $C\mathbf{X}$ are independently distributed.*

Theorem 5.6. Cochran's Theorem: Let $\mathbf{X} \sim N(\mu, \mathbf{I}_n)$ and $\mathbf{X}^T A\mathbf{X} \equiv \sum_{i=1}^{k} \mathbf{X}^T A_i\mathbf{X}$ where A_is are symmetric and A is an idempotent matrix. Then $\mathbf{X}^T A_i\mathbf{X} \sim \chi^2_{Rank(A_i), \mu^T A_i\mu}$ and they are independent. [Proof is out of syllabus]

Exercise 19. Construction of t-test and F-test: Let $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$. Define $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$
(1) Find the distribution of $\bar{X}$ and S^2.
(2) Show that they are independently distribute.
(3) Construct t-statistic from here.
(4) Construct F-statistic from here too.

Suggested Reading :

(1) Introduction to Linear Regression Analysis, 5th Edition: Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining : APPENDIX C.1,C.2.3, C.2.4
(2) Linear Algebra and Linear Models: Ravindra B. Bapat : Chapter 8 Tests of Linear Hypotheses 8.1,8.2

6. Simple linear regression

Consider a data set $D = \{(x_i, y_i) | x_i \in \mathbb{R}, y_i \in \mathbb{R}, \ \forall i = 1, 2, \cdots, n\}$ where x_is are non stochastic but y_i are stochastic and realized values of random variable Y_is respectively. If the relation between the **response variable** y and the **regressor variable** x is linear in parameter then it is called a **simple linear regression model.** For example

$$y = \beta_0 + \beta_1 x + \epsilon$$
$$y = \beta_0 + \beta_1 e^x + \epsilon$$

both are linear in parameter and hence simple linear regression model. But

$$y = \frac{1}{\beta_0 + \beta_1 x} + \epsilon$$
$$y = \Phi(\beta_0 + \beta_1 e^x + \epsilon)$$

are not linear models.

Definition 6.1. Gauss-Markov model: $y_i = \beta_0 + \beta_1 x + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Here $\beta_0 \in \mathbb{R}, \beta_i \in \mathbb{R}, \sigma > 0$ are **unknown model parameters**. Here $E(y_i) = \beta_0 + \beta_1 x_i$ and $Var(y_i) = \sigma^2$, hence

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \ \ \forall i = 1, 2, \cdots, n. \tag{6.1}$$

Estimation of model parameters: The **least squared** condition to estimate the model parameters is to minimize

$$S(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2. \tag{6.2}$$

If $(\hat{\beta}_0, \hat{\beta}_1)$ minimizes $S(\beta_0, \beta_1)$ then their values can be obtained by solving the **normal equations**

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0 \quad \Longrightarrow \quad n\hat{\beta}_0 + \hat{\beta}_1 \sum_i x_i = \sum_i y_i \tag{6.3}$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0 \quad \Longrightarrow \quad \hat{\beta}_0 \sum_i x_i + \hat{\beta}_1 \sum_i x_i^2 = \sum_i y_i x_i \tag{6.4}$$

NOTE: (1)Defining $S_{xy} = \sum_i (y_i - \bar{y})(x_i - \bar{x})$ we have the solutions as

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

(2) We have never used the normality assumption for this estimation i.e. the estimators will be the same even though ϵ_i does not follow normal distribution.

(3) **Prediction or regression line:** For any x such as old x_is or some x_{new} the prediction line is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$$

Definition 6.2. The **prediction error or residual** is defined as $e_i = y_i - \hat{y}_i$ and hence the residual sum of square residual (SSR) is

$$\sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

We estimate σ^2 by $\hat{\sigma}^2 = \frac{SSR}{n-2} \equiv MSR$, mean sum of square residual.

Definition 6.3. Linear estimator: If an estimator $T(\mathbf{y})$ can be expressed as a linear combination of $\mathbf{y}$ with non random coefficients i.e. $T(\mathbf{y}) = \sum_i \alpha_i y_i$ then $T(\mathbf{y})$ is called a linear estimator.

Remark. Note that $\hat{\beta_0}$ and $\hat{\beta_1}$ are linear estimators and SSR is a quadratic form of $\mathbf{y}$.

- ▷ Show that $\hat{\beta}_1$ is a linear estimator of β_1.
- ▷ Show that $\hat{\beta}_1$ is an unbiased estimator of β_1.
- ▷ Show that $\hat{\beta_1} \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$
- ▷ Show that $\hat{\beta}_0$ is a linear estimator of β_0.
- ▷ Show that $\hat{\beta}_0$ is an unbiased estimator of β_0.
- ▷ Show that $\hat{\beta_0} \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$
- ▷ Show that $MSR = \frac{SSR}{n-2}$ is an unbiased estimator of σ^2.
- ▷ Show that $\sum_i y_i = \sum_i \hat{y}_i$
- ▷ Show that the regression line passes through $(\bar{x}, \bar{y})$.
- ▷ Find the 95% confidence interval of β_0 and β_1.
- ▷ Test that a regression line passes through origin.
- ▷ Test that a regression line is horizontal.
- ▷ Find the prediction value, and 95% prediction interval for some new x_0.
- ▷ Obtain the maximum likelihood estimators of the model parameters $\beta_0, \beta_1, \sigma^2$. Are they same as LS estimators?

Suggested Reading :

(1) Introduction to Linear Regression Analysis, 5th Edition: Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining : 2. SIMPLE LINEAR REGRESSION

(2) Chatterjee, S., & Hadi, A. S. (2013). Regression analysis by example. John Wiley & Sons. 2: Simple linear regression

Practice problems :

- ▷ Introduction to Linear Regression Analysis, 5th Edition: Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining : Examples : 2.1, 2.2, 2.3, 2.4, 2.5, 2.6,2.7, 2.9.

7. Multiple linear regression

It is a natural extension when there are more than one regressor variables in the model. The model is written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon = \mathbf{x}^T\beta + \epsilon, \tag{7.1}$$

where $\epsilon \sim N(0, \sigma^2)$ and $\mathbf{x} = (1, x_1, x_2, \cdots, x_k)^T$, $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_k)^T$. It is trivial to notice from equation (7.1) that $E(y|\mathbf{x}) = \mathbf{x}^T\beta$ is a **hyper plane** where as the same equation represents a **straight line in simple linear regression.** When We have more than one observations from the above model then we represent then the ith observation can be represented as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i = \mathbf{x}_i^T\beta + \epsilon_i,$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ and $\mathbf{x}_i = (1, x_{1i}, x_{2i}, \cdots, x_{ki})^T$, $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_k)^T$. For n such observation we use matrix notation to represent it as follows,

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{7.2}$$

where, $\mathbf{Y} = (y_1, y_2, \cdots y_n)^T$, $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_k)^T$, $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ and $\epsilon = (\epsilon_1, \epsilon_2, \cdots \epsilon_n) \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$. **Hence there are $k+2$ unknown model parameters, $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_k)^T$ and $\sigma^2 > 0$, which are to be estimated where,**

$$\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}_n) \tag{7.3}$$

Least Squared Estimation: The least square condition to be minimized to estimate β, σ^2 is

$$S(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{Y}^T\mathbf{Y} - 2\beta^T\mathbf{X}^T\mathbf{Y} + \beta^T\mathbf{X}^T\mathbf{X}\beta \tag{7.4}$$

If $\hat{\beta}$ minimizes the least square condition then it satisfies the normal equations

$$\begin{aligned} & \frac{\partial S(\beta)}{\partial \beta}|_{\beta=\hat{\beta}} = \mathbf{0} \\ \implies & -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{0} \\ \implies & \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{y} \end{aligned} \tag{7.5}$$

So, $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{y} = P_{\mathbf{X}}\mathbf{y}$ where$P_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T$ is the orthogonal projection matrix of the column space of $\mathbf{X}$ i.e. $\mathcal{C}(\mathbf{X})$. It means $\hat{\mathbf{y}} \in \mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{X}^T\mathbf{X})$. Hence the estimated error in prediction

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - P_{\mathbf{X}})\mathbf{y} \in \mathcal{C}(\mathbf{X})^{\perp} = \mathcal{C}(\mathbf{X}^T\mathbf{X})^{\perp}.$$

Definition 1. Linear unbiased estimator (LUE): A linear estimator $l^T\mathbf{y} = \sum_i l_i y_i$ is said to be linear unbiased estimator (LUE) of $p^T\beta$ if $E(l^T\mathbf{y}) = p^T\beta$ for all $\beta \in \mathbb{R}^{k+1}$.

Definition 2. Linear zero function (LZF): A linear estimator $l^T\mathbf{y} = \sum_i l_i y_i$ is said to be linear zero function (LZF) if $E(l^T\mathbf{y}) = 0$ for all $\beta \in \mathbb{R}^{k+1}$

Definition 3. A linear parametric function $p^T\beta$ is said to be **estimable** if it has a LUE i.e. there exists $l^T\mathbf{y}$ such that $E(l^T\mathbf{y}) = p^T\beta$ for all $\beta \in \mathbb{R}^{k+1}$.

Definition 4. **BLUE:** The best linear unbiased estimator (BLUE) of a linear parametric function $p^T\beta$ is a LUE with minimum variance.

Theorem 1. *A linear function is BLUE of its expectation iff it is uncorrelated with all LZF.*

Corollary 3. *If $l^T\mathbf{y}$ is an LUE of $p^T\beta$ then the blue of $l^T P_{\mathbf{X}}\mathbf{y}$ is the BLUE of $p^T\beta$*

NOTE: If you know a LUE of a parametric function then you can get the BLUE out of it.
ANOVA: To test $H_0 : \beta_R = (\beta_1, \cdots, \beta_k) = \mathbf{0}$ vs $H_1 : (\beta_1, \cdots, \beta_k) = \mathbf{0}$ we perform the ANOVA as

$$SST = SSModel + SSRes$$

where, $SST = \mathbf{Y}^T(\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{Y} \sim \sigma^2\chi^2(df = n-1, ncp = \lambda)$
$SSModel = \mathbf{Y}^T(P_X - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{Y} \sim \sigma^2\chi^2(df = k, ncp = \lambda)$
$SSRes = \mathbf{Y}^T(\mathbf{I}_n - P_X)\mathbf{Y} \sim \sigma^2\chi^2(df = n-k-1, ncp = 0)$
under $H_0 : \lambda = 0$, otherwise $\lambda = \beta_R^T\mathbf{X}_c^T\mathbf{X}_c\beta_R$ where $\mathbf{X}_c$ is the centred regressor variables for β_R. Hence, by Cochran's theorem we have

$$\frac{SSModel/k}{SSRes/(n-k-1)} \sim F_{k,(n-k-1),ncp_1=\lambda}$$

It is a right tailed test because $\lambda = 0$ under H_0.

Coefficient of determination: For a given set of data $(\mathbf{y}, \mathbf{X})$ we first fit linear model under the standard assumptions. To measure how good the model is to explain dependencies between regression and independent variables we use (a) **Coefficient of determination** (R^2):

$$\begin{aligned} R^2 &= \frac{\text{Variation in Y explained by the model}}{\text{Total variation in Y}} \\ &= \frac{\sum_{i=1}^n(\hat{Y}_i - \bar{\hat{Y}})^2}{\sum_{i=1}^n(Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^n(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n(Y_i - \bar{Y})^2} \\ &= \frac{SSModel}{SSTotal} \\ &= 1 - \frac{SSError}{SSTotal} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n(Y_i - \bar{Y})^2} \end{aligned}$$

NOTE:

(1) $R^2 \in (0, 1)$ (7.6)

(2) R^2 is an increasing function of the number of variables. (7.7)

(3) Higher value of R^2 is an indicator of the better model. So we modify it to another measure of model adequacy checking. (7.8)

(b)**Adjusted R^2:**

$$\begin{aligned} R^2_{adjusted} &= 1 - \frac{\sum_{i=1}^{n} e_i^2/df}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2/df} \\ &= 1 - \frac{\sum_{i=1}^{n} e_i^2/(n-k-1)}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)} < R^2 \end{aligned} \tag{7.10}$$

Suggested Reading :

(1) Introduction to Linear Regression Analysis, 5th Edition: Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining : 3. MULTIPLE LINEAR REGRESSION

(2) Chatterjee, S., & Hadi, A. S. (2013). Regression analysis by example. John Wiley & Sons. 3 : MULTIPLE LINEAR REGRESSION

Practice problems :

▷ Introduction to Linear Regression Analysis, 5th Edition: Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining : Examples : 3.2, 3.3, 3.4, 3.6, 3.7,3.8,3.9.

8. Polynomial regression

We can extend the idea of multiple linear regression to polynomial regression. In polynomial regression we consider higher degrees of the components $\mathbf{x}$ but it is linear in parameters. Hence, it is a linear model too. For example,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \quad \text{for single regressor} \tag{8.1}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon \quad \text{for multiple regressors} \tag{8.2}$$

In general for k-degree polynomial for single regressor , as modeled in equation (8.1), we can write in matrix notation

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{8.3}$$

where, $\mathbf{Y} = (y_1, y_2, \cdots y_n)^T$, $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_k)^T$, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)^T$ with $\mathbf{x}_i = (1, x_i, x_i^2, \cdots, x_i^k)^T$ and $\epsilon = (\epsilon_1, \epsilon_2, \cdots \epsilon_n)^T \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

Hence there are $k+2$ unknown model parameters, $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_k)^T$ and $\sigma^2 > 0$, which are to be estimated when,

$$\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n) \tag{8.4}$$

So the least square estimate of β will be $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{y}$.

Some problems related to polynomial regression:

1. **Finding order of the model:** For that we can go by either (a) forward selection or (b) backward elimination.
2. **Extrapolation:** Beyond the range of the data the prediction may be more erroneous.
3. **Ill-conditioning:** The matrix $(\mathbf{X}^T\mathbf{X})$ may be computationally singular, specially when the magnitude of ith regressor is closed to zero.
4. **Hierarchy:** A model with all lower order terms of the highest degree is called hierarchical model. For example equation(\ref{1vpr}) and (\ref{mvpr}). But regression model need not be so. Instead of equation (\ref{mvpr}) , in design of experiment $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_5 x_1 x_2 + \epsilon$ might be sufficient considering only individual effect and interaction effect.

Orthogonal Polynomial: For a single variable polynomial regression if we increase one more degree then we need to re-estimate all the coefficients including the new one each time. To overcome this problem we introduce the notion of orthogonal polynomial. For a given set of input data $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ a set of polynomials $\{P_0, P_1, P_2, \cdots, P_k\}$ are said to be **orthogonal polynomials** if

$$P_0(x_i) = 1 \text{ and } \sum_{i=1}^{n} P_j(x_i) P_k(x_i) = 0 \ \forall j \neq k \tag{8.5}$$

Now the model will look as follows

$$y_i = \sum_{j=0}^{k} \alpha_j P_j(x_i) + \epsilon_i \ \forall i = 1, 2, \cdots n \text{ or denoting } \mathbf{Z} = ((P_j(x_i)))_{n \times (k+1)} \text{ we get} \mathbf{Y} = \mathbf{Z}\alpha + \epsilon$$

Suggested Reading :

(1) Introduction to Linear Regression Analysis, 5th Edition: Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining : 7. POLYNOMIAL REGRESSION MODELS (7.1, 7.4, 7.5)

Practice problems :

▷ Introduction to Linear Regression Analysis, 5th Edition: Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining : Examples : 7.1,7.5

Department of Mathematics, IIT Kgaragpur
E-mail address: bbanerjee@maths.iitkgp.ac.in