FAMILY NAME: ....
OTHER NAME(S): ....
STUDENT NUMBER: ....
SIGNATURE: ....

THE UNIVERSITY OF NEW SOUTH WALES

SCHOOL OF MATHEMATICS AND STATISTICS

Semester 2, 2014

MATH2089
Numerical Methods and Statistics

(1) TIME ALLOWED – 3 Hours

(2) TOTAL NUMBER OF QUESTIONS – 6

(3) ANSWER ALL QUESTIONS

(4) THE QUESTIONS ARE OF EQUAL VALUE

(5) THIS PAPER MAY **NOT** BE RETAINED BY THE CANDIDATE

(6) **ONLY** CALCULATORS WITH AN AFFIXED “UNSW APPROVED” STICKER MAY BE USED

(7) STATISTICAL FORMULAE ARE ATTACHED AT END OF PAPER
STATISTICAL TABLES ARE ATTACHED AT END OF PAPER

Part A – Numerical Methods consists of questions 1 – 3

Part B – Statistics consists of questions 4 – 6

Both parts must be answered

All answers must be written in ink. Except where they are expressly required pencils may only be used for drawing, sketching or graphical work.

Part A – Numerical Methods

1. Answer in a separate book marked Question 1

a) Give the results of the following MATLAB commands. Assume `eps` $\approx 2 \times 10^{-16}$ is the relative machine precision in MATLAB .

i)
```
ans1  = ( 1 + eps/16 ) - 1
```
ii)
```
v = [-2:2:2]
ans2 = v./v
```
iii)
```
ans3 = exp(1e100)
```

b) The computational complexity of some common operations with n by n matrices are given in Table 1.1

Operation	Flops
Matrix multiplication	$2n^3$
LU factorization	$\frac{2n^3}{3} + O(n^2)$
Cholesky factorization	$\frac{n^3}{3} + O(n^2)$
Back/forward substitution	$n^2 + O(n)$
Tridiagonal solve	$8n + O(1)$

Table 1.1: Flops for some operations with n by n matrices

i) The coefficient matrix of a linear system has no special structure. Estimate the size n of the largest linear system that can be solved in one hour on a 3 GHz quad core computer, which can do two floating point operations per clock cycle.

ii) On a different computer, multiplying two large $n \times n$ matrices takes 1000 seconds. Estimate how long it will take to solve an $n \times n$ symmetric positive definite linear system via Cholesky factorization for two different right-hand-side vectors.

c) Consider a large linear system $A\boldsymbol{x} = \boldsymbol{b}$. You know the right hand side vector $\boldsymbol{b}$ to 6 significant figures. You have the following output from MATLAB

```
nmA    = norm(A)
nmA =
   100.0000
nmAinv = norm(inv(A))
nmAinv =
   100.0000
emin   = min(eig(A))
emin =
    0.0100
emax   = max(eig(A))
emax =
   100.0000
chk    = norm(A-A')
chk =
   1.0000e-16
```

i) Is A symmetric?

ii) Estimate the 2-norm condition number $\kappa_2(A)$

iii) Estimate the 2-norm condition number $\kappa_2(A^{-1})$

iv) Now assume you know the matrix A exactly. Estimate the relative error in the computed solution $\boldsymbol{x}$.

v) How many significant figures do you expect in the computed solution?

vi) Suppose you want to ensure 5 significant figures in the computed solution. How accurately, in terms of the number of significant figures, do you need to know the right hand side vector $\boldsymbol{b}$?

d) i) The maximum eigenvalue of a matrix A is denoted $\lambda_{\max}(A)$.

Claim: The matrix 2-norm satisfies $||A||_2 = \sqrt{\lambda_{\max}(A^T A)}$.
Answer *True or False*.

ii) Orthogonal matrices Q, which by definition satisfy $Q^T Q = I$, are widely used in numerical methods and least squares problems.

Claim: One reason for this is that orthogonal matrices have the ideal condition number $\kappa_2(Q) = 1$.
Answer *True or False*, and give a short reason.

2. Answer in a separate book marked Question 2

a) You already know the possible answers $x^* = \pm 3$, but you decide to test Newton's method by using it to find a root of

$$f(x) = x^2 - 9.$$

When $J(x)$ is the derivative, Newton's method finds the new estimate x_{k+1} from the old estimate x_k by solving

$$J(x_k)(x_{k+1} - x_k) = -f(x_k).$$

You implement Newton's method on a computer for this example, which results in the following table of errors.

k	e(k)	e(k+1)/e(k)	e(k+1)/e(k)^2	e(k+1)/e(k)^3
1	1.000000000000	0.125000000000	0.125000000000	1.250000000e-01
2	0.125000000000	0.020000000000	0.160000000000	1.280000000e+00
3	0.002500000000	0.000416319734	0.166527893413	6.661115737e+01
4	0.000001040799	0.000000173232	0.166441738562	1.599172224e+05

i) Starting at your initial guess $x_1 = 4$, find the next iterate in Newton's method, x_2.

ii) Select the best answer, and give a short reason. The order of convergence is

A linear

B super linear

C quadratic

D cubic

b) The matrix A has many more rows than columns. Suppose a vector $\boldsymbol{x}$ satisfies the *normal equations*:

$$A^T A\boldsymbol{x} = A^T\boldsymbol{b}.$$

Select the best answer. You expect that the vector $\boldsymbol{x}$

A exactly solves $A\boldsymbol{x} = \boldsymbol{b}$

B minimizes $||A\boldsymbol{x} - \boldsymbol{b}||_2^2$

c) You are given the QR factorization $A = QR$, and the condition number $\kappa_2(A) = 10^6$.

i) Find $\kappa_2(A^T A)$.

ii) Find $\kappa_2(R)$.

iii) For this example, you solve a least squares problem. Would you use the normal equations or the QR factorization? Give a reason for your answer.

d) Consider the data values y_j measured at the times t_j for $j = 1, 2, 3, 4, 5$, given in the Table:

j	1	2	3	4	5
t_j	1	2	3	4	5
y_j	11.9	14.3	16.1	17.9	20.1

Table 2.1: Data

The data are stored in MATLAB column vectors `tdat` and `ydat`. You want to find the line of best fit of the form $y = \alpha + \beta t$. Suppose you use MATLAB to calculate the least squares approximation of this form. (You do not need to evaluate numerical values for α and β.) For each of the following, give MATLAB command(s) to:

i) form the matrix A.

ii) solve the appropriate linear system and store the result in a MATLAB vector `x`.

iii) calculate the residual sum of squares.

e) Two chemicals, A and B, are mixed under controlled conditions to produce a product C, according to the irreversible chemical reaction

$$A + B \longrightarrow C.$$

The process is modelled mathematically by the following system of ordinary differential equations (ODEs):

$$\begin{aligned}
\frac{dx_1}{dt} + kx_1x_2 &= 0 \\
\frac{dx_2}{dt} + kx_1x_2 &= 0 \\
\frac{dx_3}{dt} - kx_1x_2 &= 0,
\end{aligned}$$

where

- $x_1 =$ concentration of chemical A (moles/m^3)
- $x_2 =$ concentration of chemical B (moles/m^3)
- $x_3 =$ concentration of product C (moles/m^3)
- $k = 2 =$ reaction rate constant (moles/m^3-min)

Initially, there are 3 moles/m^3 of chemical A and 4 moles/m^3 of chemical B, and no product. We want to simulate the changing concentrations over time by solving the system of ODEs.

i) Write this as an initial value problem (IVP) in the standard form

$$\mathbf{x}' = \mathbf{f}(t, \mathbf{x})$$

ii) What is the initial condition $\boldsymbol{x}_0 = \boldsymbol{x}(0)$?

iii) Write

- **EITHER** a MATLAB anonymous function `myode`
- **OR** a MATLAB function M-file `myode.m`

to evaluate the vector valued function $\mathbf{f}(t, \mathbf{x})$.

iv) Use Euler's method with a step size of $h = 0.1$ to estimate $\boldsymbol{x}(0.1)$.

3. Answer in a separate book marked Question 3

The temperature $u(x,t)$ in an insulated wire of length L satisfies the partial differential equation

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2}, \quad x \in (0,L), \quad t \in (0,T]. \tag{3.1}$$

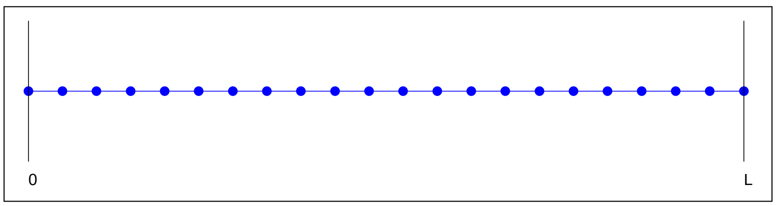
Let u_j^ℓ be your discrete finite difference approximation to the exact continuous solution $u(x_j, t_\ell)$ of (3.1) at the grid points

$$\begin{aligned} x_j &= j\Delta x \quad \text{for } j = 0, \ldots, n+1, \\ t_\ell &= \ell \Delta t \quad \text{for } \ell = 0, \ldots, m, \end{aligned}$$

where $L = (n+1)\Delta x$ and $T = m\Delta t$. At each time-step you solve the linear system:

$$(\boldsymbol{u}^{\ell+1} - \boldsymbol{u}^{\ell})/\Delta t = -A\boldsymbol{u}^{\ell+1}/(\Delta x^2) \tag{3.2}$$

The space grid and the non-zero entries in the $n \times n$ *integer* coefficient matrix A are in Figure 3.1 for $n = 20$.



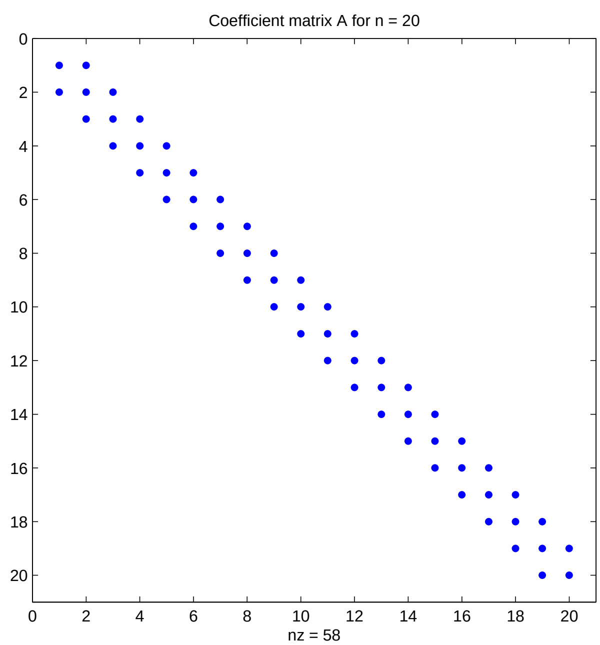


Figure 3.1: Grid and non-zero elements in coefficient matrix

You are **given** the following standard finite difference approximations for a function f of **one** variable:

$$\begin{aligned} f'(x) &= \frac{f(x+h)-f(x)}{h} + O(h), \\ f'(x) &= \frac{f(x+h)-f(x-h)}{2h} + O(h^2), \\ f''(x) &= \frac{f(x+h)-2f(x)+f(x-h)}{h^2} + O(h^2). \end{aligned}$$

The right hand side of (3.2), $-A\boldsymbol{u}^\ell/(\Delta x^2)$, is the above standard central difference approximation of $O(\Delta x^2)$ to the second derivative at the spatial grid points.

a) What additional information is needed to completely specify this problem?

b) The left hand side of (3.2) is a standard finite difference approximation to the time derivative at the grid points for the continuous problem (3.1):

$$(\boldsymbol{u}^{\ell+1} - \boldsymbol{u}^\ell)/\Delta t.$$

What is the order of accuracy of this approximation ?

c) Do the finite differences in (3.2) lead to an *implicit method* or to an *explicit method*? Give a reason for your answer.

d) Give one advantage of an implicit method compared to an explicit method.

e) You are given that $u(0,t) = u(L,t) = 0$. Give the missing numbers (represented by question marks).

   i) The first row of the matrix A is $\begin{bmatrix} ? & ? & 0 & \dots & 0 \end{bmatrix}$

   ii) The second row of the matrix A is $\begin{bmatrix} ? & ? & ? & 0 & \dots & 0 \end{bmatrix}$

   iii) The last row of the matrix A is $\begin{bmatrix} 0 & \dots & 0 & ? & ? \end{bmatrix}$

   iv) For the same matrix A, give the results of the following MATLAB commands:

```
j = 2;
v = [1:n]';
w = A*(v.^0);
ans0 = w(j)
w = A*(v.^1);
ans1 = w(j)
w = A*(v.^2);
ans2 = w(j)
```

   v) You are now also given that

$$u(x,0) = 1 - \cos\left(x\frac{2\pi}{L}\right).$$

   At the first time step ($\ell = 0$), write the full equation from (3.2) that corresponds to the first row of A. Clearly indicate known values.

f) For $n = 20$, calculate the sparsity of A.

g) Do you expect the inverse matrix, A^{-1}, to be very sparse ? (Answer yes or no.)

h) The matrix factorizes as $A = B^T B$, where B is an $m \times n$ matrix with $m > n$ and independent columns. Prove that A is positive definite.

i) Given the Cholesky factorization $A = R^T R$, explain how to solve $A\boldsymbol{x} = \boldsymbol{b}$.

j) In preparing to solve for $\boldsymbol{u}^{\ell+1}$, (3.2) can be rearranged to

$$K\boldsymbol{u}^{\ell+1} = \boldsymbol{u}^{\ell}.$$

Suppose that $\Delta t = 0.1$ and $\Delta x = 0.5$. In terms of the matrix A, find the matrix K.

Part B – Statistics

4. Answer in a separate book marked Question 4

a) Grocery stores and large supermarkets all use scanners to calculate a customer's bill. Scanners should be as accurate as possible. An independent agency regularly monitors stores by randomly selecting items and comparing the shelf price with the checkout scanner price. During one check by the agency in a given store, $n = 9$ items were randomly sampled and all were found to be incorrectly scanned. The amounts of overcharge (in cents) were:

`60,  110, 20, 40, -70, -10, 50, 205, 25`

A negative sign indicates an undercharge (the scanner price was below the shelf price). The sample mean is $\bar{x} = 47.78$ cents and the sample standard deviation is $s = 76.90$ cents.

i) Give the five-number summary of the data.

ii) Are there any outliers? Give a reason.

iii) Draw a boxplot.

iv) The store manager claims that these errors are unwilled and not intended to increase the store's profit. Does the data show enough evidence against that claim? Test the hypothesis (at significance level $\alpha = 0.05$) that the 'true' (i.e., over the whole store) mean overcharge μ is not more than 0. (*Write the detail of the test: null and alternative hypotheses, rejection criterion or observed value of the test statistic, p-value, conclusion in plain language - you may use bounds for the p-value.*)

v) What assumptions about the underlying distribution of the data are required for the above test? Explain how you would check for these assumptions.

b) The time taken to complete a certain test by a First Year student is reasonably modelled by a Normal distribution with a mean of 120 minutes and a standard deviation of 10 minutes. For the same test, a Second Year student takes a time reasonably modelled by a Normal distribution with a mean of 110 minutes and a standard deviation of 8 minutes. On a given day, 30 First Year students and 20 Second Year students sit the test.

   i) Find the probability that a randomly selected First Year student will take less than 100 minutes to complete the test.

   ii) Find the probability that a randomly selected Second Year student will take less than 100 minutes to complete the test.

   iii) Find the probability that a student, randomly selected from the 50 students sitting the test on that day, will take less than 100 minutes to complete the test.

   iv) A student has taken less than 100 minutes to complete the test. What is the probability that she is a Second Year student?

c) A random sample of 50 suspension helmets used by motorcycle riders and race-car drivers was subjected to an impact test, and on 18 of these helmets some damage was observed.

   i) Find a 95% two-sided confidence interval on the true proportion π of helmets of this type that would show damage from this test.

   ii) State any assumptions you need to make to determine the above confidence interval, and explain whether they seem reasonable in this situation.

   iii) How many helmets must be tested to be 95% confident that the error in estimating π is less than 0.02, regardless of the true value of π?

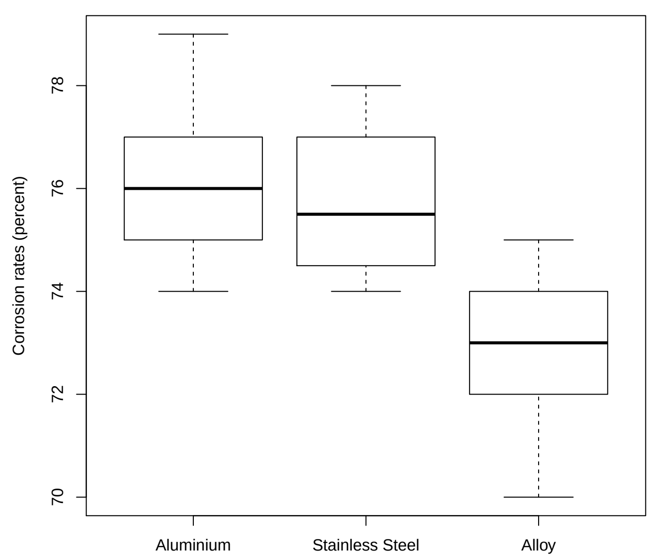
Please see over ...

5. Answer in a separate book marked Question 5

Corrosion rates (percent) were measured for 3 different metals that were immersed in a highly corrosive solution:

Aluminium	Stainless Steel	Alloy
75	74	73
77	76	74
76	75	72
79	78	74
74	74	70
77	77	73
75	75	74
	77	71
		75
$n_1 = 7$	$n_2 = 8$	$n_3 = 9$
$\bar{x}_1 = 76.14$	$\bar{x}_2 = 75.75$	$\bar{x}_3 = 72.89$
$s_1 = 1.68$	$s_2 = 1.49$	$s_3 = 1.62$

Comparative boxplots are given in the figure below.



a) What do the boxplots tell you about the corrosion rates for the different metals?

b) State three assumptions that need to be valid for an Analysis of Variance (ANOVA) to be an appropriate analysis. Comment on the suitability of these assumptions here, where applicable.

Assume from now on that these assumptions are valid.

c) An ANOVA table was partially constructed to summarise the data:

Source	df	SS	MS	F
Treatment	**(1)**	**(2)**	**(3)**	10.39
Error	**(4)**	**(5)**	2.536	
Total	**(6)**	105.96		

Copy the ANOVA table in your answer booklet. Complete the table by determining the missing values (1)–(6) **without using the value of** $F = 10.39$, stating how you computed the missing entries. Confirm the value of $F = 10.39$ and explain how this is obtained from the rest of the table.

d) Using a significance level of $\alpha = 0.05$, carry out the ANOVA F-test to determine whether there is a difference in corrosion rates among the three metals. (*You can use the numerical values found in the above table; however, you are required to write the detail of the test: null and alternative hypotheses, rejection criterion or observed value of the test statistic and p-value, conclusion in plain language - you may use bounds for the p-value.*)

e) From the previous results, construct a 95% two-sided confidence interval on the difference between the 'true' corrosion rates for Aluminium and for Stainless Steel, that is, $\mu_1 - \mu_2$. Would you conclude that there is a significant difference between these two metals? Explain.

f) A two-sample t-test is carried out for comparing the 'true' mean corrosion rates for Aluminum and Alloy. The p-value obtained is 0.0006. Another two-sample t-test is then carried out to compare Stainless Steel and Alloy, and the obtained p-value is 0.0013. Does simultaneously analysing the three pairwise comparisons (these two t-tests and the confidence interval in (e)) allow you to come to the same conclusion as the ANOVA F-test in (d), at overall significance level $\alpha = 0.05$? Explain.

6. Answer in a separate book marked Question 6

The Leaning Tower of Pisa is an architectural wonder. Engineers concerned about the tower's stability have done extensive studies of its increasing tilt. Measurements of the lean of the tower over time provide much useful information.

The following table gives measurements for the years 1975 to 1987. The variable *Lean* represents the difference between where a point on the tower would be if the tower were straight and where it actually is. The data are coded as tenths of a millimeter (mm) in excess of 2.9 meters, so that the 1975 lean, which was 2.9642 meters, appears in the table as 642. Only the **last two digits** of the year were entered into the computer:

Year (x_i)	75	76	77	78	79	80	81	82	83	84	85	86	87
Lean (y_i)	642	644	656	667	673	688	696	698	713	717	725	742	757

The engineers obtained the following summary statistics:

$$\bar{x} = 81 \quad \text{and} \quad s_{xx} = \sum_{i=1}^{13}(x_i - \bar{x})^2 = 182$$

The response variable Y is *Lean* and the predictor variable X is *Year*. The regression model to be fitted is given by:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Use the following regression output to answer the questions below.

```
Regression Analysis: Y versus X

Estimated Coefficients:

                   Coef    SE Coef         T           P
(Intercept)     -61.121      25.13   -2.4322    0.033279
X                9.3187    0.30991    30.069      0.0000

Number of observations: 13, Error degrees of freedom: 11

S: 4.18                     R-squared: 0.988
```

a) i) What is the linear regression equation when fitted to the data above?

ii) What percentage of variation in the response variable *Lean* (Y) is explained by the predictor variable *Year* (X)? Comment on the strength of this linear relationship.

iii) State carefully what the slope (β_1) tells you about the relationship between *Year* and *Lean*.

iv) Give the estimated value of σ, the standard deviation of the error term ϵ.

b) The engineers were interested in whether there is significant evidence that the lean changes each year.

i) Perform a hypothesis test to determine whether the variable X is significant in this model, at the 5% level of significance. *(You can use the numerical values found in the above output, however you are required to write the details of the test: null and alternative hypotheses, rejection criterion or observed value of the test statistics and p-value, conclusion in plain language.)*

ii) Give a 99% two-sided confidence interval for the slope of the population regression line.

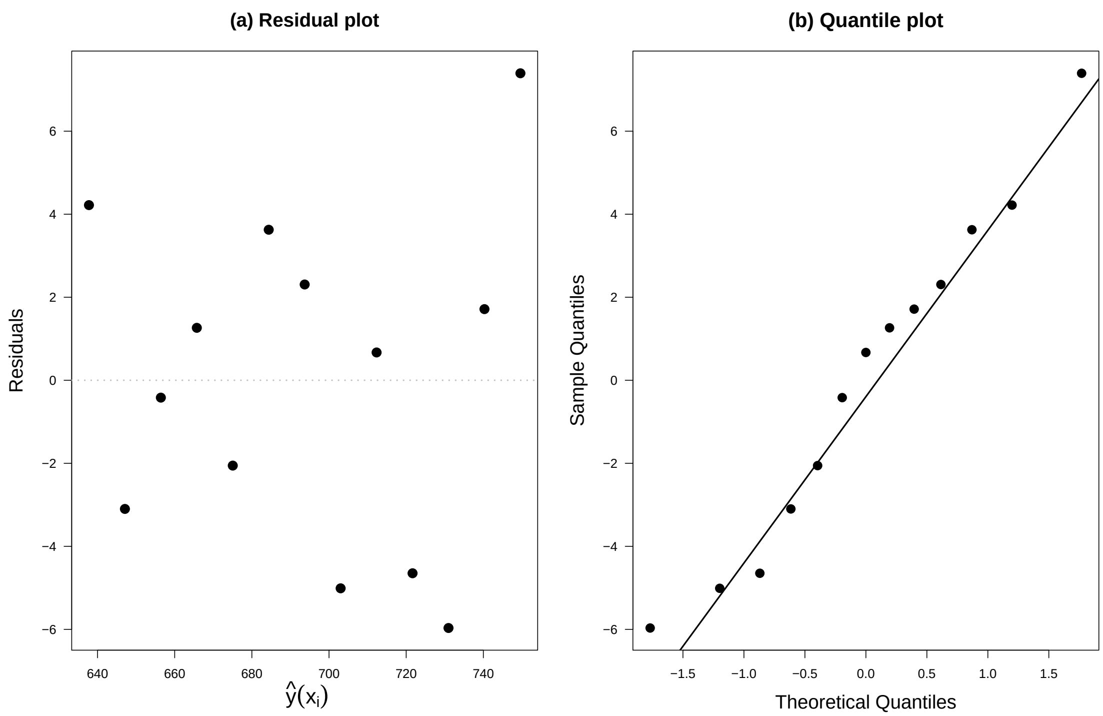
c) The engineers working on the Leaning Tower of Pisa were most interested in how much the tower would lean if no corrective action was taken.

i) Use the linear regression equation in a) i) to predict the tower's lean in the year 1989. Give this in tenths of a mm.

ii) Find an interval which you are 95% confident that the actual lean value for the year 1989 will be contained in.

d) The engineers conducted a residual analysis by plotting: (a) the residual versus fitted values; and (b) a normal quantile plot below.

List three essential assumptions that the error ϵ in the model must satisfy for the above regression analysis to be valid. Comment on their validity using the plots below.


(a) Residual plot
Residuals
ŷ(xᵢ)
(b) Quantile plot
Sample Quantiles
Theoretical Quantiles