REGRESSION & TIME SERIES MODELS

Buddhananda Banerjee
Department of Mathematics
&
Centre for Artificial Intelligence
Indian Institute of Technology Kharagpur
e-mail:bbanerjee@maths.iitkgp.ac.in

Last updated on August 26, 2019

Contents

Part 1. Recap

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

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

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.f f_θ or F_θ through the estimation of θ itself. Parametric estimation has two aspects, namely, (a) **Point estimation** and (b) **Interval estimation** .

Definition 1. 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.

Definition 2. 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 3. 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 4. It does not require$T(\mathbf{x}) = g(\theta)$ to be hold or it may hold with probability zero.

Definition 5. 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 6. Asymptotically unbiased estimator: Denoting $T_n = T(X_1, X_2, \cdots, X_n)$ an estimator T_n is said to be asymptotically unbiased of $g(\theta)$ if

$$\lim_{n\to\infty} B_{g(\theta)}(T_n) = \lim_{n\to\infty} E(T_n - g(\theta)) = 0$$

Definition 7. Consistent estimator: An estimator T_n is said to be consistent estimator $g(\theta)$ if $T_n \xrightarrow{P} g(\theta)$ i.e.

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

Definition 8. Mean squared error (MSE): The MSE of an estimator $T(\mathbf{X})$ while estimating a parametric function $g(\theta)$ is

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

Remark 9. $MSE_{g(\theta)}(T(\mathbf{X})) = Var(T(\mathbf{X})) + B^2_{g(\theta)}(T(\mathbf{X}))$

Remark 10. If $MSE_{g(\theta)}(T_n(\mathbf{X})) \downarrow 0$ as $n \uparrow \infty$ then $(T_n(\mathbf{X}))$ is a consistent estimator.

Remark 11. Asymptotic unbiasedness and consistency are large sample properties and both are based on L_1 norm. . MSE is defined based on L_2 norm.

Remark 12. 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 $S_2^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. Then
(a) $T_n(\mathbf{X})$ is an unbiased estimator of μ.
(b) S_1^2 is a biased estimator of σ^2
(c) S_2^2 is an asymptotically unbiased estimator of σ^2

Remark 13. 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.

Definition 14. 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.

Remark 15. We can not use MME to estimate the parameters of $C(\mu, \sigma)$, because the moments does not exists for Cauchy distribution.

Definition 16. 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 17. 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 18. 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.

Propertied 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 19. 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 20. 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

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

Example 21. If $X_1, X_2, \ldots, X_n$ 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

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

2. Testing of Hypothesis

Definition 22. 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.

Definition 23. 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.

Definition 24. 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.

Definition 25. Level-α test: For any $\alpha \in (0,1)$, a test is said to be level-α test if

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

Definition 26. Size-α test: For any $\alpha \in (0,1)$, a test is said to be size-α test if

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

Definition 27. Power-function: A power function is a function

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

Remark 28. 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 29. Type-I error: The event $\mathbf{X} \in C$ when $\theta \in \Theta_0$ is known as Type-I error.

Definition 30. 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).

Lemma 31. ***Neyman-Pearson Lemma (1933):*** *To test $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$ reject H_0 in favour of H_1 at level/ size α if*

$$\Lambda(\mathbf{x}) = \frac{f(\mathbf{x}|\theta_0)}{f(\mathbf{x}|\theta_1)} \leq \xi \quad \textit{such that} \quad P_{\theta_0}(\Lambda(\mathbf{X}) \leq \xi) = \alpha$$

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 32. 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

```
library("TeachingDemos")
n<-10
mu_true<-10.5
sd_true<-1.2
x<-rnorm(10,mu_true,sd_true) # generate data
#########################
print(x)
##  [1] 11.207673 10.461317  9.396939 11.300244 11.895166 10.858552 10.481322
##  [8]  8.430529 10.267865 11.659670
cat("Unbiased estimate of mean =",mean(x), "\n")
## Unbiased estimate of mean = 10.59593
cat("Unbiased estimate of variance =",var(x), "\n")
## Unbiased estimate of variance = 1.113804
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 = 1.5704, n = 10.00000, Std. Dev. = 1.20000, Std. Dev. of the
## sample mean = 0.37947, p-value = 0.1163
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   9.852174 11.339682
## sample estimates:
## mean of x
##  10.59593
##(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 = 1.7856, df = 9, p-value = 0.1078
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   9.840962 11.350893
## sample estimates:
## mean of x
##  10.59593
##(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 = 10.024, df = 9, p-value = 0.6971
## alternative hypothesis: true variance is not equal to 1
## 95 percent confidence interval:
```

```
##  0.5269601 3.7121456
## sample estimates:
## var of x
## 1.113804
```

Exercise 33. 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 34. 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 $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Part 2. Preliminaries

3. Linear Algebra

Definition 35. 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 36. (a) $\mathbb{R}$, (b)$\mathbb{R}^n$, (c) $\mathbb{C}^n$, (d) $\mathbb{P}_n$: all polynomials with degree less or equal to n.

Definition 37. Subspace: If $\mathbf{S} \subseteq \mathbf{V}$ is a vector space then $\mathbf{S}$ is a subspace of $\mathbf{V}$.

Remark 38. **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 39. (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 $\mathbb{P}_7$

Definition 40. 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.

Definition 41. 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 $\{\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_k\}$.

Definition 42. 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 $\{\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_k\}$.

Definition 43. 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 44. (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.

Definition 45. 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 46. (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 47. 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 48. *A projection matrix is an idempotent matrix.* ***[Prove it]***

Theorem 49. *An idempotent matrix has eigen values 0 and 1.* ***[Prove it]***

Theorem 50. *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* $P_s = \sum_{i=1}^{k} v_i v_i^T$

Definition 51. 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

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

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

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]**

(5) If A has n-rows then $dim(\mathcal{C}(A)^{\perp}) = n - Rank(A)$

(6) $Rank(AB) \leq \min\{Rank(A), Rank(B)\}$

(7) $Rank(A + B) \leq Rank(A) + Rank(B)$

Definition 52. 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 $|\mathbf{A}| \geq 0$.

Definition 53. **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** $A(A^TA)^-A^T$.

Suggested reading from

Linear Algebra and Linear Models By Ravindra B. Bapat

Linear Algebra and its Application by D C Lay

4. Multivariate Analysis

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

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

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 55. Prove (3). It will imply (1).

Exercise 56. Prove (5). It will imply (2) and (4).

Exercise 57. Show that $D(\mathbf{X})$ is a p.s.d. matrix.

Definition 58. 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 59. Prove the above theorem.

Theorem 60. *Let $\mathbf{X}$ be a random vector with n-components such that $E(\mathbf{X}) = \mu$ and $D(\mathbf{X}) = \Sigma$ with $Rank(\Sigma) = r \leq n$. Also assume that $\Sigma = BB^T$ where B is a $(n \times r)$ matrix and C is a left inverse of B i.e. $CB = \mathbf{I}_r$. Define $\mathbf{Y} = C(\mathbf{X} - \mu)$. Show that*
(i) $E(\mathbf{Y}) = \mathbf{0}$
(ii) $D(\mathbf{Y}) = \mathbf{I}_r$
(iii) $\mathbf{X} = \mu + B\mathbf{Y}$ with probability 1.

Exercise 61. 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 62. 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:

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

(2) If $\mathbf{X} \sim N(\mu, \Sigma)$ then there exists B and its left inverse C such that

$\mathbf{Y} = C(\mathbf{X} - \mu) \sim N(\mathbf{0}, \mathbf{I}_r)$ and $\mathbf{X} = \mu + B\mathbf{Y}$ with probability one.

Definition 63. 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 64. If $\mathbf{X} \sim N(\mu, \mathbf{I}_n)$, show that $E(\mathbf{X}^T\mathbf{X}) = n + \mu^T\mu$

Theorem 65. *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 $\mathbf{X}^T A\mathbf{X} \sim \chi^2_{df=Rank(A), ncp=\mu^T A\mu}$*

Corollary 66. *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 67. *If A is symmetric and $CA = \mathbf{0}$ then $\mathbf{X}^T A\mathbf{X}$ and $C\mathbf{X}$ are independently distributed.*

Theorem 68. ***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.*

Exercise 69. Construction of t-test and F-test: Let $X_i \overset{iid}{\sim} N(0, \sigma^2)$. 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 from

Linear Algebra and Linear Models By Ravindra B. Bapat

Linear Models : An Integrated Approach By Debasis Sengupta, Sreenivasa Rao Jammalamadaka

5. 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 70. 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

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)\ \ \forall i = 1, 2, \cdots, n. \tag{5.1}$$

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

$$S(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2. \tag{5.2}$$

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

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0 \quad \Longrightarrow \quad n\hat\beta_0 + \hat\beta_1 \sum_i x_i = \sum_i y_i \tag{5.3}$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0 \quad \Longrightarrow \quad \hat\beta_0 \sum_i x_i + \hat\beta_1 \sum_i x_i^2 = \sum_i y_i x_i \tag{5.4}$$

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

$$\hat\beta_1 = \frac{S_{xy}}{S_{xx}} \ \text{ and } \ \hat\beta_0 = \bar y - \hat\beta_1 \bar x$$

(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.

Definition 71. Prediction or regression line: For any x such as old x_is or some x_{new} the prediction line is $\hat y = \hat\beta_0 + \hat\beta_1 x$

Definition 72. The **prediction error or residual** is defined as $e_i = y_i - \hat{y}_i$. and hence the residual sum of square (SSR) is

$$\begin{aligned}
\sum_i e_i^2 &= \sum_i (y_i - \hat{y}_i)^2 \\
&= \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&= \sum_i [(y_i - (\bar{y} - \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)]^2 \\
&= \sum_i [(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})]^2 \\
&= S_{yy} + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1 S_{xy} \\
&= S_{yy} - \hat{\beta}_1 S_{xy} \\
&= S_{yy} - \frac{S_{xy}^2}{S_{xx}}
\end{aligned} \tag{5.5}$$

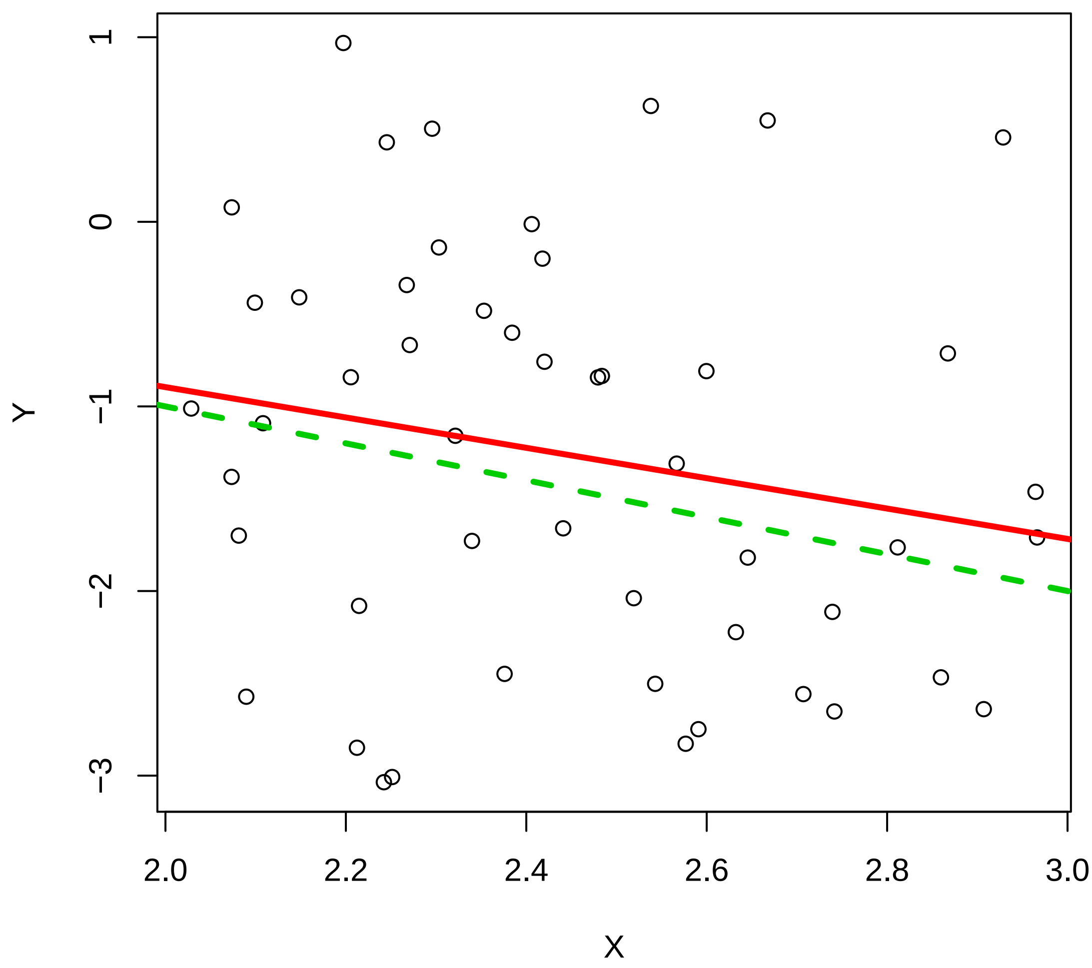
NOTE: We estimate σ^2 by $\hat{\sigma}^2 = \frac{SSR}{n-2} \equiv MSR$, mean residual sum of square. [Explanation will be studied under Multiple Linear Regression]

Definition 73. 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.

```
# Simple linear regression
####################################
N <- 50
b0 <- 1
b1 <- -1
X <- runif(N,2,3)
epsilon <- rnorm(N, 0,1)
Y <- b0 + b1 * X + epsilon
rr<-lm(Y~X)
print(rr)
```

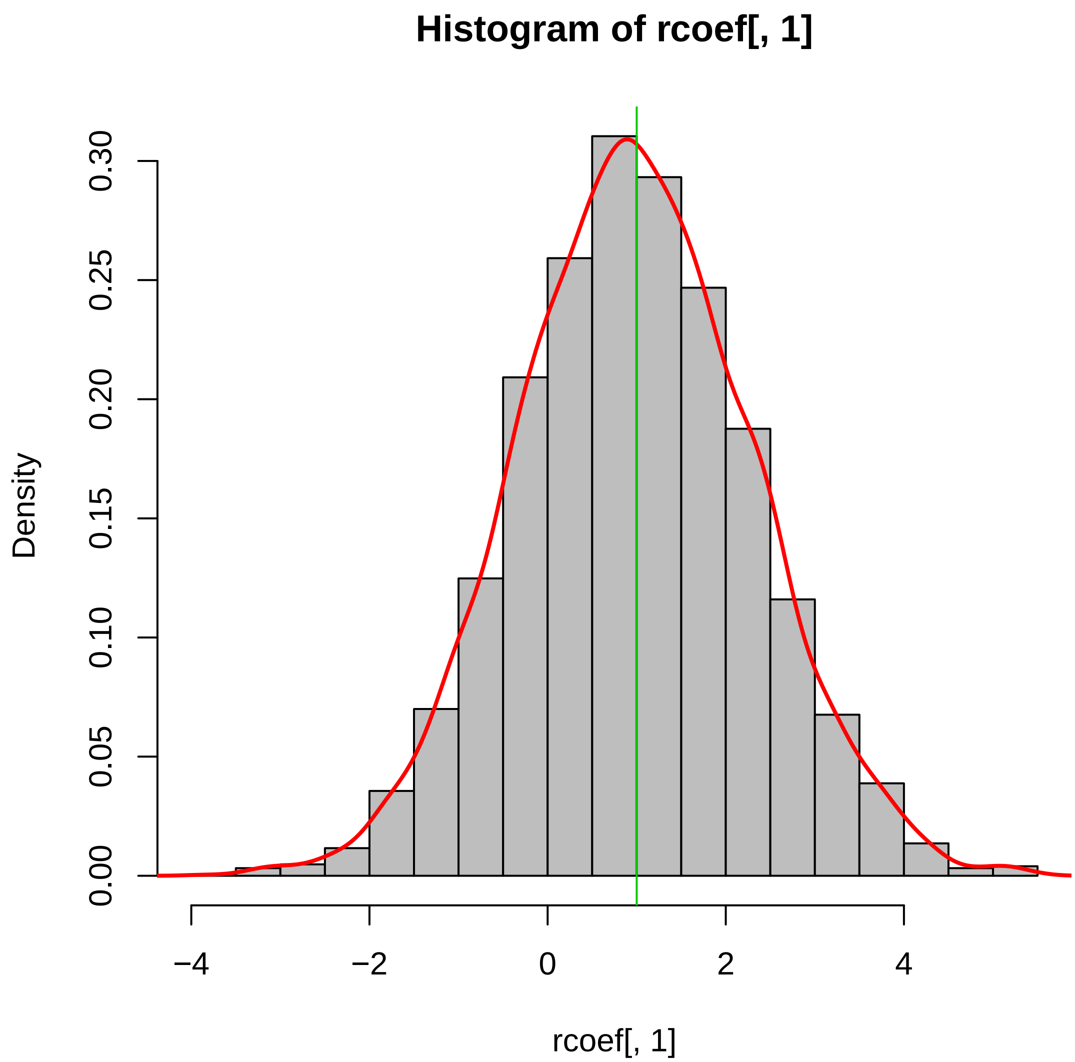
```
##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
```

```
## (Intercept)           X
##      0.7489     -0.8222
plot(X,Y)
abline(b0,b1, lty=2, col=3, lwd=3)
abline(rr$coefficients[1],rr$coefficients[2], col="red", lwd=3)
```

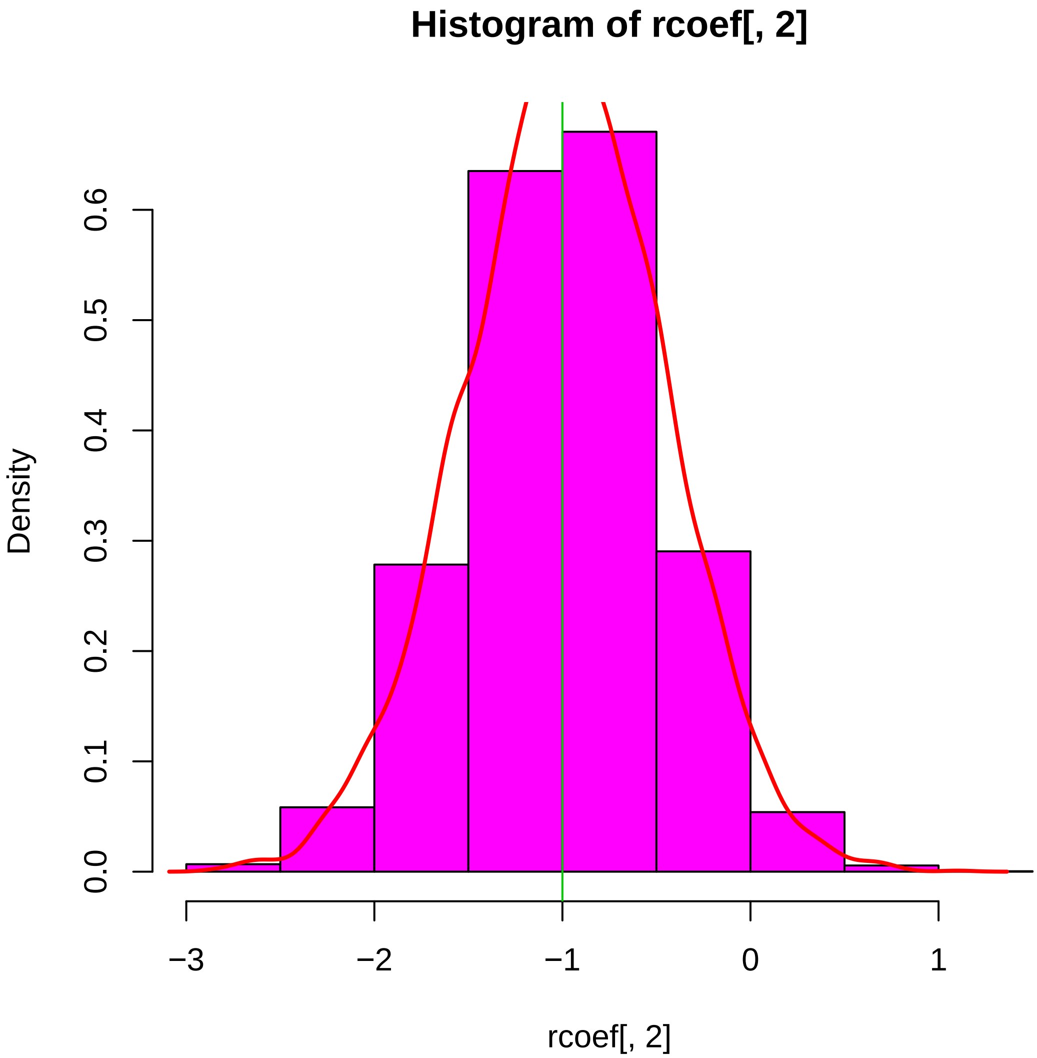



```
print(summary(rr))
##
## Call:
## lm(formula = Y ~ X)
```

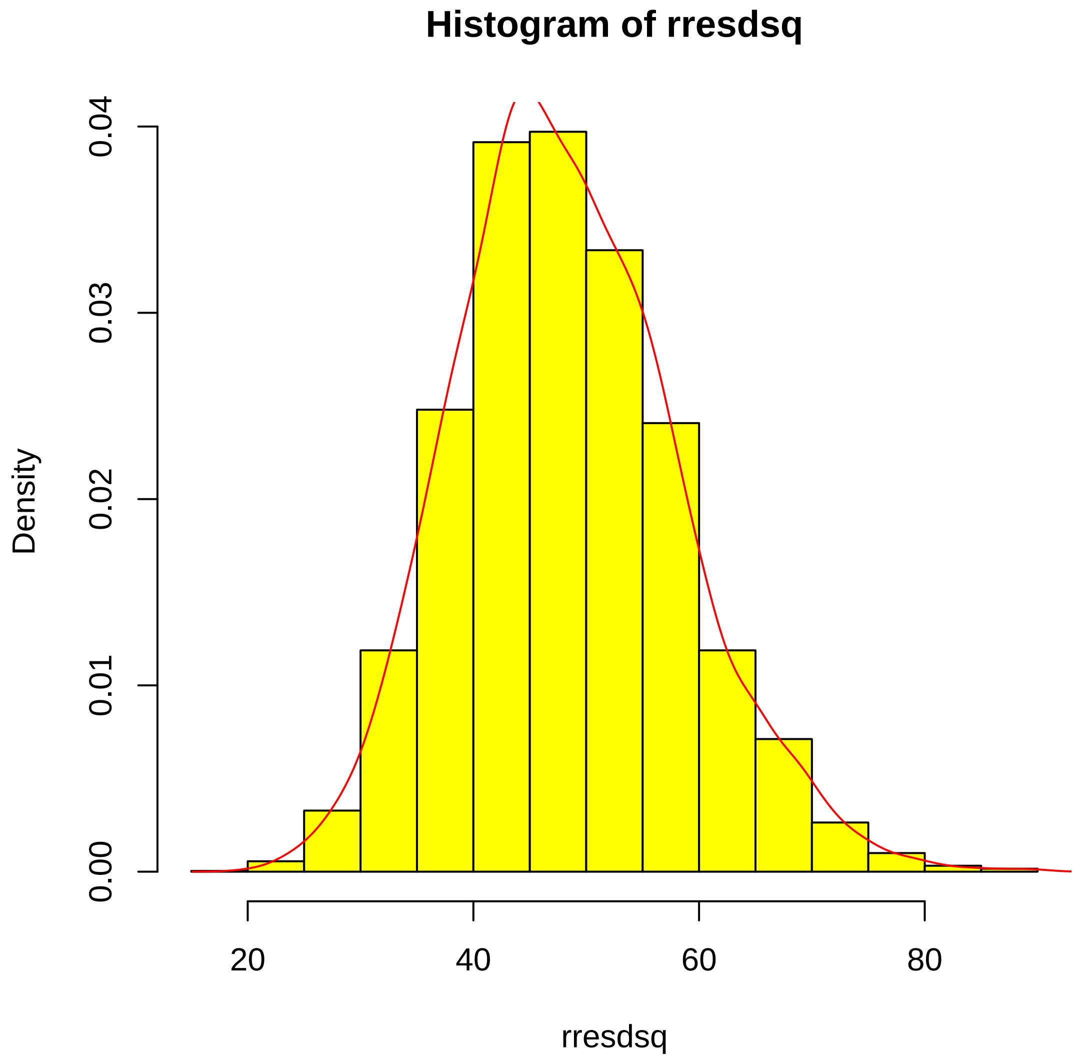
```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94092 -0.85067 -0.00954  0.68062  2.11614
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.7489     1.4572   0.514    0.610
## X            -0.8222     0.5928  -1.387    0.172
##
## Residual standard error: 1.104 on 48 degrees of freedom
## Multiple R-squared:  0.03853,Adjusted R-squared:  0.0185
## F-statistic: 1.924 on 1 and 48 DF,  p-value: 0.1719
# Distributions of  estimated parameters

itrn<-5000
rcoef<-array(0,dim=c(itrn,2))
rresdsq<-array(0,dim=c(itrn))
for(i in 1: itrn ){
epsilon <- rnorm(N, 0,1)
Y <- b0 + b1 * X + epsilon
r<-lm(Y~X)
rcoef[i,]<-r$coefficients
rresdsq[i]<-sum((r$residuals)^2)
}
hist(rcoef[,1], probability = T, col=8)
lines(density(rcoef[,1]), col=2, lwd=2)
abline(v=b0, col=3)
```




```r
hist(rcoef[,2], probability = T, col=6)
lines(density(rcoef[,2]), col=2, lwd=2)
abline(v=b1, col=3)
```




```
hist(rresdsq, probability = T, col=7)
lines(density(rresdsq), col=2)
```


Histogram of rresdsq
Density
0.00
0.01
0.02
0.03
0.04
20
40
60
80
rresdsq

6. Multiple linear regression

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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon = \mathbf{x}^T\beta + \epsilon, \tag{6.1}$$

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 (6.1) 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,

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{6.2}$$

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,**

$$\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}_n) \tag{6.3}$$

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

$$S(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{Y}^T\mathbf{Y} - 2\beta^T\mathbf{X}^T\mathbf{Y} + \beta^T\mathbf{X}^T\mathbf{X}\beta \tag{6.4}$$

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

$$\begin{aligned} & \frac{\partial S(\beta)}{\partial \beta}|_{\beta=\hat{\beta}} = \mathbf{0} \\ \implies & -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{0} \\ \implies & \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{y} \end{aligned} \tag{6.5}$$

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 74. 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 75. 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 76. 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 77. BLUE: The best linear unbiased estimator (BLUE) of a linear parametric function $p^T\beta$ is a LUE with minimum variance.

Theorem 78. *A linear function is BLUE of its expectation iff it is uncorrelated with all LZF.*

Corollary 79. *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

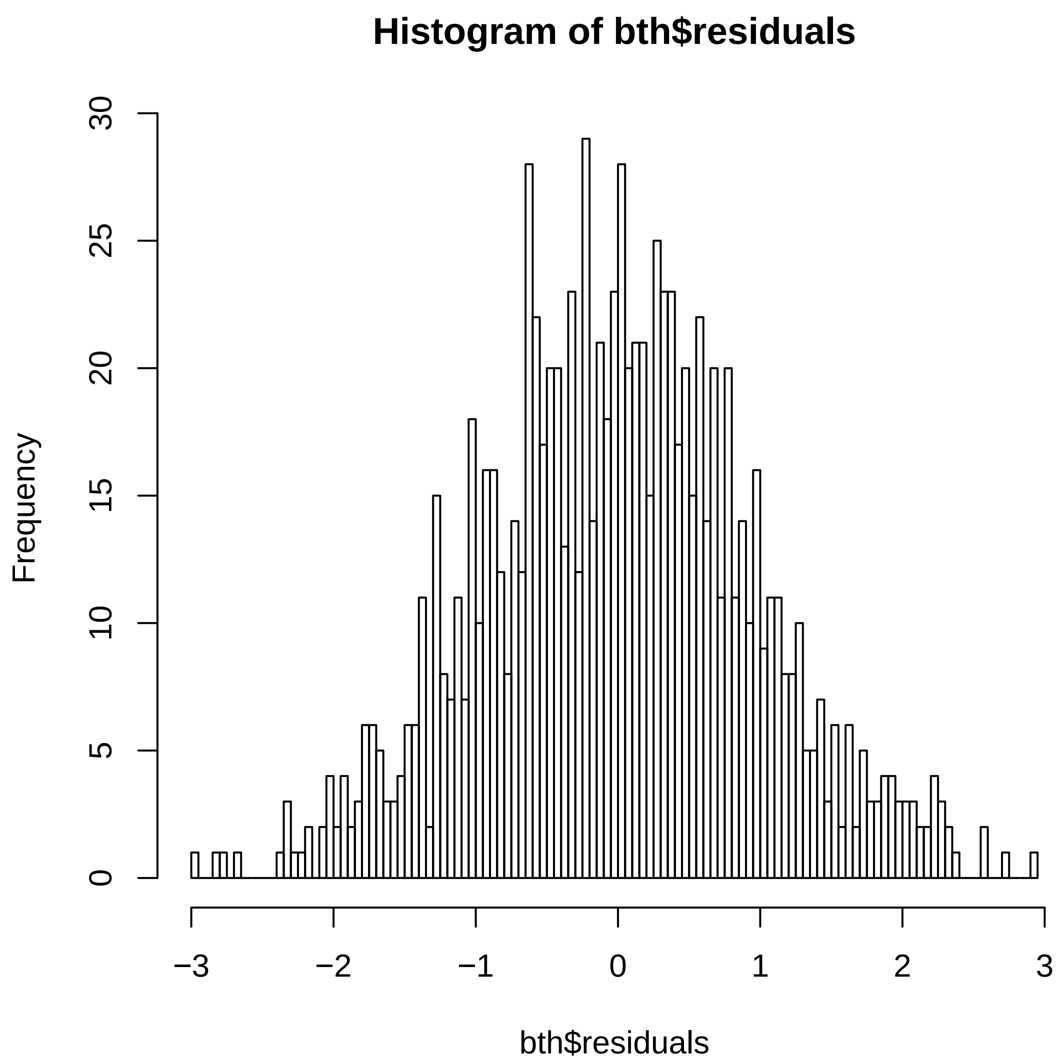
$$\frac{SSModel/k}{SSRes/(n-k-1)} \sim F_{k,(n-k-1),ncp_1=\lambda}$$

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

```
# Multiple linear regression
####################################
k<-3 # indepencent variables
n<- 1000 # number of observations
X<-array(0,dim=c(n,(k+1)))
bt<-c(1,2.3, 1.5,0.05)


x1<-rgamma(n,2,3)
x2<-rbinom(n,10,0.7)
x3<-rbeta(n,0.5,0.5)
X[,1]<-1
X[,2]<-x1
X[,3]<-x2
```

```
X[,4]<-x3
eps<-rnorm(n,0,1)
y<-X%*%bt+eps ##### OR  y<-bt[1]+bt[2]*x1+bt[3]*x2+bt[4]*x3+eps
d<- data.frame(y,x1,x2,x3)
write.table(d,"data.txt")

print(d[1:10,])
##           y        x1 x2          x3
## 1  18.48191 0.9819996 10 0.115169983
## 2  16.35794 0.8201889  9 0.004061756
## 3  14.45718 0.6187825  8 0.047126432
## 4  13.25062 0.7428446  7 0.093569209
## 5  18.64918 1.9021928  9 0.999988338
## 6  11.23165 0.4738346  7 0.995279240
## 7  14.66262 0.1891390  7 0.867041732
## 8  11.08802 0.3839012  6 0.197883604
## 9  16.80316 1.5258726  8 0.477394076
## 10 16.17364 0.8521659  9 0.810154224
# #------Estimation-------------
#
bth1<-solve(t(X)%*%X)%*%t(X)%*%y  # value of beta_hat using formula

print(bth1)
##           [,1]
## [1,] 0.9232149
## [2,] 2.3517364
## [3,] 1.4973246
## [4,] 0.1257661
# Create the relationship model.

bth<-lm(y~x1+x2+x3,data=d)

cat("trur beta=",bt, "sigma=",1,'\n')
## trur beta= 1 2.3 1.5 0.05 sigma= 1
```

```
cat("beta_hat=",coefficients(bth), "sigma_hat=",summary(bth)$sigma,'\n') # model coefficients
## beta_hat= 0.9232149 2.351736 1.497325 0.1257661 sigma_hat= 0.9541874
par(mfrow=c(1,1))
hist(bth$residuals, breaks = 100 )
```

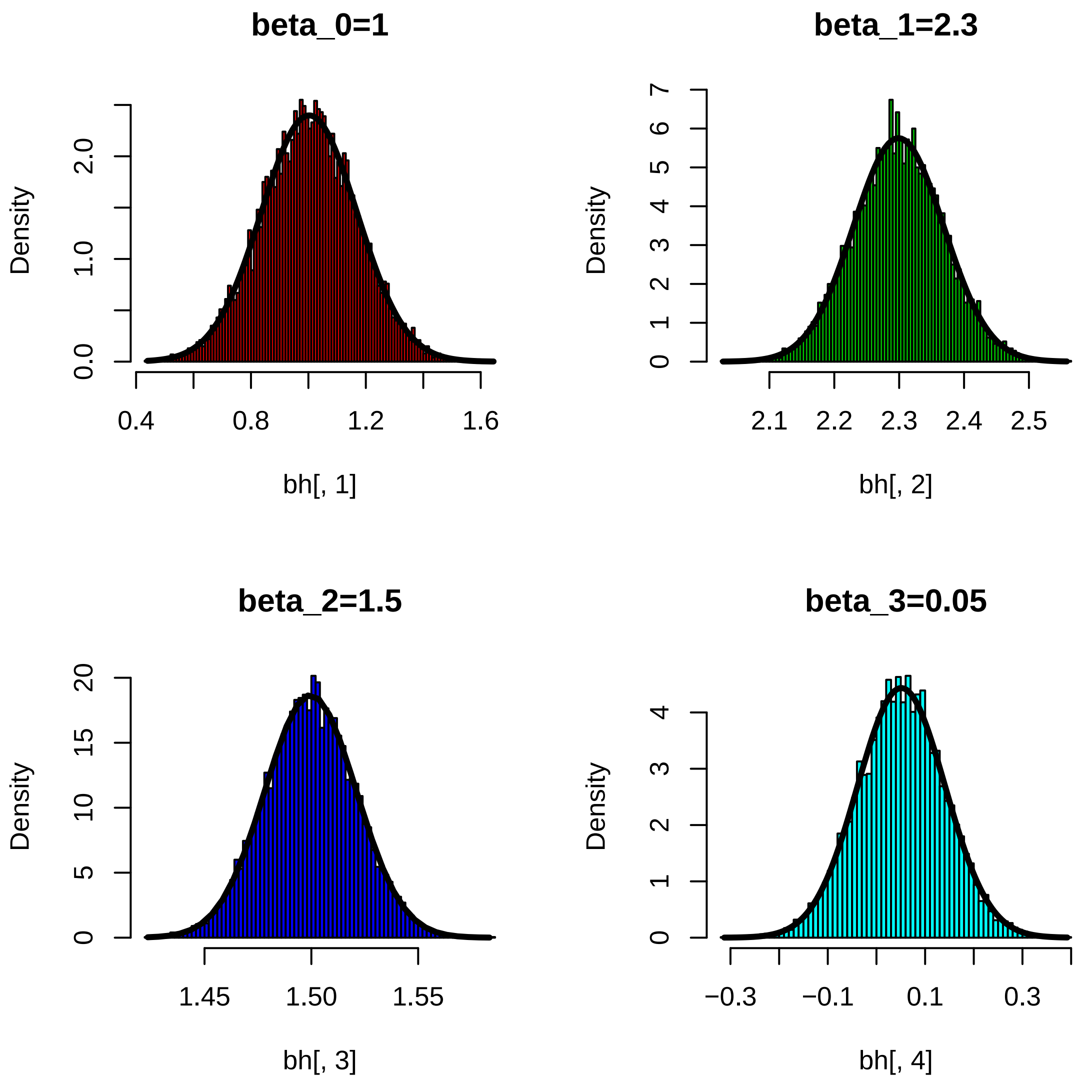



```
print(summary(bth))
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = d)
##
```

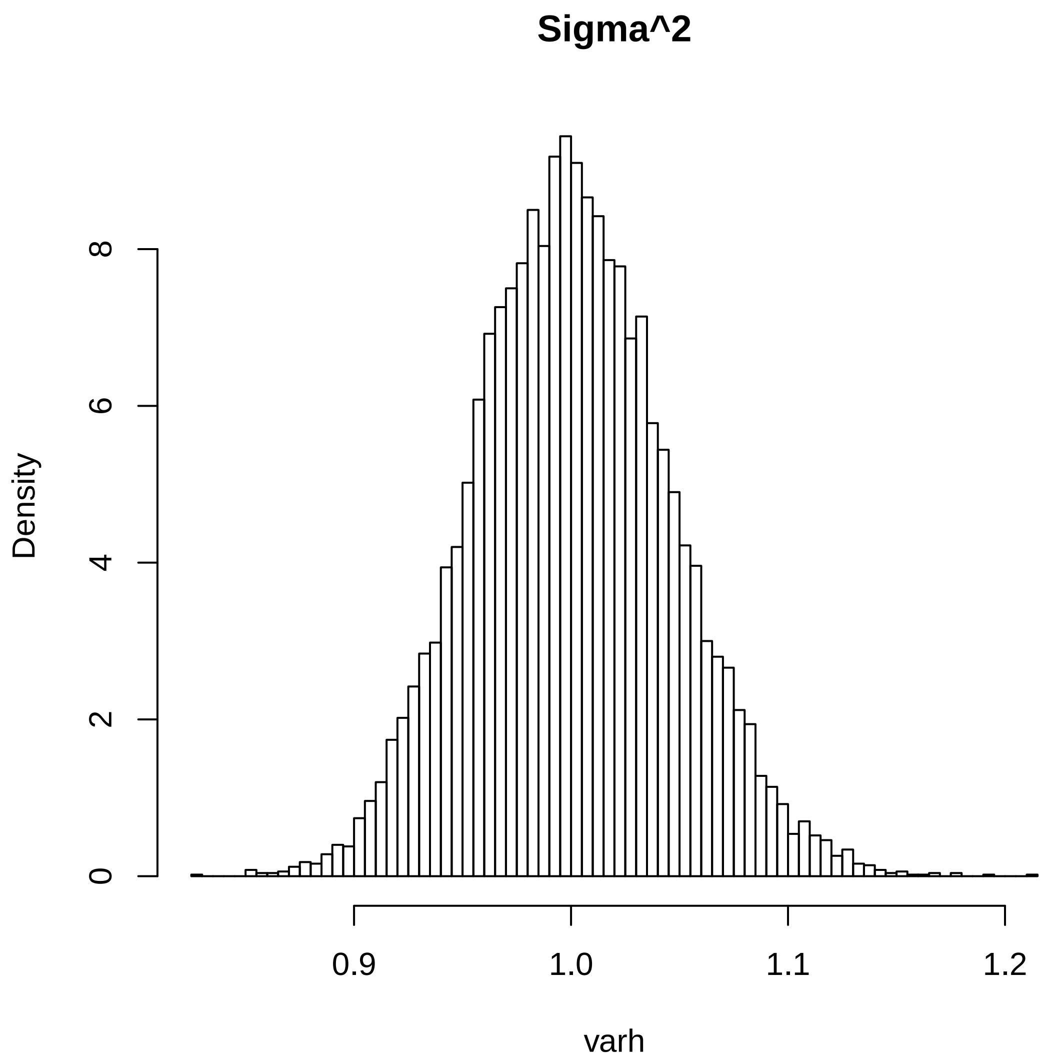
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.98867 -0.62048  0.01058  0.61311  2.91625
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.92321    0.15847   5.826 7.67e-09 ***
## x1           2.35174    0.06609  35.584  < 2e-16 ***
## x2           1.49732    0.02035  73.594  < 2e-16 ***
## x3           0.12577    0.08626   1.458    0.145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9542 on 996 degrees of freedom
## Multiple R-squared:  0.8682,Adjusted R-squared:  0.8678
## F-statistic:  2187 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
itrn<-10000
bh<-array(0,dim=c(itrn,(k+1)))
varh<-array(0,dim=c(itrn))

for (i in 1 : itrn){
  eps<-rnorm(n,0,1)
  y<-X%*%bt+eps ##### OR  y<-bt[1]+bt[2]*x1+bt[3]*x2+bt[4]*x3+eps
  d<- data.frame(y,x1,x2,x3)
  bth<-lm(y~x1+x2+x3,data=d)
  bh[i,]<-coefficients(bth)
  varh[i]<-(summary(bth)$sigma)^2

}

cat("trur beta=",bt,'\n')
## trur beta= 1 2.3 1.5 0.05
cat("mean of estimated  beta=",colMeans(bh),'\n')
## mean of estimated  beta= 1.002452 2.299017 1.49962 0.05106693
```

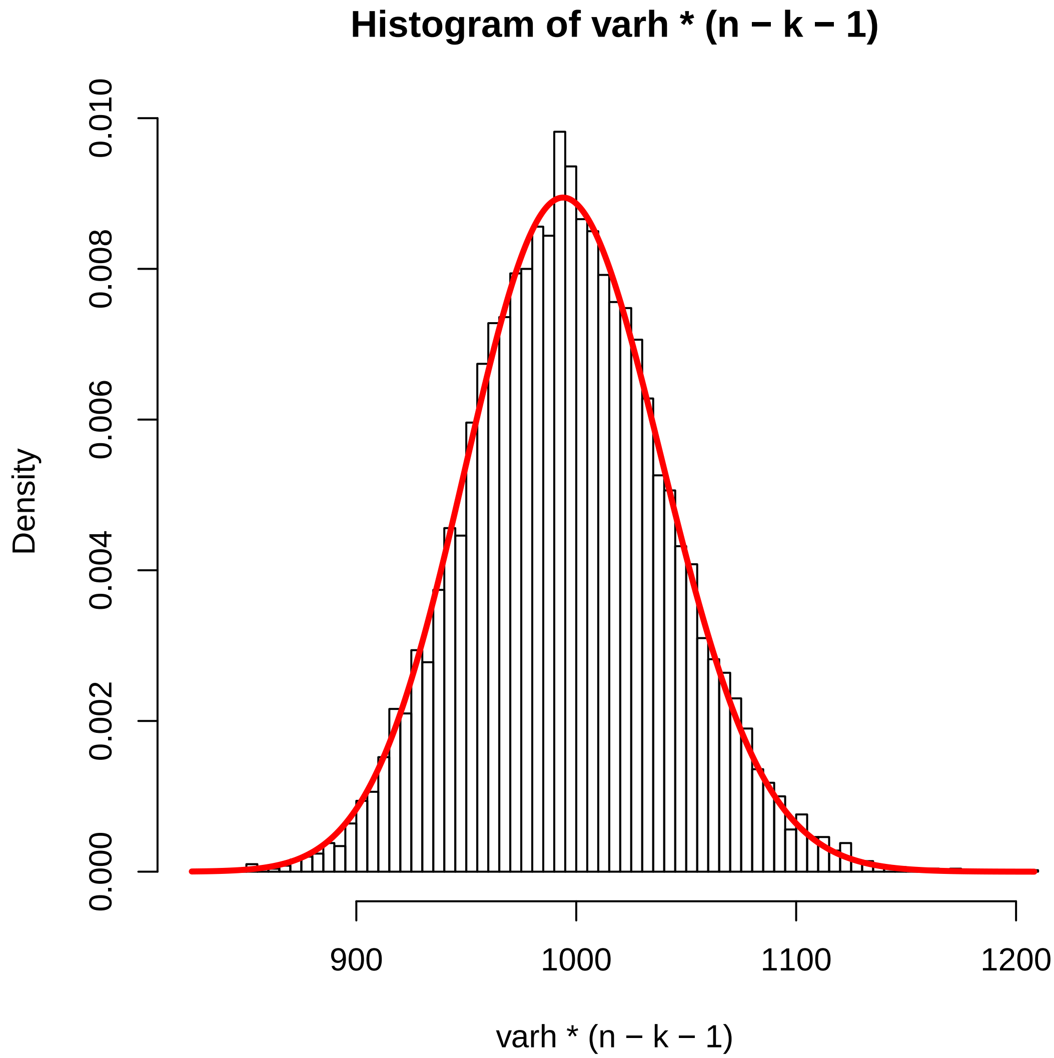
```
par(mfrow=c(2,2))
hist(bh[,1], breaks = 100, col=2, probability = T , main = "beta_0=1")
s<-seq(min(bh[,1]),max(bh[,1]),by=0.005)
lines(dnorm(s, mean=mean(bh[,1]), sd=sd(bh[,1]))~s, col=1, lwd=3)

hist(bh[,2], breaks = 100, col=3, probability = T, main = "beta_1=2.3")
s<-seq(min(bh[,2]),max(bh[,2]),by=0.005)
lines(dnorm(s, mean=mean(bh[,2]), sd=sd(bh[,2]))~s, col=1, lwd=3)

hist(bh[,3], breaks = 100, col=4, probability = T, main = "beta_2=1.5")
s<-seq(min(bh[,3]),max(bh[,3]),by=0.005)
lines(dnorm(s, mean=mean(bh[,3]), sd=sd(bh[,3]))~s, col=1, lwd=3)

hist(bh[,4], breaks = 100, col=5, probability = T, main = "beta_3=0.05")
s<-seq(min(bh[,4]),max(bh[,4]),by=0.005)
lines(dnorm(s, mean=mean(bh[,4]), sd=sd(bh[,4]))~s, col=1, lwd=3)
```



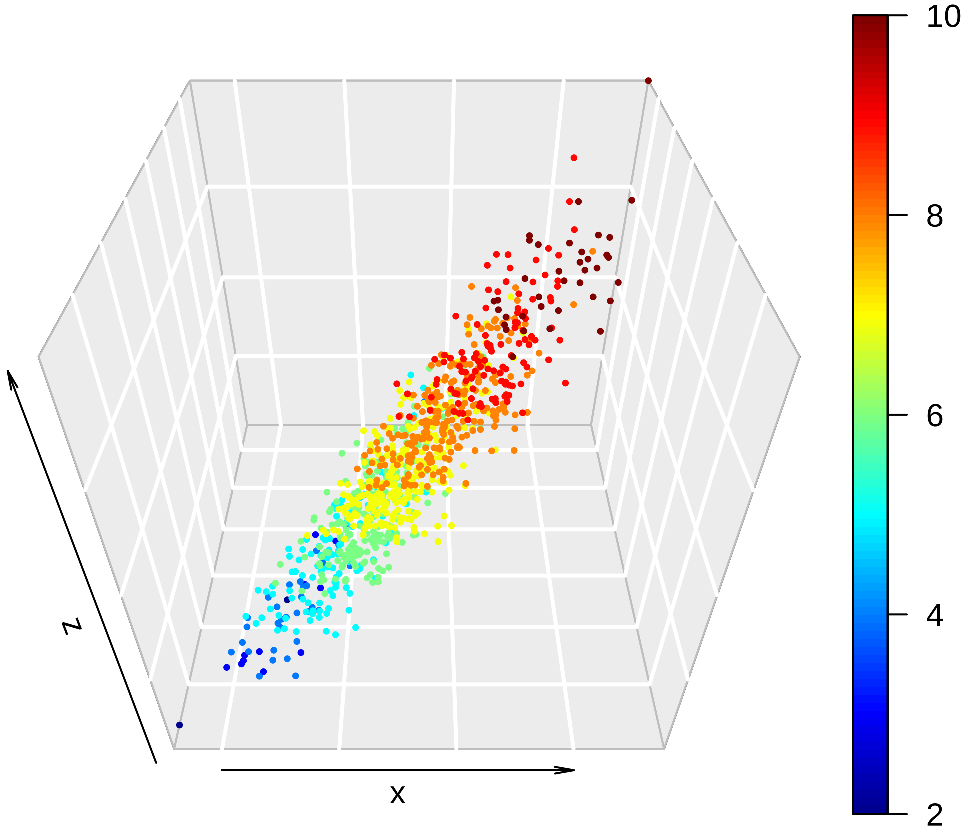

```
par(mfrow=c(1,1))
hist(varh, breaks = 100, probability = T , main = "Sigma^2")
```




```
par(mfrow=c(1,1))
hist(varh*(n-k-1), breaks = 100, probability = T )
s<-seq(min(varh*(n-k-1)),max(varh*(n-k-1)),by=0.05)
lines(dchisq(s, df=(n-k-1))~s, col=2, lwd=3)
```

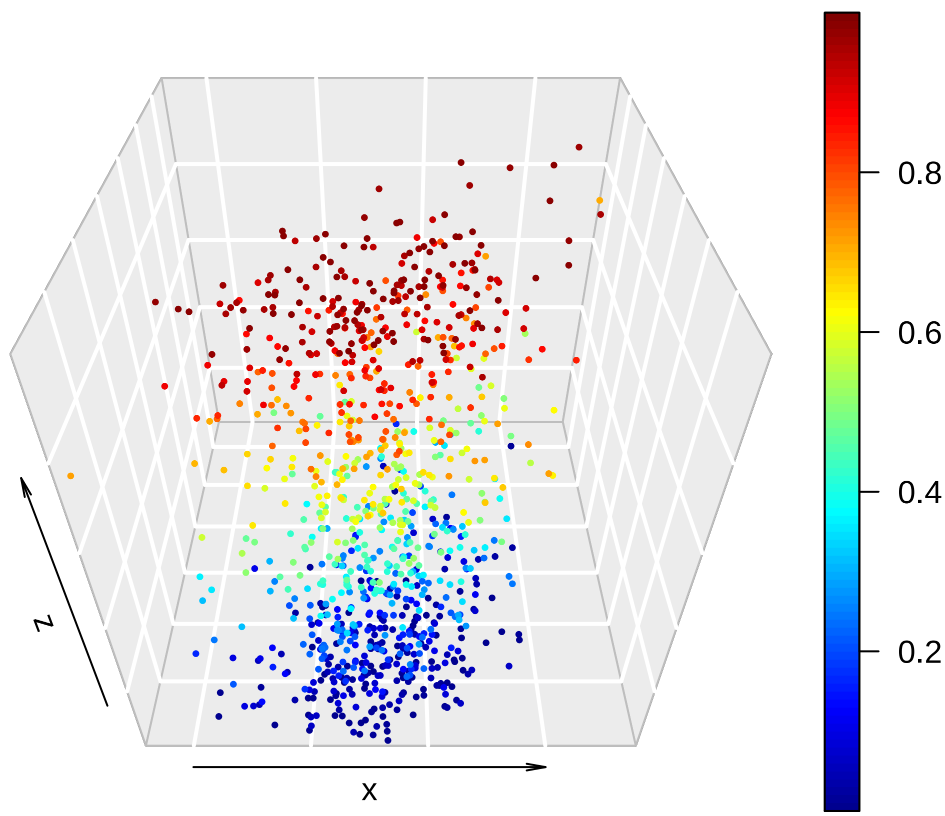



```
#Data presntation


d1<-data.frame(y,x1,x2)
library("plot3D")
scatter3D(d1[,1],d1[,2],d1[,3], theta = 00, phi = 40,  bty = "g", pch = 20, cex = 0.5)# ticktype = "det
```

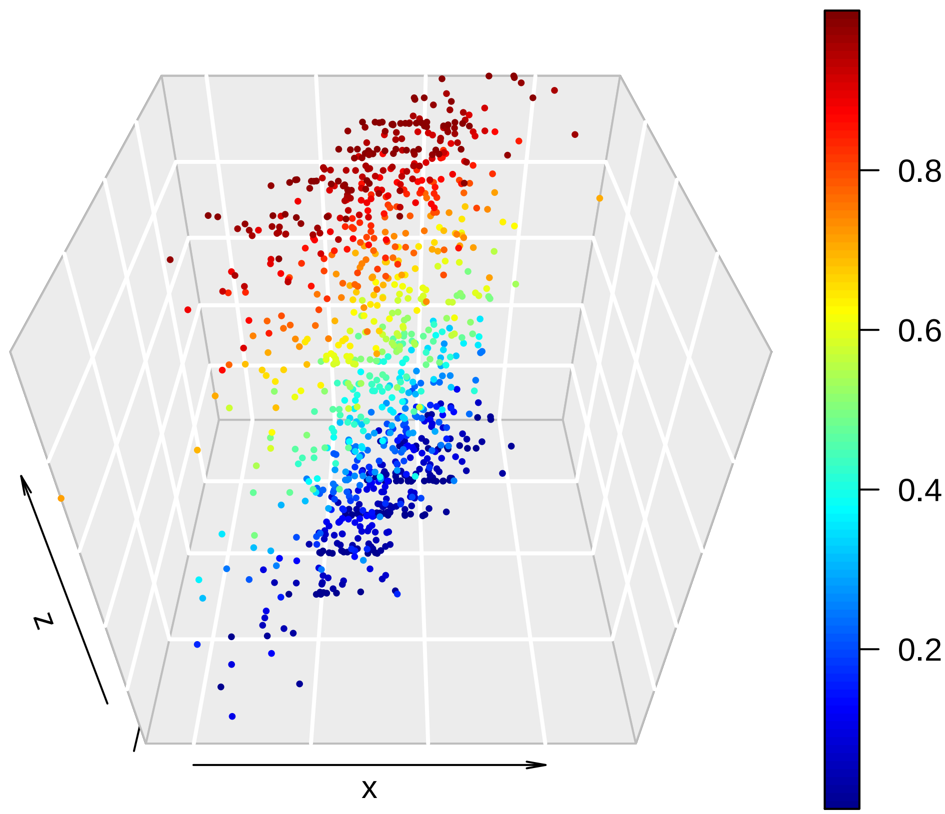



```
library("plot3Drgl")

## Loading required package:  rgl

## Warning:  package 'rgl' was built under R version 3.4.4

plotrgl()

#



d1<-data.frame(y,x1,x3)

library("plot3D")
```

```
scatter3D(d1[,1],d1[,2],d1[,3], theta = 00, phi = 40,  bty = "g", pch = 20, cex = 0.5)# ticktype = "det
```




```
library("plot3Drgl")
plotrgl()

d1<-data.frame(y,x2,x3)
library("plot3D")
scatter3D(d1[,1],d1[,2],d1[,3], theta = 00, phi = 40,  bty = "g", pch = 20, cex = 0.5)# ticktype = "det
```



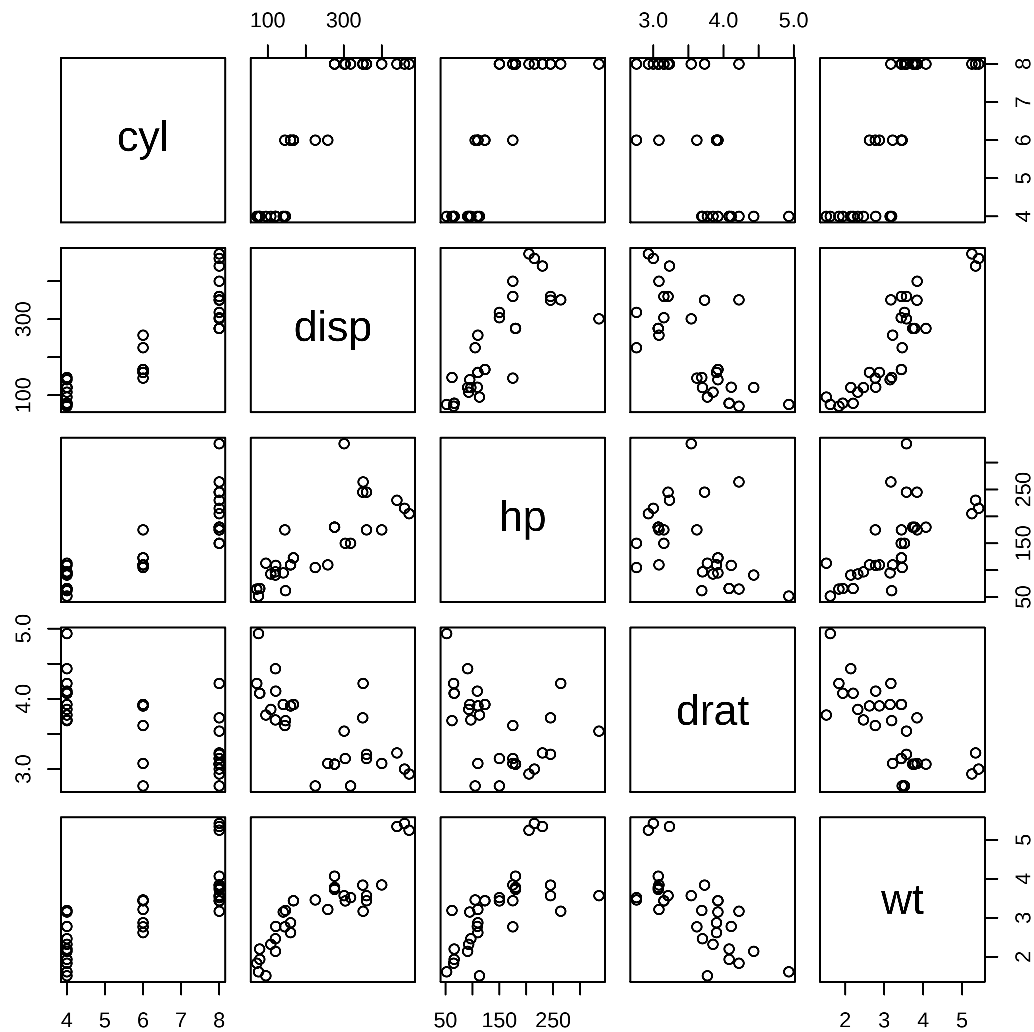

```r
library("plot3Drgl")
plotrgl()
```

```r
# Multiple linear regression
###################################

# Motor Trend Car Road Tests
#
# Description
#
```

```
# The data was extracted from the 1974 Motor Trend US magazine,
# and comprises fuel consumption and 10 aspects of automobile design
# and performance for 32 automobiles (1973-74 models).
# A data frame with 32 observations on 11 variables.
#
# [, 1] mpg    Miles/(US) gallon
# [, 2] cyl    Number of cylinders
# [, 3] disp  Displacement (cu.in.)
# [, 4] hp    Gross horsepower
# [, 5] drat  Rear axle ratio
# [, 6] wt    Weight (1000 lbs)
# [, 7] qsec  1/4 mile time
# [, 8] vs    V/S
# [, 9] am    Transmission (0 = automatic, 1 = manual)
# [,10] gear  Number of forward gears
# [,11] carb  Number of carburetors
# Source
#
# Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391

print(mtcars[,c("cyl","disp","hp","drat","wt")])

##                     cyl  disp  hp drat    wt
## Mazda RX4             6 160.0 110 3.90 2.620
## Mazda RX4 Wag         6 160.0 110 3.90 2.875
## Datsun 710            4 108.0  93 3.85 2.320
## Hornet 4 Drive        6 258.0 110 3.08 3.215
## Hornet Sportabout     8 360.0 175 3.15 3.440
## Valiant               6 225.0 105 2.76 3.460
## Duster 360            8 360.0 245 3.21 3.570
## Merc 240D             4 146.7  62 3.69 3.190
## Merc 230              4 140.8  95 3.92 3.150
## Merc 280              6 167.6 123 3.92 3.440
## Merc 280C             6 167.6 123 3.92 3.440
## Merc 450SE            8 275.8 180 3.07 4.070
```

```
## Merc 450SL            8 275.8 180 3.07 3.730
## Merc 450SLC           8 275.8 180 3.07 3.780
## Cadillac Fleetwood    8 472.0 205 2.93 5.250
## Lincoln Continental   8 460.0 215 3.00 5.424
## Chrysler Imperial     8 440.0 230 3.23 5.345
## Fiat 128              4  78.7  66 4.08 2.200
## Honda Civic           4  75.7  52 4.93 1.615
## Toyota Corolla        4  71.1  65 4.22 1.835
## Toyota Corona         4 120.1  97 3.70 2.465
## Dodge Challenger      8 318.0 150 2.76 3.520
## AMC Javelin           8 304.0 150 3.15 3.435
## Camaro Z28            8 350.0 245 3.73 3.840
## Pontiac Firebird      8 400.0 175 3.08 3.845
## Fiat X1-9             4  79.0  66 4.08 1.935
## Porsche 914-2         4 120.3  91 4.43 2.140
## Lotus Europa          4  95.1 113 3.77 1.513
## Ford Pantera L        8 351.0 264 4.22 3.170
## Ferrari Dino          6 145.0 175 3.62 2.770
## Maserati Bora         8 301.0 335 3.54 3.570
## Volvo 142E            4 121.0 109 4.11 2.780

summary(mtcars[,c("cyl","disp","hp","drat","wt")])

##       cyl             disp             hp             drat
##  Min.   :4.000   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :6.000   Median :196.3   Median :123.0   Median :3.695
##  Mean   :6.188   Mean   :230.7   Mean   :146.7   Mean   :3.597
##  3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
##  Max.   :8.000   Max.   :472.0   Max.   :335.0   Max.   :4.930
##        wt
##  Min.   :1.513
##  1st Qu.:2.581
##  Median :3.325
##  Mean   :3.217
##  3rd Qu.:3.610
```

```
##  Max.   :5.424
plot((mtcars[,c("cyl","disp","hp","drat","wt")]))
```




```
fit1<-lm(mpg~cyl+disp+hp+drat+wt, data = mtcars)
summary(fit1)
##
## Call:
## lm(formula = mpg ~ cyl + disp + hp + drat + wt, data = mtcars)
##
## Residuals:
```

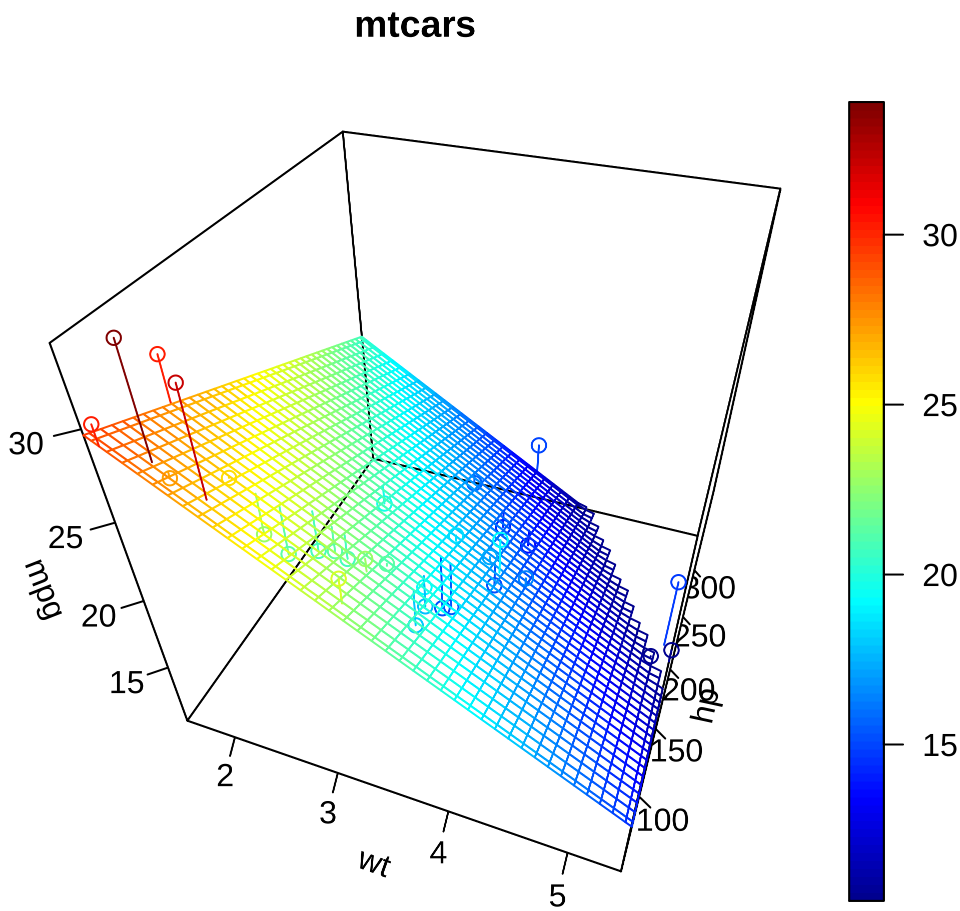
```
##      Min       1Q   Median       3Q      Max
## -3.7014 -1.6850 -0.4226  1.1681  5.7263
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.00836    7.57144   4.756  6.4e-05 ***
## cyl         -1.10749    0.71588  -1.547  0.13394
## disp         0.01236    0.01190   1.039  0.30845
## hp          -0.02402    0.01328  -1.809  0.08208 .
## drat         0.95221    1.39085   0.685  0.49964
## wt          -3.67329    1.05900  -3.469  0.00184 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 26 degrees of freedom
## Multiple R-squared:  0.8513,Adjusted R-squared:  0.8227
## F-statistic: 29.77 on 5 and 26 DF,  p-value: 5.618e-10
```

```
fit2<-lm(mpg~hp+wt, data = mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ hp + wt, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.941 -1.600 -0.182  1.050  5.854
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
## hp          -0.03177    0.00903  -3.519  0.00145 **
## wt          -3.87783    0.63273  -6.129 1.12e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
x <- mtcars$wt
y <- mtcars$hp
z <- mtcars$mpg
fit <- lm(z ~ x + y)
grid.lines = 40
x.pred <- seq(min(x), max(x), length.out = grid.lines)
y.pred <- seq(min(y), max(y), length.out = grid.lines)
xy <- expand.grid( x = x.pred, y = y.pred)
z.pred <- matrix(predict(fit, newdata = xy), nrow = grid.lines, ncol = grid.lines)
fitpoints <- predict(fit)
library("plot3D")
scatter3D(x, y, z, pch = 1, cex = 1, theta = 20, phi = 40, ticktype = "detailed",xlab = "wt", ylab = "hp
```




```
library("plot3Drgl")
plotrgl()
```

7. Polynomial regression

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

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \quad \text{for single regressor} \tag{7.1}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon \quad \text{for multiple regressors} \tag{7.2}$$

In general for k-degree polynomial for single regressor , as modelled in equation (7.1), we can write in matrix notation

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{7.3}$$

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,**

$$\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n) \tag{7.4}$$

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 th 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(7.1) and (7.2). But regression model need not be so. Instead of equation (7.2) , 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

$$P_0(x_i) = 1 \text{ and } \sum_{i=1}^{n} P_j(x_i) P_k(x_i) = 0 \ \forall j \neq k \tag{7.5}$$

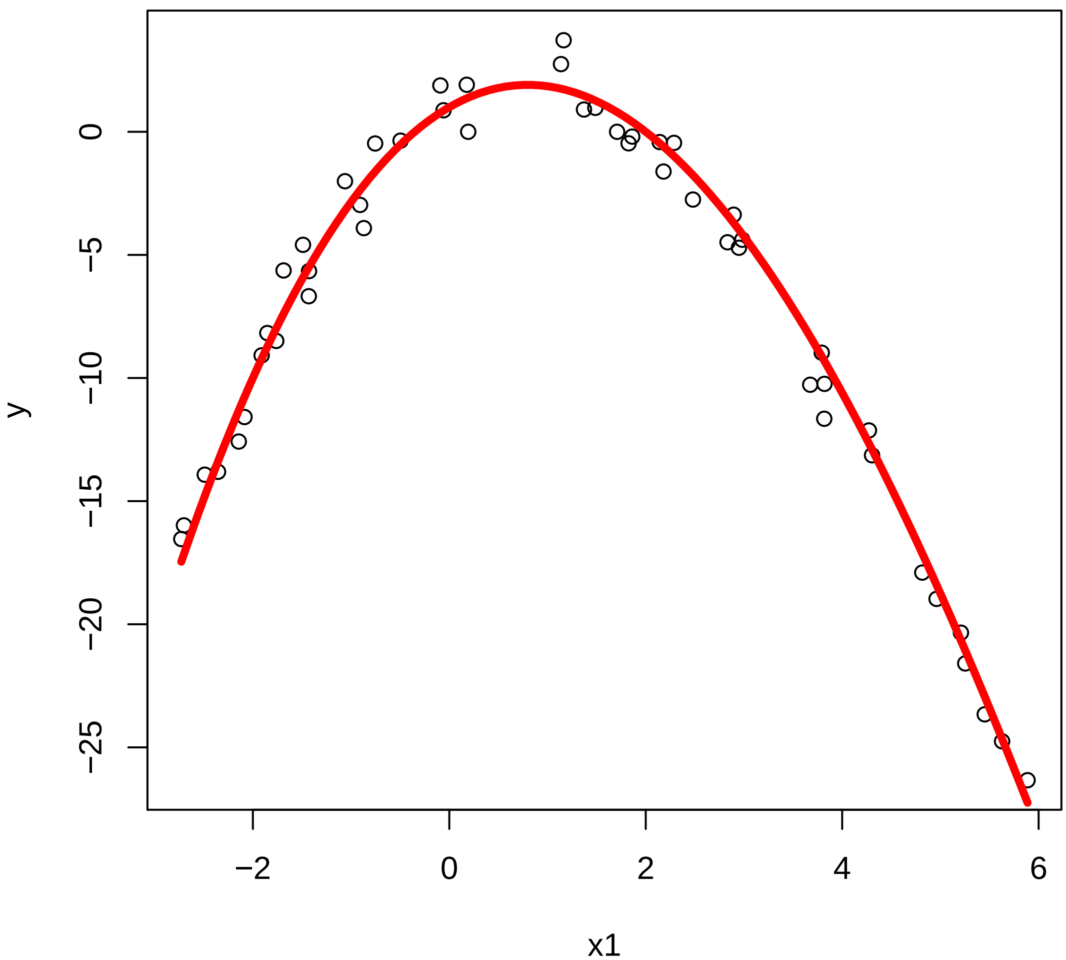
Now the model will look as follows

$$y_i = \sum_{j=0}^{k} \alpha_j P_j(x_i) + \epsilon_i \ \forall i = 1, 2, \cdots n \tag{7.6}$$

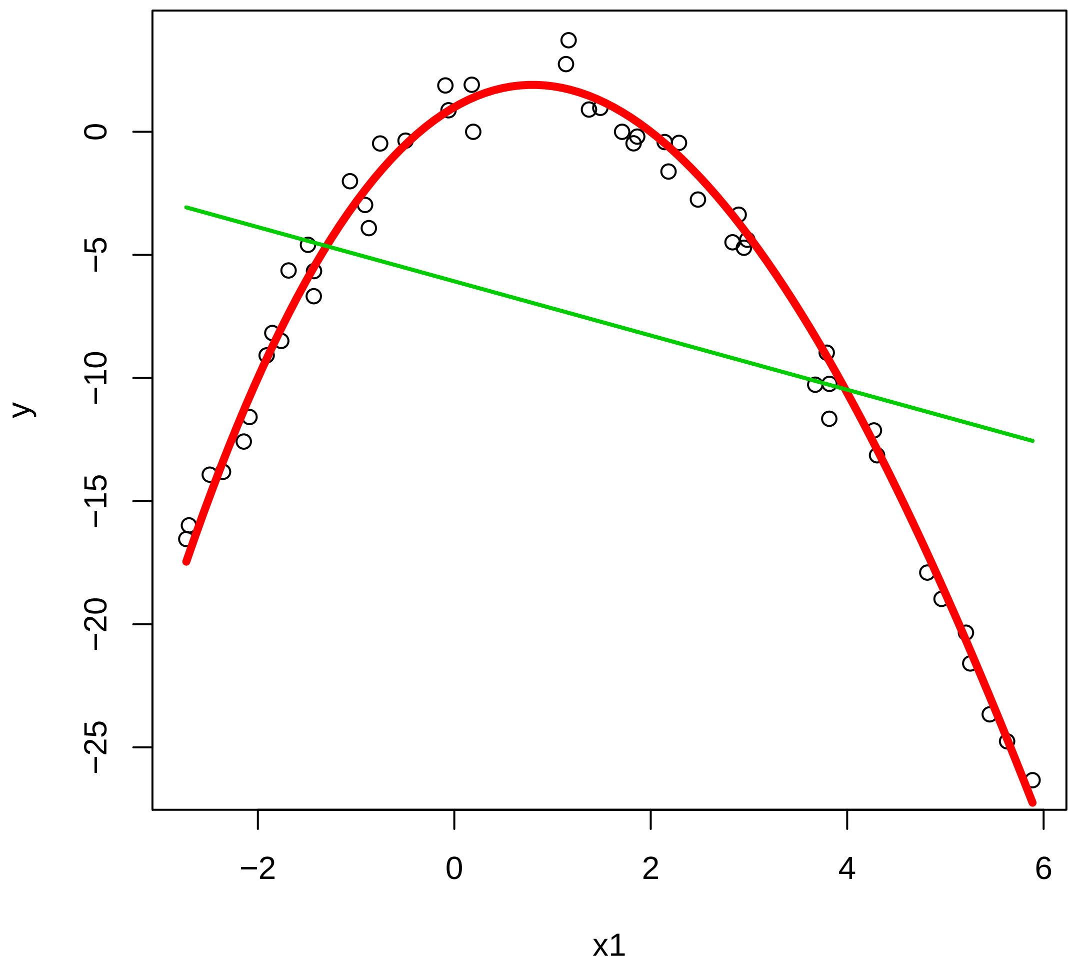
or denoting $\mathbf{Z} = ((P_j(x_i)))_{n \times (k+1)}$

$$\mathbf{Y} = \mathbf{Z}\alpha + \epsilon$$

```
# Polynomial  regression
####################################
k<-3 # indepencent variables
n<- 50 # number of observations
X<-array(0,dim=c(n,(k+1)))
bt<-c(1,2.3, -1.5,0.05)

x1<-sort(runif(n,-3,6))
x2<-x1^2
x3<-x1^3
X[,1]<-1
X[,2]<-x1
X[,3]<-x2
X[,4]<-x3
eps<-rnorm(n,0,1)
y<-X%*%bt+eps ##### OR  y<-bt[1]+bt[2]*x1+bt[3]*x2+bt[4]*x3+eps
d<- data.frame(y,x1,x2,x3)
write.table(d,"data.txt")

print(d[1:10,])
##             y        x1       x2         x3
## 1  -16.538785 -2.728578 7.445138 -20.314639
## 2  -15.983342 -2.701610 7.298694 -19.718223
## 3  -13.923523 -2.489812 6.199164 -15.434755
## 4  -13.808264 -2.354764 5.544915 -13.056969
## 5  -12.579717 -2.143547 4.594793  -9.849153
## 6  -11.583103 -2.085180 4.347974  -9.066307
## 7   -9.078760 -1.910441 3.649783  -6.972694
## 8   -8.171739 -1.852746 3.432669  -6.359865
## 9   -8.492925 -1.762082 3.104932  -5.471144
## 10  -5.630671 -1.687479 2.847585  -4.805240
plot(y~x1)
s<-seq(min(x1), max(x1),length=100)
lines(bt[1]+bt[2]*s+bt[3]*s^2+bt[4]*s^3~s,col=2, lwd=4, lty=1)
```



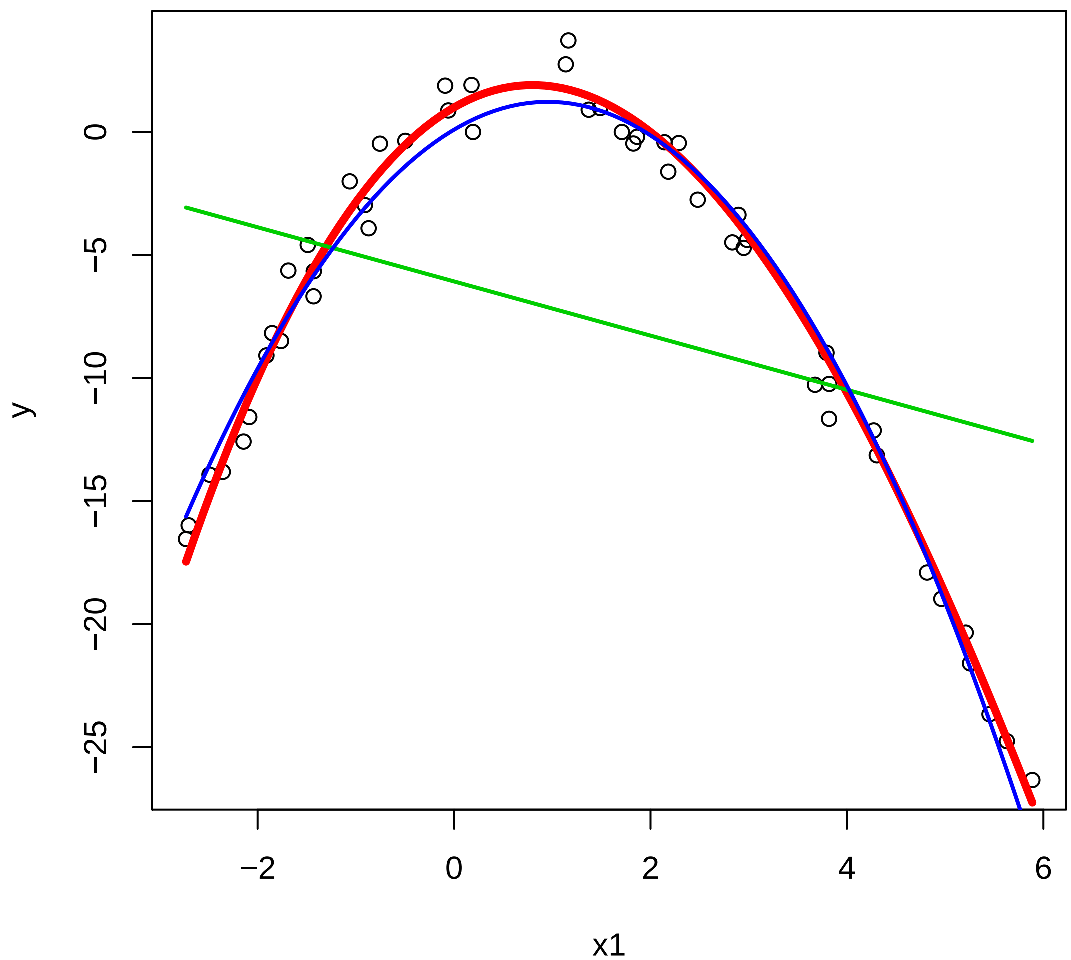

```
plot(y~x1)
s<-seq(min(x1), max(x1),length=100)
lines(bt[1]+bt[2]*s+bt[3]*s^2+bt[4]*s^3~s,col=2, lwd=4, lty=1)
bth<-lm(y~x1,data=d)
lines(bth$coefficients[1]+bth$coefficients[2]*s~s,col=3, lwd=2, lty=1)
```




```
cat("trur beta=",bt, "sigma=",1,'\n')

## trur beta= 1 2.3 -1.5 0.05 sigma= 1

cat("beta_hat=",coefficients(bth), "sigma_hat=",summary(bth)$sigma,'\n') # model coefficients

## beta_hat= -6.073633 -1.100642 sigma_hat= 7.478236

print(summary(bth))

##
## Call:
## lm(formula = y ~ x1, data = d)
##
```

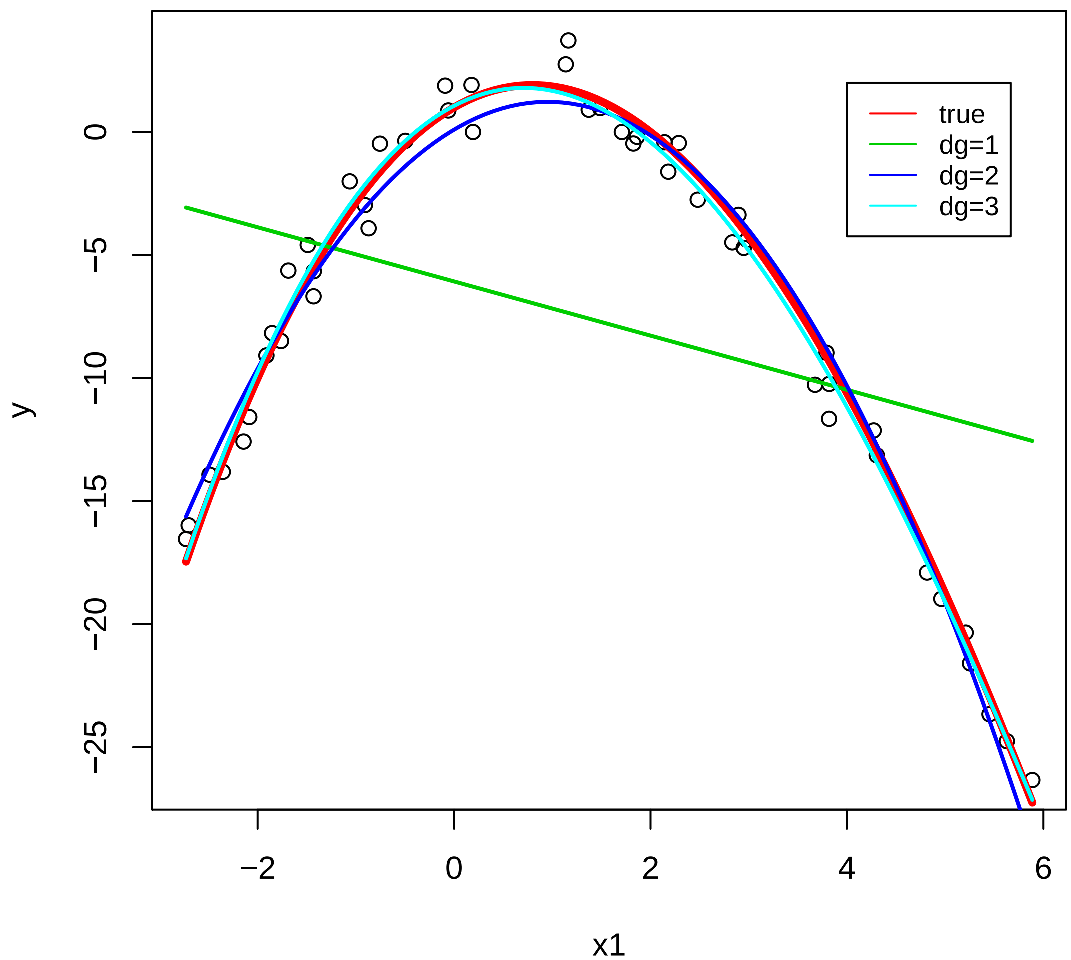
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.7789  -6.1719   0.6227   6.7169  11.0742
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.0736     1.1705  -5.189 4.23e-06 ***
## x1           -1.1006     0.4029  -2.732  0.00879 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.478 on 48 degrees of freedom
## Multiple R-squared:  0.1345,Adjusted R-squared:  0.1165
## F-statistic: 7.462 on 1 and 48 DF,  p-value: 0.008791
```

```
plot(y~x1)
s<-seq(min(x1), max(x1),length=100)
lines(bt[1]+bt[2]*s+bt[3]*s^2+bt[4]*s^3~s,col=2, lwd=4, lty=1)
bth<-lm(y~x1,data=d)
lines(bth$coefficients[1]+bth$coefficients[2]*s~s,col=3, lwd=2, lty=1)
bth<-lm(y~x1+x2,data=d)
lines(bth$coefficients[1]+bth$coefficients[2]*s+bth$coefficients[3]*s^2~s,col=4, lwd=2, lty=1)
```




```r
cat("trur beta=",bt, "sigma=",1,'\n')

## trur beta= 1 2.3 -1.5 0.05 sigma= 1

cat("beta_hat=",coefficients(bth), "sigma_hat=",summary(bth)$sigma,'\n') # model coefficients

## beta_hat= 0.09420014 2.370362 -1.242888 sigma_hat= 1.236286

print(summary(bth))

##
## Call:
## lm(formula = y ~ x1 + x2, data = d)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6835 -0.8128 -0.1453  0.8281  2.7005
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.09420    0.24434   0.386    0.702
## x1           2.37036    0.10717  22.118   <2e-16 ***
## x2          -1.24289    0.03006 -41.344   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 47 degrees of freedom
## Multiple R-squared:  0.9768,Adjusted R-squared:  0.9759
## F-statistic: 991.2 on 2 and 47 DF,  p-value: < 2.2e-16
plot(y~x1)
s<-seq(min(x1), max(x1),length=100)
lines(bt[1]+bt[2]*s+bt[3]*s^2+bt[4]*s^3~s,col=2, lwd=4, lty=1)
bth<-lm(y~x1,data=d)
lines(bth$coefficients[1]+bth$coefficients[2]*s~s,col=3, lwd=2, lty=1)
bth<-lm(y~x1+x2,data=d)
lines(bth$coefficients[1]+bth$coefficients[2]*s+bth$coefficients[3]*s^2~s,col=4, lwd=2, lty=1)
bth<-lm(y~x1+x2+x3,data=d)
lines(bth$coefficients[1]+bth$coefficients[2]*s+bth$coefficients[3]*s^2+bth$coefficients[4]*s^3~s,col=5


legend(4, 2, legend=c( "true", "dg=1", "dg=2", "dg=3"),
       col=c(2:5), lty=c(1,1,1,1), cex=0.8)
```

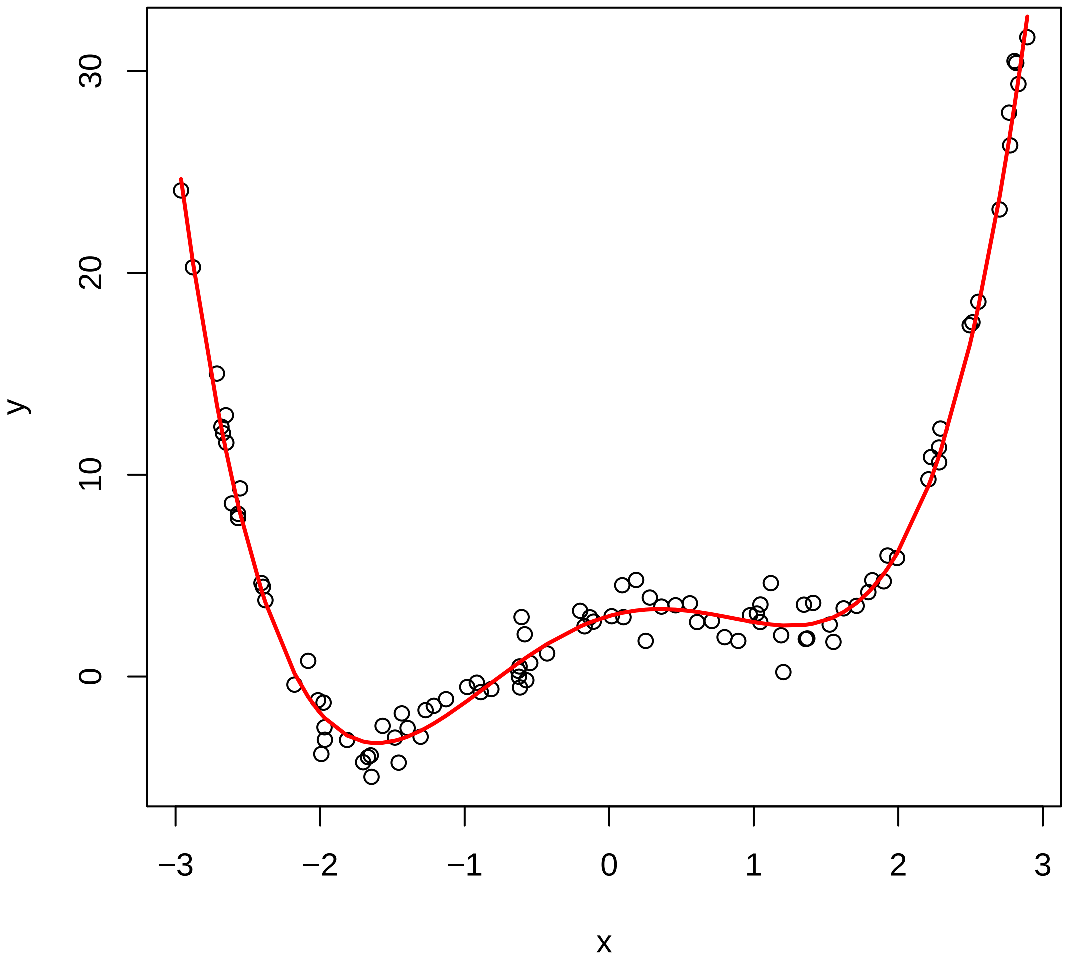



```
cat("trur beta=",bt, "sigma=",1,'\n')

## trur beta= 1 2.3 -1.5 0.05 sigma= 1

cat("beta_hat=",coefficients(bth), "sigma_hat=",summary(bth)$sigma,'\n') # model coefficients

## beta_hat= 1.064168 2.081987 -1.53616 0.06266661 sigma_hat= 0.9366696

print(summary(bth))

##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = d)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.95998 -0.55766 -0.09151  0.77623  2.21431
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.06417    0.24596   4.327 8.08e-05 ***
## x1           2.08199    0.09440  22.056  < 2e-16 ***
## x2          -1.53616    0.05400 -28.447  < 2e-16 ***
## x3           0.06267    0.01046   5.990 3.00e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9367 on 46 degrees of freedom
## Multiple R-squared:  0.987,Adjusted R-squared:  0.9861
## F-statistic:  1163 on 3 and 46 DF,  p-value: < 2.2e-16

# Orthogonal polynomial regression

set.seed(123)
n<- 100 # number of observations
bt<-c(3, 2 , -3, 0, .7)
k<-length(bt)-1   # indepencent variables
X<-array(1,dim=c(n,(k+1)))
XO<-array(1,dim=c(n,(k+1)))
sigma<-1
eps<-rnorm(n,0,sigma)
x<-sort(runif(n, -3,3))
XM<-as.matrix(poly(x,degree = k, raw = T))
X[,2:(k+1)]<-XM[,1:k]
z<-X%*%bt
y<-z+eps

fit <- lm(y~poly(x,1,raw=T))
print(fit$coefficients)
```

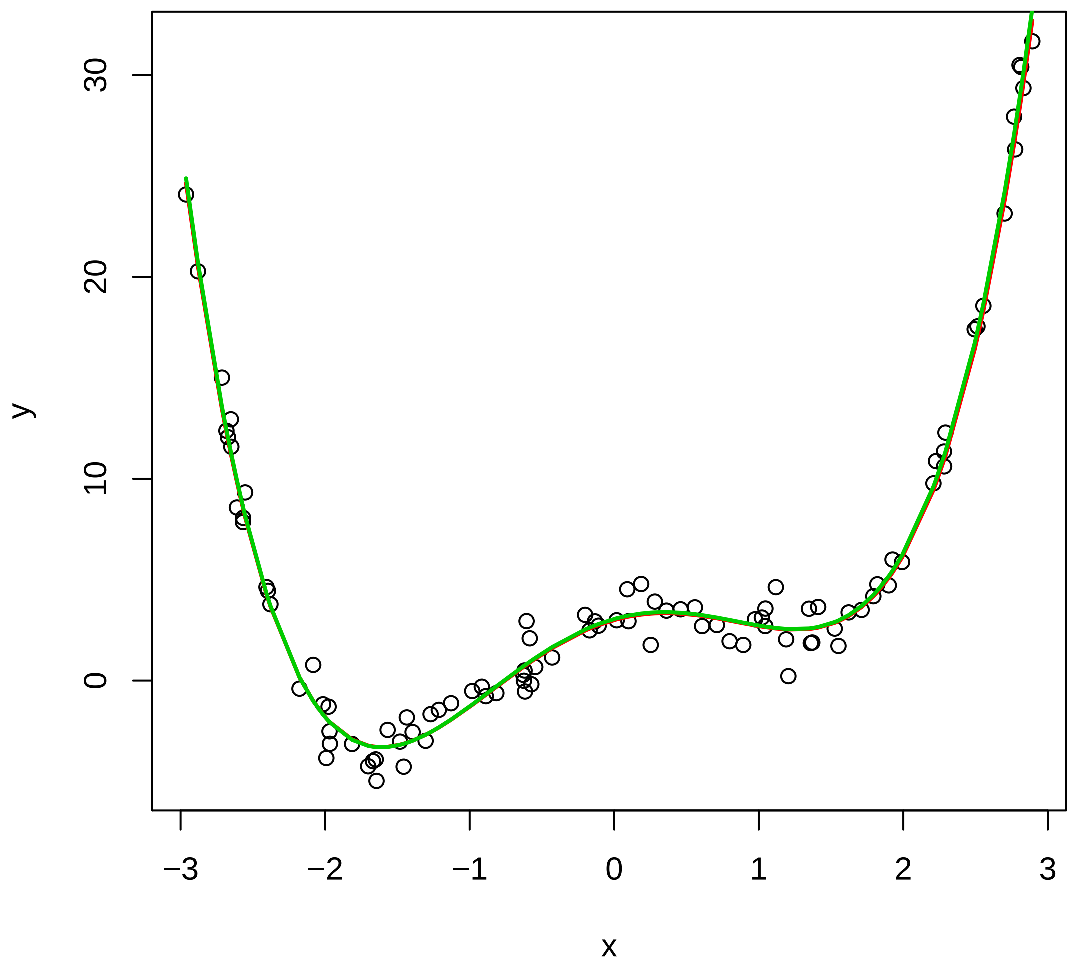
```
##         (Intercept) poly(x, 1, raw = T)
##            5.509150            2.084683
fit <- lm(y~poly(x,2,raw=T))
print(fit$coefficients)
##          (Intercept) poly(x, 2, raw = T)1 poly(x, 2, raw = T)2
##            -1.945595             2.155262             2.410901
fit <- lm(y~poly(x,3,raw=T))
print(fit$coefficients)
##          (Intercept) poly(x, 3, raw = T)1 poly(x, 3, raw = T)2
##          -1.94850117           2.02152232           2.41101853
## poly(x, 3, raw = T)3
##           0.02475886
fit <- lm(y~poly(x,4,raw=T))
print(fit$coefficients)
##          (Intercept) poly(x, 4, raw = T)1 poly(x, 4, raw = T)2
##          3.053927196          1.994692160         -3.040129157
## poly(x, 4, raw = T)3 poly(x, 4, raw = T)4
##          0.008905088          0.709775643
fitortho<- lm(y~poly(x,1))
print(fitortho$coefficients)
## (Intercept)  poly(x, 1)
##    5.334684   36.630831
fitortho<- lm(y~poly(x,2))
print(fitortho$coefficients)
## (Intercept) poly(x, 2)1 poly(x, 2)2
##    5.334684   36.630831   64.370147
fitortho<- lm(y~poly(x,3))
print(fitortho$coefficients)
## (Intercept) poly(x, 3)1 poly(x, 3)2 poly(x, 3)3
##   5.3346835  36.6308309  64.3701469   0.9977942
fitortho<- lm(y~poly(x,4))
print(fitortho$coefficients)
## (Intercept) poly(x, 4)1 poly(x, 4)2 poly(x, 4)3 poly(x, 4)4
```

```
##   5.3346835  36.6308309  64.3701469   0.9977942  40.8276005
par(mfrow=c(1,1))
plot(y~x)
par(mfrow=c(1,1))
plot(y~x)
#z<-X%*%bt
lines(z~x, col=2, lwd=2)
```



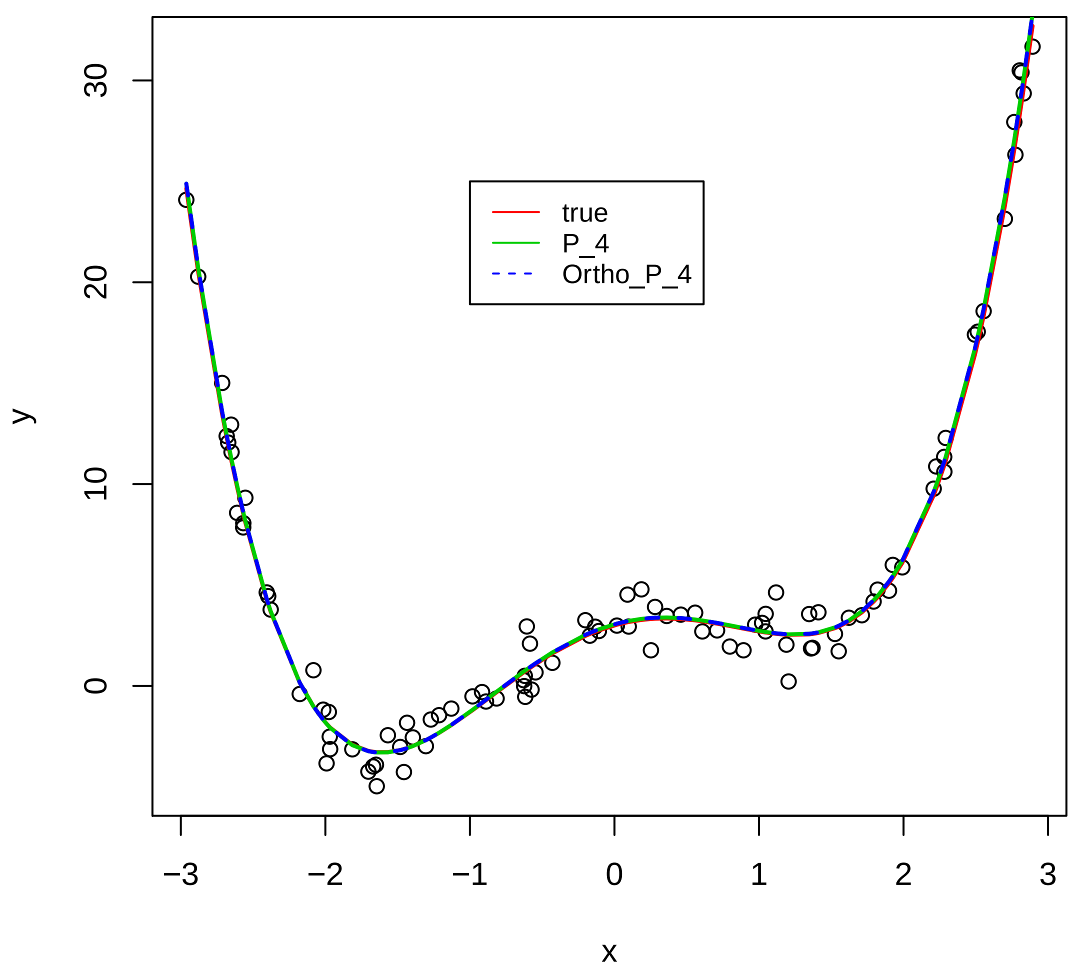

```
z1<-X%*%fit$coefficients
par(mfrow=c(1,1))
```

```
plot(y~x)
lines(z~x, col=2, lwd=2)
lines(z1~x, col=3, lwd=2)
```




```
XX<-as.matrix(poly(x,k,raw=F))
X0[,2:(k+1)]<-XX[,1:k]
#cf<-fitortho£coefficients[1:(k+1)]
z2<-X0%*%fitortho$coefficients
par(mfrow=c(1,1))
plot(y~x)
```

```
lines(z~x, col=2, lwd=2)
lines(z1~x, col=3, lwd=2)
lines(z2~x, col=4, lwd=2,lty=2)
legend(-1, 25, legend=c( "true", "P_4", "Ortho_P_4"),
       col=c(2:4), lty=c(1,1,2), cex=0.8)
```



8. Model Adequacy Checking & Diagnostics

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)**:

$$\begin{aligned} R^2 &= \frac{\text{Variation in Y explained by the mdel}}{\text{Total variation in Y}} \\ &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})}{\sum_{i=1}^n (Y_i - \bar{Y})} \\ &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})}{\sum_{i=1}^n (Y_i - \bar{Y})} \\ &= \frac{SSModel}{SSTotal} \\ &= 1 - \frac{SSRes}{SSTotal} \\ &= 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})} \end{aligned} \tag{8.1}$$

NOTE: (1) $R^2 \in (0, 1)$

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

(3) Higher value of R^2 is an indicator of the better model

So we modify it to another measure of model adequacy checking (b)**Adjusted R^2:**

$$\begin{aligned} R^2_{adjusted} &= 1 - \frac{\sum_{i=1}^n e_i^2/df}{\sum_{i=1}^n (Y_i - \bar{Y})/df} \\ &= 1 - \frac{\sum_{i=1}^n e_i^2/(n-k-1)}{\sum_{i=1}^n (Y_i - \bar{Y})/(n-1)} \end{aligned} \tag{8.2}$$

8.1. **Residual Analysis.** The residual vector $\mathbf{e} \sim N(\mathbf{0}, (\mathbf{I}_n - P_X)\sigma^2)$. Define $P_X = ((h_{ij}))$. Hence $e_i \sim N(0, \sigma^2(1 - h_{ii}))$ and $cov(e_i, e_j) = -h_{ij}\sigma^2$. Now

$$r_i = \frac{e_i}{\sqrt{MSRes(1 - h_{ii})}} \sim N(0, 1) \quad \text{asymptotically and marginally.}$$

Use $q - q$ plot to check the normality. It must overlap with the line passing through origin with slope 1. This is known as **studentized residual**.

In **Leave one out / Jack-knife** ith observation is removed from the the data and model parameters are estimated as $\hat{\beta}_{(i)}$ and $\hat{\sigma}^2 = S^2_{(i)}$. So the corresponding predicted value of y_i is $\hat{y}_{(i)}$. Then the prediction error or studentized **PRESS** residual is

$$\sum_{i=1}^n \hat{e}^2_{(i)} = \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^n \frac{e_i^2}{\sigma^2(1 - h_{ii})} \quad \forall i = 1, 2, \cdots n$$

If we use $\hat{\sigma}^2 = MSRes$ then it becomes same as studentized residual. But if we use

$$\hat{\sigma}^2 = S^2_{(i)} = \frac{(n-k-1)MSRes - e_i^2/(1 - h_{ii})}{n-k-2}$$

then **externally studentize residual** or **R-student**

$$t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}}$$

can be used for **outlier** detection.

8.2. **Diagnostics of leverage and influence point.**

Definition 8.1. Leverage point: A data point which is staying out of the cluster affecting the values of R^2 or $\hat{\sigma}^2$, but not much affecting the value of $\hat{\beta}$.

An indicator of leverage can be $h_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$. Trace of H is $k+1$ and H is idempotent. So the average of the diagonal is $(k+1)/n$. A thumb-rule to consider an observation as a leverage point if $h_{ii} > 2(k+1)/n$ provided $2(k+1)/n < 1$.

Definition 8.2. Influential point: Observations with large hat diagonal and large residual can be an influential point.

Measure of influence point:

Cook's Distance[1977,79] It measures the deviation on $\hat{\mathbf{y}}$ because of the ith observation by

$$D_i = \frac{(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})^T(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})}{(k+1)MSRes} = \frac{e_i^2\ h_{ii}}{(k+1)MSRes(1-h_{ii})^2} = \frac{r_i^2\ h_{ii}}{(k+1)(1-h_{ii})}$$

and if D_i is larger than the median of $F_{k+1,n-k-1} \approxeq 1$, it is considered as influence point.

Belsley, Kuh, Welsch's [1980] measures: It measures the change in $\hat{\beta}_j$ because of the *ith* observation

$$DFBETAS_{j,(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S^2_{(i)}C_{j,j}}} = \frac{r_{j,i}\ e_i}{\sqrt{\mathbf{r_j^T r_j} S^2_{(i)}(1-h_{ii})^2}} = \frac{r_{j,i}\ t_i}{\sqrt{\mathbf{r_j^T r_j}(1-h_{ii})}}$$

where, $((C_{ij})) = (X^TX)^{-1}$ and $\mathbf{R} = (X^TX)^{-1}X^T$ and hence, $((C_{i,j})) = \mathbf{R}^T\mathbf{R}$.

Consider *ith* observation as an influential factor if $|DFBETAS_{j,(i)}| > 2/\sqrt{n}$. It measures both leverage and effect on residual. To measures the change in $\hat{y}_i$ because of deletion of the *ith* observation use

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S^2_{(i)}h_{ii}}} = t_i\sqrt{\left(\frac{h_{ii}}{1-h_{ii}}\right)}$$

If the data is an outlier or has high leverage $DFFITS_i$ will have high magnitude. So consider the *ith* observation as influence point if $|DFFITS_i| > 2\sqrt{(k+1)/n}$ The precision of the estimation is measured by

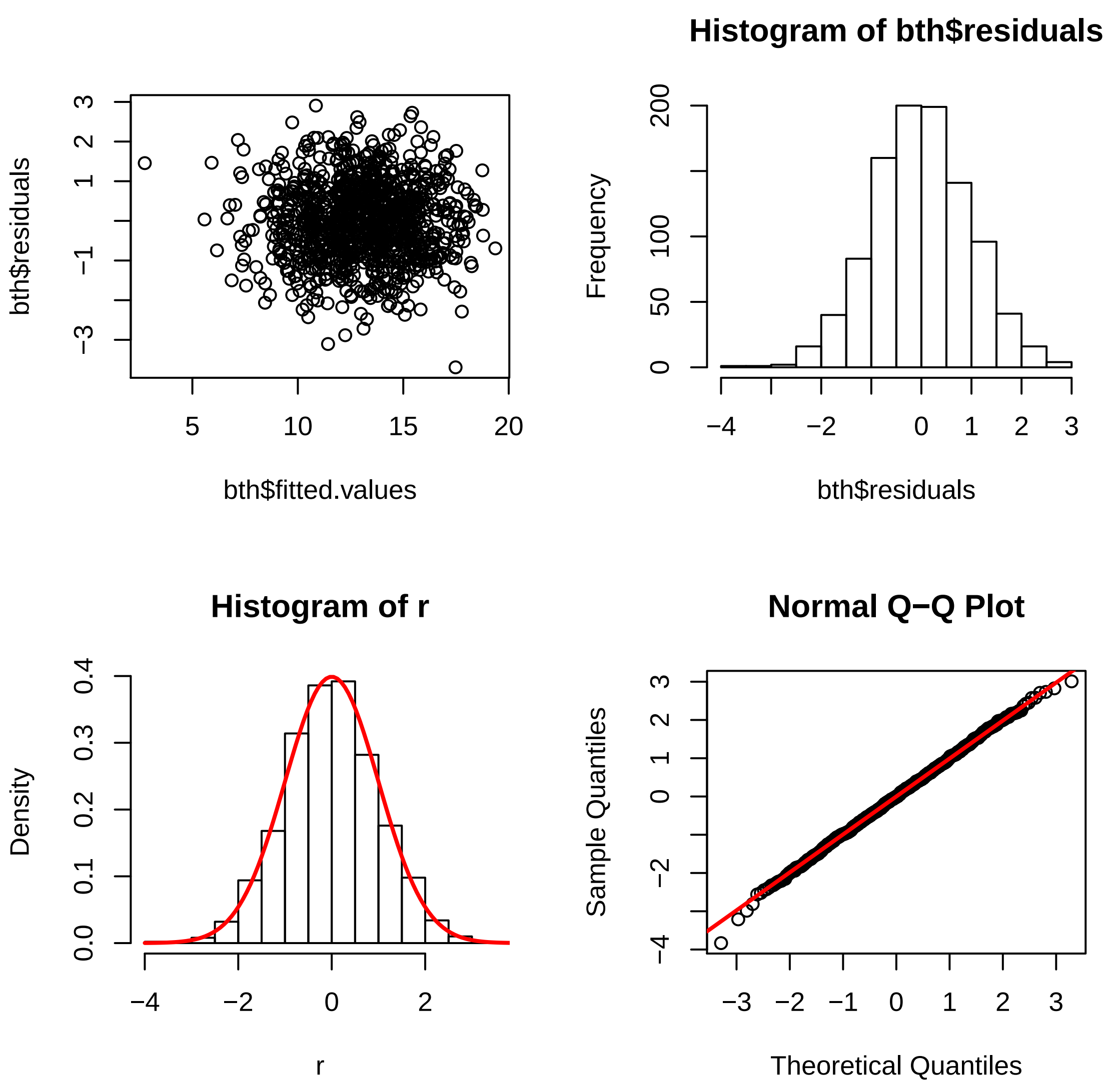
$$COVRATIO_i = \frac{|\widehat{Var(\hat{\beta}_{(i)})}|}{|\widehat{Var(\hat{\beta})}|} = \frac{|(X^T_{(i)}X_{(i)})^{-1}S^2_{(i)}|}{|(X^TX)^{-1}MSRes|} = \left(\frac{S^2_{(i)}}{MSRes}\right)^{k+1}\left(\frac{1}{1-h_{ii}}\right)$$

Consider the *ith* observation as an high leverage point if $COVRATIO_i > 1+3(k+1)/n$ or $COVRATIO_i < 1-3(k+1)/n$ provided $1 > 3(k+1)/n$ **Suggested reading Ch 4 and Ch 6 of Introduction to Linear**

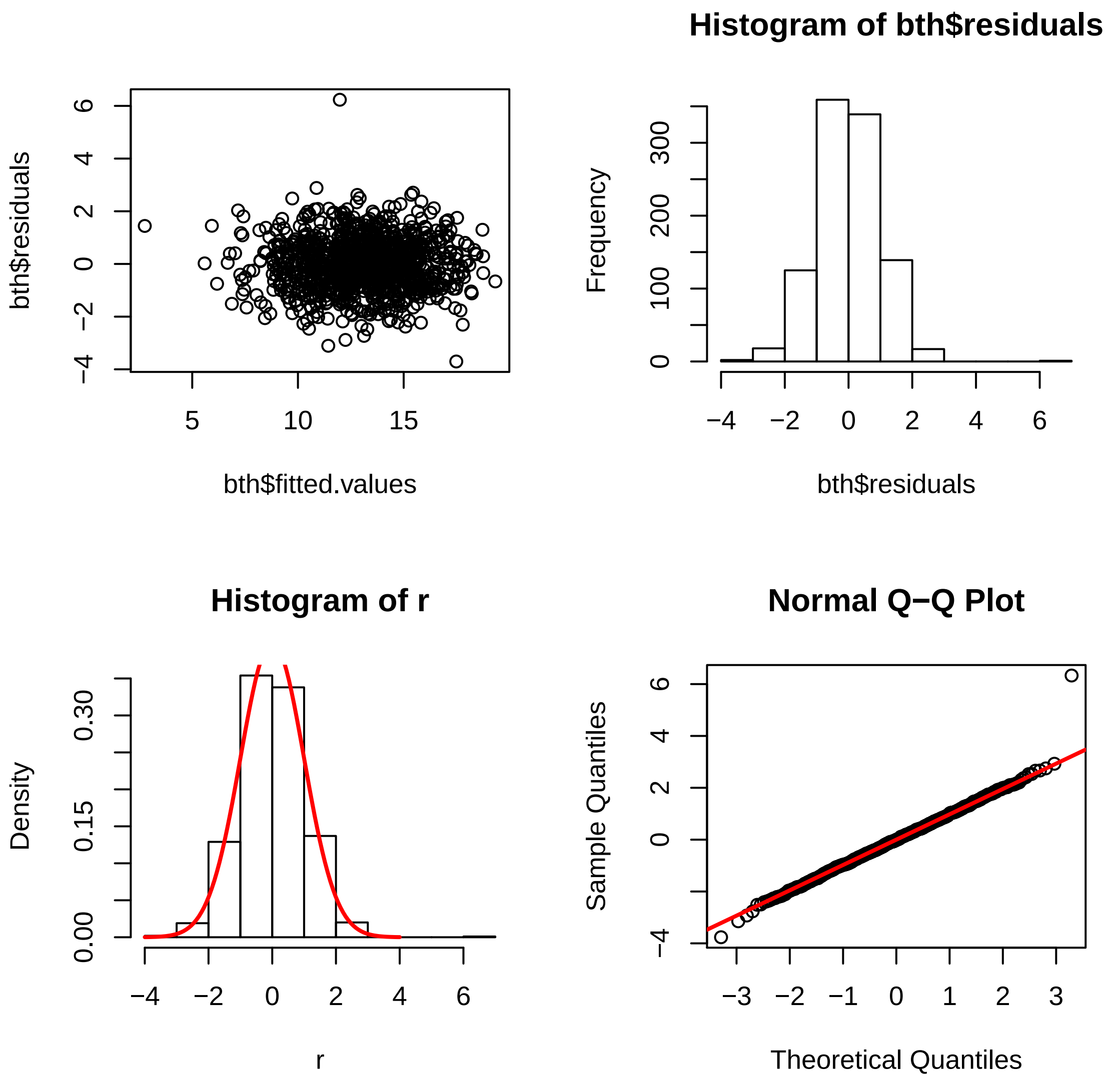
Regression Analysis, 5th Edition By Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining

```
# Multiple linear regression
####################################
k<-3 # indepencent variables
n<- 1000 # number of observations
X<-array(0,dim=c(n,(k+1)))
bt<-c(1,2.3, 1.5,0.05)

x1<-rgamma(n,2,3)
x2<-rbinom(n,10,0.7)
x3<-rbeta(n,0.5,0.5)
X[,1]<-1
X[,2]<-x1
X[,3]<-x2
X[,4]<-x3
eps<-rnorm(n,0,1)
y<-X%*%bt+eps ##### OR  y<-bt[1]+bt[2]*x1+bt[3]*x2+bt[4]*x3+eps
d<- data.frame(y,x1,x2,x3)

# #------Estimation-------------

# Create the relationship model.

bth<-lm(y~x1+x2+x3,data=d)
sigmah<-summary(bth)$sigma
H<-(X%*%solve(t(X)%*%X)%*%t(X))
C<-diag(diag(1, nrow = n)-H)
r<-bth$residuals/(sigmah*C)

cat("trur beta=",bt, "sigma=",1,'\n')
## trur beta= 1 2.3 1.5 0.05 sigma= 1
cat("beta_hat=",coefficients(bth), "sigma_hat=",summary(bth)$sigma,'\n') # model coefficients
## beta_hat= 1.041935 2.17495 1.498535 0.06988376 sigma_hat= 0.9687723
```

```
par(mfrow=c(2,2))
plot(bth$residuals~bth$fitted.values)
hist(bth$residuals)
hist(r, breaks = 10, probability = T )
s<-seq(-4,4,length=100)
lines(dnorm(s,0,1)~s, col=2, lwd=2)
qqnorm(r)
qqline(r, col=2,lwd=2)
```

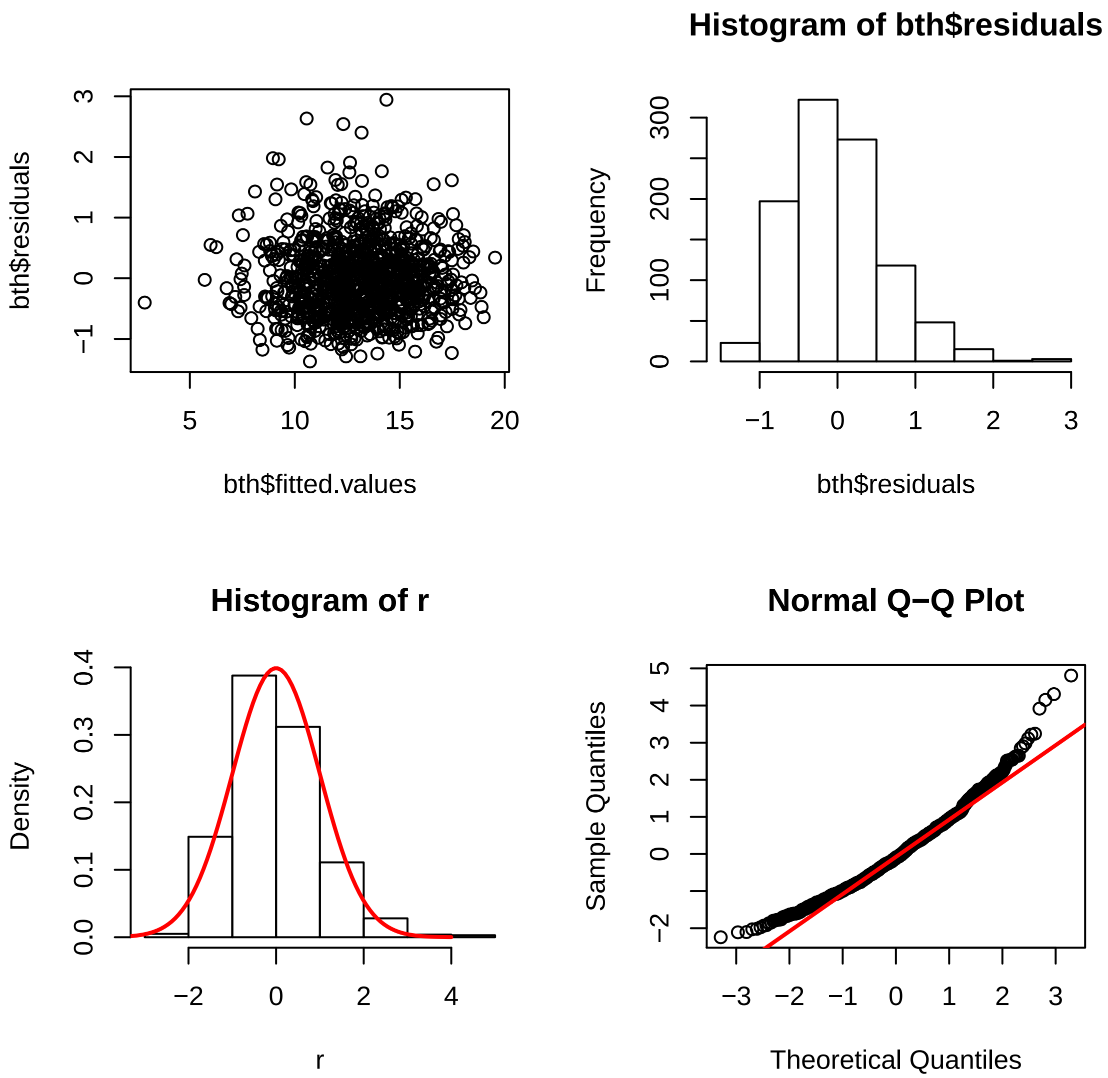


```r
# with outlier
y[10]<-y[10]+7
d<- data.frame(y,x1,x2,x3)

bth<-lm(y~x1+x2+x3,data=d)
sigmah<-summary(bth)$sigma
H<-(X%*%solve(t(X)%*%X)%*%t(X))
C<-diag(diag(1, nrow = n)-H)
r<-bth$residuals/(sigmah*C)

cat("trur beta=",bt, "sigma=",1,'\n')
## trur beta= 1 2.3 1.5 0.05 sigma= 1
cat("beta_hat=",coefficients(bth), "sigma_hat=",summary(bth)$sigma,'\n') # model coefficients
## beta_hat= 1.049491 2.156919 1.498163 0.09732973 sigma_hat= 0.9885
par(mfrow=c(2,2))
plot(bth$residuals~bth$fitted.values)
hist(bth$residuals)
hist(r, breaks = 10, probability = T )
s<-seq(-4,4,length=100)
lines(dnorm(s,0,1)~s, col=2, lwd=2)
qqnorm(r)
qqline(r, col=2,lwd=2)
```



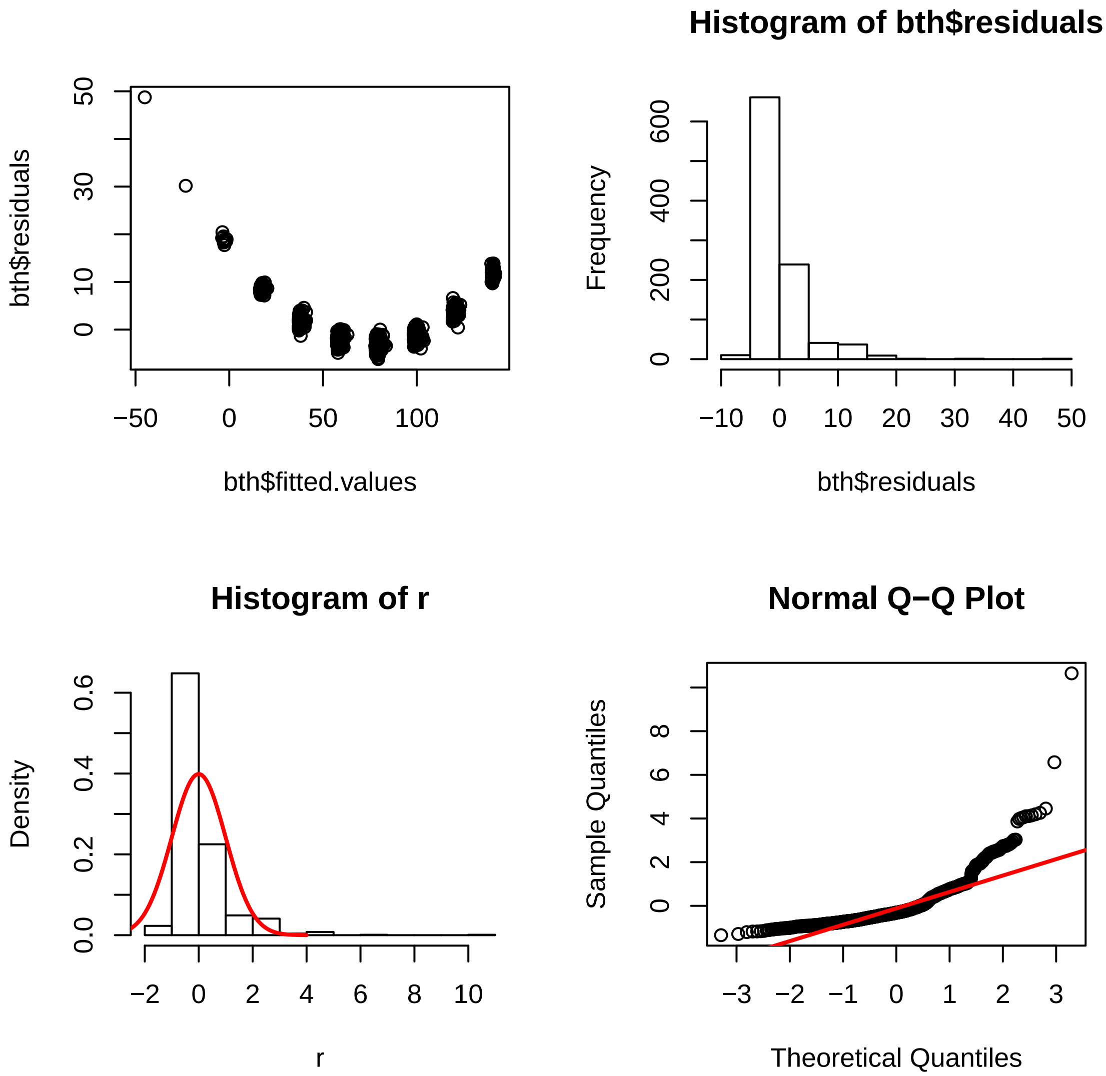

```
# non normal error

eps1<-rgamma(n,shape = 10,rate = 5)-2
y<-X%*%bt+eps1 ##### OR  y<-bt[1]+bt[2]*x1+bt[3]*x2+bt[4]*x3+eps
d<- data.frame(y,x1,x2,x3)

# #------Estimation-------------

# Create the relationship model.
```

```
bth<-lm(y~x1+x2+x3,data=d)
sigmah<-summary(bth)$sigma
H<-(X%*%solve(t(X)%*%X)%*%t(X))
C<-diag(diag(1, nrow = n)-H)
r<-bth$residuals/(sigmah*C)

cat("trur beta=",bt, "sigma=",1,'\n')
## trur beta= 1 2.3 1.5 0.05 sigma= 1
cat("beta_hat=",coefficients(bth), "sigma_hat=",summary(bth)$sigma,'\n') # model coefficients
## beta_hat= 1.166347 2.265942 1.482583 0.03178569 sigma_hat= 0.6142851
par(mfrow=c(2,2))
plot(bth$residuals~bth$fitted.values)
hist(bth$residuals)
hist(r, breaks = 10, probability = T )
s<-seq(-4,4,length=100)
lines(dnorm(s,0,1)~s, col=2, lwd=2)
qqnorm(r)
qqline(r, col=2,lwd=2)
```




```
## Wrong model
##
X[,1]<-1
X[,2]<-x1
X[,3]<-x2^2 ## model chanded
X[,4]<-x3
eps<-rnorm(n,0,1)
y<-X%*%bt+eps ##### OR  y<-bt[1]+bt[2]*x1+bt[3]*x2+bt[4]*x3+eps
d<- data.frame(y,x1,x2,x3)
```

```
bth<-lm(y~x1+x2+x3,data=d)
sigmah<-summary(bth)$sigma
H<-(X%*%solve(t(X)%*%X)%*%t(X))
C<-diag(diag(1, nrow = n)-H)
r<-bth$residuals/(sigmah*C)

cat("trur beta=",bt, "sigma=",1,'\n')
## trur beta= 1 2.3 1.5 0.05 sigma= 1
cat("beta_hat=",coefficients(bth), "sigma_hat=",summary(bth)$sigma,'\n') # model coefficients
## beta_hat= -65.83592 1.948325 20.51509 0.4095548 sigma_hat= 4.619644
par(mfrow=c(2,2))
plot(bth$residuals~bth$fitted.values)
hist(bth$residuals)
hist(r, breaks = 10, probability = T )
s<-seq(-4,4,length=100)
lines(dnorm(s,0,1)~s, col=2, lwd=2)
qqnorm(r)
qqline(r, col=2,lwd=2)
```


bth$residuals
bth$fitted.values
Histogram of bth$residuals
Frequency
bth$residuals
Histogram of r
Density
r
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles

9. Exercise

(1) In each of the following, find precisely which axioms in the definition of a vector space are violated. Take $V = \mathbb{R}^2$ and $F = \mathbb{R}$ throughout.

(a) $(x_1, x_2) + (y_1, y_2) = (x_1 + y_1, 0), \alpha(x_1, x_2) = (\alpha x_1, 0)$

(b) $(x_1, x_2) + (y_1, y_2) = (x_1 + y_1, x_2 + y_2), \alpha(x_1, x_2) = (\alpha x_1, 0)$

(c) $(x_1, x_2) + (y_1, y_2) = (x_1 + y_1, x_2 + y_2), \alpha(x_1, x_2) = (\alpha x_1, 2\alpha x_2)$

(d) $(x_1, x_2) + (y_1, y_2) = (x_1 + y_1, x_2 + y_2), \alpha(x_1, x_2) = (\alpha + x_1, \alpha + x_2)$

(2) Show that the set of all positive real numbers forms a vector space over $\mathbb{R}$ if the sum of x and y is defined to be the usual product xy and α times x is defined to be x^α

(3) Show that the set of all positive real numbers forms a vector space over $\mathbb{R}$ if the sum of x and y is defined to be the usual product xy and α times x is defined to be x^α

(4) In each of the following, find out whether the subsets given form subspaces of the vector space V.

(a) $V = \mathbb{R}^3, S = \{(x_1, x_2, x_3) : 2x_1 + x_2 + x_3 = 1\}$.

(b) $V = \mathbb{R}^2, S=$ the set of all (x_1, x_2) such that $x_1 \geq 0$ and $x_2 \geq 0$ and $T =$ the set of all (x_1, x_2) such that $x_1 x_2 \geq 0$.

(5) (a) Consider the vectors $x_1 = (1, 3, 2)$ and $x_2 = (-2, 4, 3)$ in $\mathbb{R}^3$. Show that the span of $\{x_1, x_2\}$ is

$\{(\xi_1, \xi_2, \xi_3) : \xi_1 - 7\xi_2 + 10\xi_3 = 0\} = \{(\alpha, \beta, (-\alpha + 7\beta)/10) : \alpha, \beta \in \mathbb{R}\}$

(b)Consider the vectors $x_1 = (1, 2, 1, -1)$, $x_2 = (2, 4, 1, 1)$, $x_3 = (-1, -2, -2, -4)$ and $x_4 = (3, 6, 2, 0)$ in $\mathbb{R}^4$. Show that the span of $\{x_1, ..., x_4\}$ is

$\{(\xi_1, \xi_2, \xi_3, \xi_3) : 2\xi_1 - \xi_2 = 0, 2\xi_1 - 3\xi_3 - \xi_4 = 0\}$

Show that the subspace can be written as $\{(\alpha, 2\alpha, \beta, 2\alpha - 3\beta) : \alpha, \beta \in \mathbb{R}\}$

(6) Find a basis of each of the following subspaces of $\mathbb{R}^4$. Also express S_3 in the form $\{(x_1, x_2, x_3, x_4) : \frac{x_1}{a_1} = \frac{x_2}{a_2} = \frac{x_3}{a_3} = \frac{x_4}{a_4}\}$.

(a) $S_1 = \{(x_1, x_2, x_3, x_4) : x_1 - 2x_3 + x_4 = 0\}$

(b) $S_2 = \{(x_1, x_2, x_3, x_4) : x_1 + x_2 - x_3 = 0, x_2 + 2x_3 - x_4 = 0, 2x_1 + 3x_2 - x_4 = 0\}$

(c) $S_3 = \{(x_1, x_2, x_3, x_4) : x_1 + x_2 - x_3 = 0, x_1 + x_2 + 2x_3 + x_4 = 0, x_1 - 3x_2 - x_3 + 2x_4 = 0\}$.

(7) If A is an n x n matrix and x, y are n x 1 vectors, show that $x^T Ay = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i y_j$. Without using this, prove that $x^T Ay = y^T Ax$ if A is symmetric.

(8) Let A be a square matrix and $B = \frac{1}{2}\ (A + A^T)$. Prove the following

(a) B is symmetric.

(b) $x^T Bx = x^T Ax$ for all n x 1 vectors x.

(c) If C is a symmetric matrix such that $x^T Cx = x^T Ax$ for all x, then $C = B$.

(9) Let P the n x n matrix,

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

(a) If $x = (x_1, x_2, ..., x_n)^T$, show that $Px = (x_n, x_{n-1}, ..., x_1)^T$.

(b) Show that $P^T = P$ and $P^2 = I$.

(c) If $A = ((a_{ij}))$ is an n x n matrix, what is the (i,j)-th element of PAP? Write down PAP when $n = 3$ and when $n = 4$

(d) If A is upper triangular, what can you say about PAP?

(10) Let A be a symmetric nonnegative matrix. Show that A is positive definite if and only if A^{-1} is positive definite.

NOTE: Consider Gauss-Markov multiple linear regression model and answer the following questions.

(11) Show that $\hat{\beta}_1$ is a linear estimator of β_1.

(12) Show that $\hat{\beta}_1$ is an unbiased estimator of β_1.

(13) Show that $\hat{\beta}_0$ is a linear estimator of β_0.

(14) Show that $\hat{\beta}_0$ is an unbiased estimator of β_0.

(15) Show that $MSR = \frac{SSR}{n-2}$ is an unbiased estimator of σ^2.

(16) Show that $\sum_i y_i = \sum_i \hat{y}_i$

(17) Show that the regression line passes through $(\bar{x}, \bar{y})$.

(18) find the 95% confidence interval of β_0 and β_1.

(19) Test that a regression line passes through origin.

(20) Test that a regression line is horizontal.

(21) Find the prediction value, and its interval for some new x_0.

(22) Obtain the maximum likelihood estimators of the model parameters $\beta_0, \beta_1, \sigma^2$. Are they same as LS estimators?

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

(24) Perform a large sample test for $H_0 : \rho = 0$ Vs $H_1 : \rho \neq 0$.

NOTE: Consider Gauss-Markov multiple linear regression model and answer the following questions.

(25) Show that $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$ if $|\mathbf{X}^T\mathbf{X}| \neq 0$.

(26) Show that $SSR = (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}}) \sim \sigma^2\chi^2_{n-k-1}$. Hence find an unbiased estimator of σ^2.

(27) Find the maximum likelihood estimate of β and σ^2. Are they same with the least square estimator?

(28) Show that $l^T\mathbf{y}$ is an LUE of $p^T\beta$ iff $\mathbf{X}^T l = p$.

(29) Show that $l^T\mathbf{y}$ is an LZF iff $\mathbf{X}^T l = \mathbf{0}$.

(30) Prove that the necessary and sufficient condition for $p^T\beta$ to be estimable is $p \in \mathcal{C}(\mathbf{X}^T)$.

(31) Show that every estimable linear parametric function has unique blue.

(32) 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 $p^T\beta$

(33) Construct a test for $H_0 : \beta_i = b_i$ vs $H_1 : \beta_i \neq b_i$.

(34) Construct a test for $H_0 : \beta_i - \beta_j = k$ vs $H_1 : \beta_i - \beta_j \neq k$.

(35) Construct a test for $H_0 : \beta_i - \theta\beta_j = k$ vs $H_1 : \beta_i - \theta\beta_j \neq k$.

(36) Construct a test for $H_0 : \mathbf{A}\beta = b$ vs $H_1 : \mathbf{A}\beta \neq b$ where $Rank(\mathbf{A}) = r < k+1$

(37) For a new vector $\mathbf{x}_0 = (1, x_{01}, x_{02}, \cdots, x_{0k})^T$

(a)Predict the value of y_0

(b) Estimate the prediction error.

(c) Find a $100(1-\alpha)\%$ prediction interval of the predicted value.

(38) Perform the ANOVA for multiple linear regression.

NOTE: Consider orthogonal polynomial model

(39) Show that $\hat{\alpha}_0 = \bar{y}$ and $\hat{\alpha}_j = \frac{\sum_{i=1}^n P_j(x_i)y_i}{\sum_{i=1}^n P_j^2(x_i)}$

(40) Show that $SSR = SST - \sum_{j=1}^k (\hat{\alpha}_j \sum_{i=1}^n P_j(x_i)y_i)$. Hence, construct a test for $H_0 : \alpha_j = 0$ vs $H_1 : \alpha_j \neq 0$

(41) Show that $R^2_{adjusted} < R^2$.

(42) Prove that that the PRESS residual $\hat{e}_{(i)} = y_i - \hat{y}_{(i)} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$ $\forall i = 1, 2, \cdots n$

(43) Prove Cook's distance

$$D_i = \frac{(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})^T(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})}{(k+1)MSRes} = \frac{e_i^2\ h_{ii}}{(k+1)MSRes(1-h_{ii})^2} = \frac{r_i^2\ h_{ii}}{(k+1)(1-h_{ii})}$$

(44) Prove that

$$DFBETAS_{j,(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S^2_{(i)}C_{j,j}}} = \frac{r_{j,i}\ e_i}{\sqrt{\mathbf{r_j^T r_j} S^2_{(i)}(1-h_{ii})^2}} = \frac{r_{j,i}\ t_i}{\sqrt{\mathbf{r_j^T r_j}(1-h_{ii})}}$$

(45) Prove that

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S^2_{(i)}h_{ii}}} = t_i\sqrt{\left(\frac{h_{ii}}{1-h_{ii}}\right)}$$

(46) Prove that

$$COVRATIO_i = \frac{|\widehat{Var(\hat{\beta}_{(i)})}|}{|\widehat{Var(\hat{\beta})}|} = \frac{|(X_{(i)}^T X_{(i)})^{-1}S^2_{(i)}|}{|(X^TX)^{-1}MSRes|} = \left(\frac{S^2_{(i)}}{MSRes}\right)^{k+1}\left(\frac{1}{1-h_{ii}}\right)$$