Please print in pen:

Waterloo Student ID Number:

WatIAM/Quest Login Userid:

Times: Friday 2019-08-09 at 09:00 to 11:30
Duration: 2 hours 30 minutes (150 minutes)
Exam ID: 4097543
Sections: STAT 331 LEC 001
Instructors: Nathaniel Stevens

UNIVERSITY OF WATERLOO

Examination
Final
Spring 2019
STAT 331

Special Materials

Candidates may bring only the listed aids.

· Calculator - Pink Tie
· Study Notes - Double-Sided 8.5x11

Instructions:

- This test consists of 20 pages including this cover page.
- Page 17 contains information relevant to Question 3.
- Page 18 contains tables of quantiles from the $t_{(10)}$, $t_{(45)}$, $F_{(1,45)}$, $F_{(2,45)}$, $F_{(3,45)}$ and $F_{(4,45)}$ distributions.
- Pages 19 and 20 contain additional space for rough work. If you use these pages for work that you would like to have marked, you must clearly indicate this.
- For your convenience you may remove pages 17-20.
- All numeric answers should be rounded to four decimal places (unless the answer is exact to fewer than four decimal places).
- Incorrect answers may receive partial credit if your work is shown. An incorrect answer with no work shown will receive 0 points.

Question	Points
Q1	/20
Q2	/25
Q3	/32
Q4	/5
Q5	/11
Q6	/7
Total	/100

Signature:

- Please identify yourself by signing here: ____

Please initial:

Question 1 [20 points]

(a) [1] Suppose that you wish to use forward selection to choose among $q = 7$ explanatory variables for inclusion in a model. In the *worst-case scenario* how many separate models would you have to fit when applying this algorithm?
   i. 28
   ii. 29
   iii. 21
   iv. 22

(b) [1] The regression model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$ is a *linear* regression.
   i. True
   ii. False

(c) [1] In the context of a simple linear regression relating y to x, the point $(\bar{x}, \bar{y})$ lies on the line-of-best fit.
   i. True
   ii. False

(d) [1] A 99% prediction interval for y_0 is narrower than the 99% confidence interval for μ_0.
   i. True
   ii. False

(e) [1] In the context of a linear regression model relating to y to x_1, x_2,…, x_p the null hypothesis corresponding to the test of *overall significance* in the linear regression is

$$H_0\colon \beta_0 = \beta_1 = \cdots = \beta_p = 0$$

   i. True
   ii. False

(f) [1] Suppose that the explanatory variables x_1, x_2 and x_3 are together in a model. The variance inflation factor for x_1 is $VIF_1 = 2$. What percentage of the variation of x_1 is explained by x_2 and x_3?
   i. 10%
   ii. 20%
   iii. 50%
   iv. 80%

(g) [1] Suppose that you wish to identify observations that have a large influence on your analysis. Which of the following metrics is best-suited for this purpose?
   i. Cook's D-Statistics
   ii. Leverages
   iii. Studentized Residuals

(h) [1] Suppose that a model with $p = 77$ explanatory variables and $n = 100$ observations is fit and the resulting R^2 value is 0.9. What is R^2_{adj}? Write your answer in the space below.

____

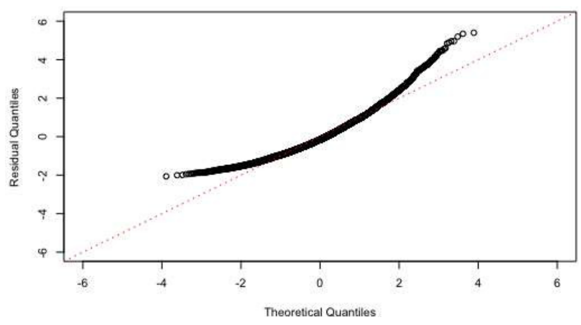
(i) [1] The addition of an explanatory variable to a linear regression model always decreases the error sum of squares.
   i. True
   ii. False

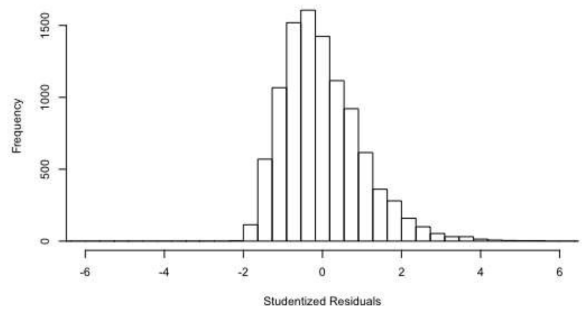
(j) [1] *Leave-one-out* cross validation is equivalent to:
   i. 1-fold cross validation
   ii. n-fold cross validation
   iii. $(n-1)$-fold cross validation

Please initial:

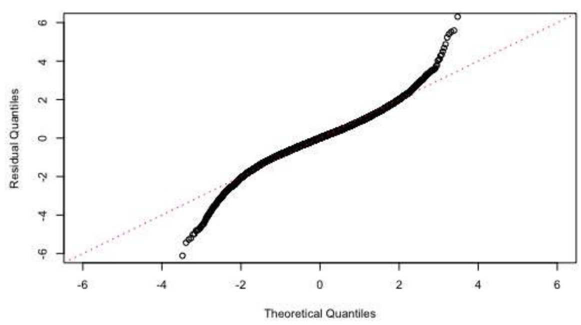
(k) [4] Four datasets gave rise to the following histograms and QQ-plots. Match each QQ-plot with the histogram that is most likely based on the same data. To identify your matching, beside each histogram indicate which QQ-plot – either i., ii., iii., or iv. – matches it.

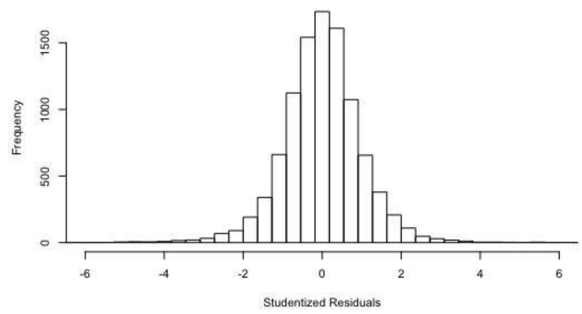
i.



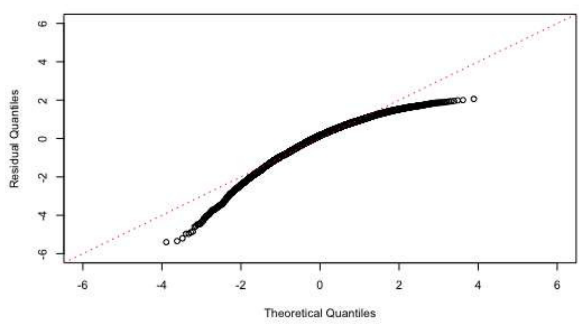


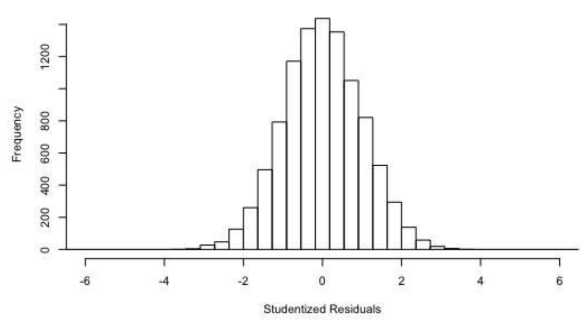
ii.



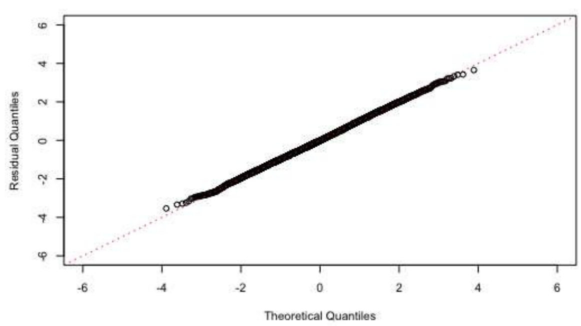


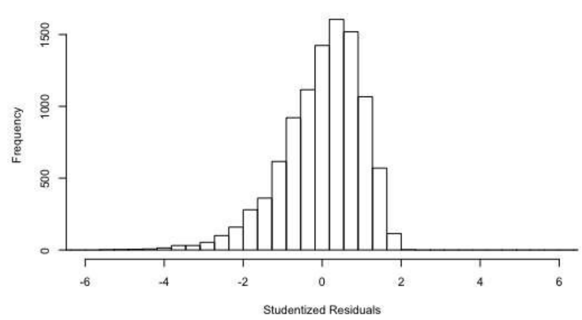
iii.





iv.





(l) [1] Suppose that we fit the following model: $\ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$. Suppose also that the estimate of β_1 is 0.4. By what factor do we expect y to increase if x_1 is increased by 3 units (and x_2 and x_3 are held fixed)? Write your answer in the space below.

____

(m) [1] In *every* linear regression model, the average residual $\bar{e}$ is zero.

i. True
ii. False

The context of questions (n)-(q) is the following model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i$$

for $i = 1,2,\dots,50$.

(n) [1] The null t-distribution associated with a test of $\mathrm{H_0}: \beta_4 = 0$ vs. $\mathrm{H_A}: \beta_4 \neq 0$ has how many degrees of freedom? Write your answer in the space below.

____

(o) [1] The rejection region associated with the test in (n) is given by:

$$\{t | t \geq 2.015 \text{ or } t \leq -2.015\}$$

If $\hat{\beta}_4 = 1.07$ and $\mathrm{SE}[\hat{\beta}_4] = 0.06$, is the hypothesis rejected? In the space below, state YES or NO.

____

(p) [1] The null F-distribution associated with a test of $\mathrm{H_0}: \beta_3 = \beta_4 = \beta_5 = 0$ has how many degrees of freedom? Write your answer in the space below.

____

(q) [1] The least squares estimate of σ^2 is $\hat{\sigma}^2 = 5$ and the sample variance of the response observations is 20. What is the value of R^2? Write your answer in the space below.

____

Question 2 [25 points]

Consider the following regression equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

for $i = 1,2,\dots,13$.

(a) [3] This model may be equivalently written in matrix form as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Define each of $\boldsymbol{y}$, X and $\boldsymbol{\beta}$ in this case (i.e., when $n = 13$ and $p = 2$).

Please initial:

(b) [9] Suppose that the following summaries are available:

$$(X^T X)^{-1} = \begin{bmatrix} 2 & -5 & -3 \\ -5 & 1 & -1 \\ -3 & -1 & 5 \end{bmatrix} \quad \text{and} \quad \widehat{\boldsymbol{\beta}} = \begin{bmatrix} 52 \\ 10 \\ 3 \end{bmatrix} \quad \text{and} \quad SSE = 30$$

i. [1] Write down the equation for the fitted model.

ii. [1] Compute $\hat{\sigma}^2$.

iii. [1] Compute the estimated variance of $\hat{\beta}_1$, $\mathrm{Var}[\hat{\beta}_1]$.

iv. [3] Test the following hypothesis at a 5% level of significance. [N.B. State the value of the test statistic and draw your conclusion by referring to the relevant quantiles found on page 18].

$$\mathrm{H}_0: \beta_1 = 0 \text{ vs. } \mathrm{H}_A: \beta_1 \neq 0$$

v. [3] Calculate a 95% confidence interval for β_2. [N.B. It will be useful for you to refer to the relevant quantiles found on page 18].

Please initial:

(c) [11] Interest lies in doing inference for the response variable for specific values of x_1 and x_2.

i. [3] Consider the row vector $\boldsymbol{x_0} = [1 \quad x_{01} \quad x_{02}]$. Given $\widehat{\boldsymbol{\beta}} \sim \text{MVN}(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1})$, derive the distribution of $\hat{\mu}_0 = \boldsymbol{x_0}\widehat{\boldsymbol{\beta}}$. For full points you must name the distribution and derive general expressions for the expected value and variance of $\hat{\mu}_0 = \boldsymbol{x_0}\widehat{\boldsymbol{\beta}}$.

- Distribution name:
- $\text{E}[\hat{\mu}_0] =$
- $\text{Var}[\hat{\mu}_0] =$

ii. [2] By substituting the estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2)$ and $x_1 = 1$ and $x_2 = 3$ into the equations above, calculate estimates of the $\text{E}[\hat{\mu}_0]$ and $\text{Var}[\hat{\mu}_0]$ when $x_1 = 1$ and $x_2 = 3$ (i.e., when $\boldsymbol{x_0} = [1 \quad 1 \quad 3]$).

Please initial:

iii. [3] Calculate a 90% *confidence* interval for the expected response observation when $x_1 = 1$ and $x_2 = 3$. [N.B. It will be useful for you to refer to the relevant quantiles found on page 18].

iv. [3] Calculate a 90% *prediction* interval for the predicted response observation y_0 when $x_1 = 1$ and $x_2 = 3$. [N.B. It will be useful for you to refer to the relevant quantiles found on page 18].

v. [2] Briefly explain why the widths of these intervals differs.

Question 3 [32 points]
There is a great deal of interest in determining the traits of a high school that predict the success of its students. In the US, student success beyond high school is often linked to high SAT (scholastic aptitude test) scores. A recent study aimed to identify attributes of a high school that may be used to predict the SAT scores of its students. In this study the following information was collected for each high school in a sample of $n = 50$ American high schools:

- `SAT`: average SAT (scholastic aptitude test) score (y)
- `expend`: average expenditures per student (in \$1000s) (x_1)
- `ratio`: average student-teacher ratio (x_2)
- `salary`: average annual salary of teachers (in \$1000s) (x_3)
- `takers`: percentage of eligible students that actually took the SAT test (x_4)

The linear regression model that relates `SAT` to the other variables may be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

Partial R output for this model is shown on page 17. You may refer to this in the questions that follow.

(a) [2] It is usually believed that students thrive in smaller classes.

- [1] Interpret $\hat{\beta}_2$ in the context of the problem.

- [1] Does the value of $\hat{\beta}_2$ suggest that students perform better when class sizes are smaller? Circle one.

YES NO

(b) [6] It is also typically believed that students perform better in schools that pay their teachers higher salaries.

- [1] Interpret $\hat{\beta}_3$ in the context of the problem

- [1] Does the value of $\hat{\beta}_3$ suggest that students perform better when their teachers are paid more money? Circle one.

YES NO

- [4] Using an appropriate hypothesis test, at a 5% level of significance, test whether higher salaries are associated with a *significantly higher* expected SAT score. Draw your conclusions in the context of the problem.
  [N.B. Clearly state the hypothesis you are testing; state the value of the test statistic; draw your conclusion by referring to the relevant quantiles found on page 18].

(c) [1] What percentage of the variability in `SAT` scores is *not* explained by `expend`, `ratio`, `salary` and `takers`?

(d) [3] Complete the following ANOVA table:

Source	*df*	SS	MS	*F*
Regression				
Error				
Total				

SPACE LEFT FOR ROUGH WORK:

Please initial:

(e) [2] The F-ratio from the ANOVA table in part (d) tests the following hypothesis:

$$\text{H}_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \text{ vs. } \text{H}_\text{A}: \beta_j \neq 0 \text{ for some } j = 1,2,3,4$$

At a 1% level of significance, do you reject the null hypothesis? State YES or NO and provide a brief justification.
[N.B. Draw your conclusion by referring to the relevant quantiles found on page 18]

(f) [3] To evaluate the extent of multicollinearity, four separate regressions were fit between the explanatory variables. These models and their R^2 values are summarized in the table below. In the space provided, state the value of the variance inflation factor for each explanatory variable.

Response	Explanatory Variables	R^2	VIF
`expend`	`ratio, salary, takers`	0.894	
`ratio`	`expend, salary, takers`	0.589	
`salary`	`expend, ratio, takers`	0.892	
`takers`	`expend, ratio, salary`	0.430	

Does multicollinearity appear to be a problem for this data? Circle one.

YES NO

SPACE LEFT FOR ROUGH WORK:

Please initial:

(g) [5] Interest lies in finding a subset of the explanatory variables (and hence a reduced model) that adequately accounts for variation in SAT scores. The information provided in Table 1 on page 17 gives the model summary of *all possible regressions*, which in this case corresponds to $2^4 = 16$ different models. Note: all models contain an intercept. Using the information in this table, answer the following questions.

i. [2] Perform *backward elimination* using the AIC as a basis for eliminating variables from the model. Specifically, indicate the order in which variables exit the model and state the final model.

ii. [2] Perform *forward selection* using the AIC as a basis for adding variables into the model. Specifically, indicate the order in which variables enter the model and state the final model.

iii. [1] The best overall model among *all possible regressions* is the one with the smallest AIC. Did either of these stepwise selection techniques choose the best overall model? Circle one.

YES NO

(h) [4] The optimal model according to *all possible regressions* and the AIC metric is the reduced model

$$y = \beta_0 + \beta_1 x_1 + \beta_4 x_4 + \varepsilon$$

This model arises if the following null hypothesis is true:

$$\mathrm{H}_0: \beta_2 = \beta_3 = 0 \tag{1}$$

i. [1] This null hypothesis may be equivalently stated as

$$\mathrm{H}_0: A\boldsymbol{\beta} = \mathbf{0} \tag{2}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^T$. Define the matrix A that makes hypotheses (1) and (2) equivalent.

ii. [3] Using the additional sum of squares principle, formally test this hypothesis at a 5% level of significance.
[N.B. State the value of the test statistic and draw your conclusion by referring to the relevant quantiles found on page 18].

(i) [6] The six plots in Figure 1 (on page 17) were constructed using the reduced model from part (h). For each of the questions below answer YES or NO and provide a one-sentence justification. In each case, be sure to indicate which plot(s) you used to make your decision.

i. Does the independence assumption appear to be met?

ii. Does the constant variance assumption appear to be met?

iii. Does the normality assumption appear to be met?

iv. Do there appear to be any outliers in the y-dimension?

v. Do there appear to be any outliers in the x-dimension?

vi. Do there appear to be any highly influential observations?

Question 4 [5 points]
In class we showed that when $\mathrm{SD}[y]$ is proportional to μ (where $\mu = \mathrm{E}[y]$) that the log-transformation is a variance stabilizing transformation. We also showed that when $\mathrm{Var}[y]$ is proportional to μ that the square-root transformation is variance stabilizing. By the same methods, find the transformation $g(\mu)$ that stabilizes the variance when $\mathrm{SD}[y] \propto \mu^2$.

Please initial:

Question 5 [11 points]
The *Gauss-Markov Theorem* states that in the linear regression model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

the least squares estimator of $\boldsymbol{\beta}$ is the *best linear unbiased estimator* (BLUE) because among all possible linear and unbiased estimators of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$ has the smallest variance. In this question, you will prove this fact by considering another linear transformation $\widehat{\boldsymbol{\beta}}^* = M\boldsymbol{y}$ where M is any $(p+1) \times n$ matrix of fixed numbers. The matrix $D = M - (X^T X)^{-1} X^T$ summarizes the difference between the two transformations.

(a) [5] Show that if $DX = O$ (where O is the $(p+1) \times (p+1)$ matrix of zeros), then

$$\mathrm{E}\left[\widehat{\boldsymbol{\beta}}^*\right] = \boldsymbol{\beta}$$

(b) [5] Show that if $\widehat{\boldsymbol{\beta}}^*$ is unbiased, then

$$\mathrm{Var}\left[\widehat{\boldsymbol{\beta}}^*\right] = \mathrm{Var}\left[\widehat{\boldsymbol{\beta}}\right] + \sigma^2 D D^T$$

(c) [1] Why does the result in part (b) imply that $\mathrm{Var}\left[\widehat{\boldsymbol{\beta}}\right] \leq \mathrm{Var}\left[\widehat{\boldsymbol{\beta}}^*\right]$?

Question 6 [7 points]

(a) [5] Suppose that we wish to model the relationship between a response variable y and p explanatory variables: $x_1, x_2, \dots, x_p$ and we observe the data in two batches: we observe n_1 observations initially, and then another n_2 observations at a later point in time.
For time point $k = 1,2$, let

- $\boldsymbol{y}_k$ be the $n_k \times 1$ response vector containing the response observations
- X_k be the $n_k \times (p+1)$ matrix containing the explanatory variable observations
- $\widehat{\boldsymbol{\beta}}_k = (X_k^T X_k)^{-1} X_k^T \boldsymbol{y}_k$ be the least squares estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$

Show that the least squares estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ based on all of the data can be written as

$$\widehat{\boldsymbol{\beta}} = (X_1^T X_1 + X_2^T X_2)^{-1} \left(X_1^T X_1 \widehat{\boldsymbol{\beta}}_1 + X_2^T X_2 \widehat{\boldsymbol{\beta}}_2\right)$$

Hint: the full X matrix and $\boldsymbol{y}$ vector (based on all $n = n_1 + n_2$ observations) can be obtained by stacking the respective time-specific matrices and vectors on top of each other.

(b) [2] Now suppose that you've collected data in $K > 2$ batches, and with each batch you estimate $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ by $\widehat{\boldsymbol{\beta}}_k = (X_k^T X_k)^{-1} X_k^T \boldsymbol{y}_k$, $k = 1, 2, \dots, K$. Write an expression for the least squares estimate of $\boldsymbol{\beta}$ based on all $n = n_1 + n_2 + \cdots + n_K$ observations and in terms of $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \dots, \widehat{\boldsymbol{\beta}}_K$.

Please initial:

This page is left blank.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1045.9715    52.8698   ?????   ?????
expend         4.4626    10.5465   ?????   ?????
ratio         -3.6242     3.2154   ?????   ?????
salary         1.6379     0.2387   ?????   ?????
takers        -2.9045     0.2313   ?????   ?????
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.7 on ?? degrees of freedom
Multiple R-squared:  0.8246,    Adjusted R-squared:  0.809
F-statistic: ????? on ?? and ?? DF,  p-value: ?????
```

Partial R Output for the full model in Question 3

Variables in Model	AIC	SSE
None (intercept only)	576.39	274307.68
x_1	570.57	234585.62
x_2	578.06	272496.65
x_3	567.64	221229.86
x_4	501.07	58433.15
x_1 x_2	572.33	233442.95
x_1 x_3	569.64	221225.01
x_1 x_4	494.80	49520.06
x_2 x_3	569.23	219441.23
x_2 x_4	500.14	55096.98
x_3 x_4	498.51	53338.38
x_1 x_2 x_3	570.63	216811.94
x_1 x_2 x_4	495.89	48627.32
x_1 x_3 x_4	496.76	49482.54
x_2 x_3 x_4	495.57	48315.37
x_1 x_2 x_3 x_4	497.37	48123.90

Table 1: AIC and SSE for various models in Question 3(g)

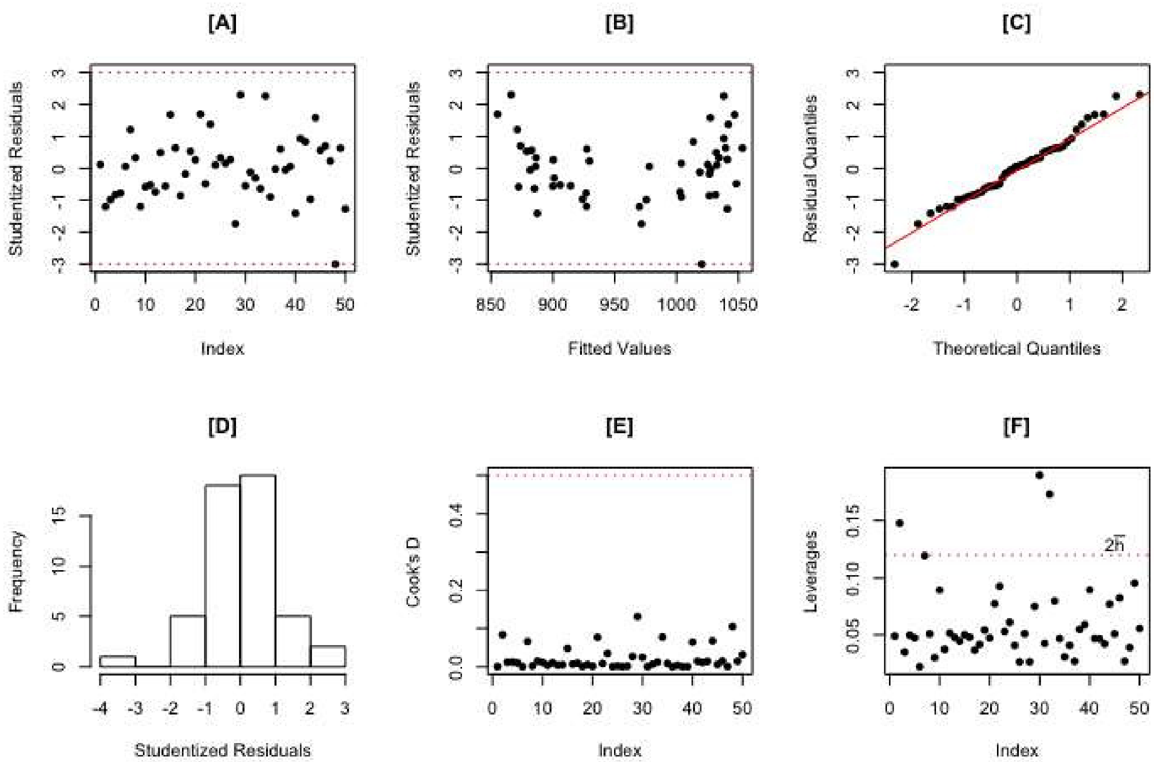


Figure 1: Diagnostic plots for the reduced model in Question 3(h)

For the indicated value of p, the following tables provide x^* where $P(X \geq x^*) = p$

$\boldsymbol{X \sim t_{(10)}}$

p	x^*
0.005	3.1693
0.01	2.7638
0.025	2.2281
0.05	1.8125
0.1	1.3722

$\boldsymbol{X \sim t_{(45)}}$

p	x^*
0.005	2.6896
0.01	2.4121
0.025	2.0141
0.05	1.6794
0.1	1.3006

$\boldsymbol{X \sim F_{(1,45)}}$

p	x^*
0.005	8.7148
0.01	7.2339
0.025	5.3773
0.05	4.0566
0.1	2.8205

$\boldsymbol{X \sim F_{(2,45)}}$

p	x^*
0.005	5.9741
0.01	5.1103
0.025	4.0085
0.05	3.2043
0.1	2.4245

$\boldsymbol{X \sim F_{(3,45)}}$

p	x^*
0.005	4.8918
0.01	4.2492
0.025	3.4224
0.05	2.8115
0.1	2.2097

$\boldsymbol{X \sim F_{(4,45)}}$

p	x^*
0.005	4.2941
0.01	3.7674
0.025	3.0860
0.05	2.5787
0.1	2.0742

LEFT BLANK FOR ROUGH WORK

LEFT BLANK FOR ROUGH WORK

Please initial: