THE UNIVERSITY OF MELBOURNE

Student Number	

Semester 1 Assessment, 2020

School of Mathematics and Statistics

MAST30025 Linear Statistical Models

This exam consists of 24 pages (including this page)

Authorised materials: printed one-sided copy of the Exam or the Masked Exam made available earlier (or an offline electronic PDF reader), any amount of handwritten material, a Casio FX82 calculator, and blank A4 paper.

Instructions to Students

- During exam writing time you may only interact with the device running the Zoom session with supervisor permission. The screen of any other device must be visible in Zoom from the start of the session.
- If you have a printer, print out the exam single-sided and hand write your solutions into the answer spaces.
- If you do not have a printer, or if your printer fails on the day of the exam,

  (a) download the exam paper to a second device (not running Zoom), disconnect it from the internet as soon as the paper is downloaded and read the paper on the second device;

  (b) write your answers on the Masked Exam PDF if you were able to print it single-sided before the exam day.

  If you do not have the Masked Exam PDF, write single-sided on blank sheets of paper.
- If you are unable to answer the whole question in the answer space provided then you can append additional handwritten solutions to the end of your exam submission. If you do this you MUST make a note in the correct answer space or page for the question, warning the marker that you have appended additional remarks at the end.
- Assemble all the exam pages (or template pages) in correct page number order and the correct way up, and add any extra pages with additional working at the end.
- Scan your exam submission to a single PDF file with a mobile phone or a scanner. Scan from directly above to avoid any excessive keystone effect. Check that all pages are clearly readable and cropped to the A4 borders of the original page. Poorly scanned submissions may be impossible to mark.
- Upload the PDF file via the Canvas Assignments menu and submit the PDF to the GradeScope tool by first selecting your PDF file and then clicking on Upload PDF.
- Confirm with your Zoom supervisor that you have GradeScope confirmation of submission before leaving Zoom supervision.
- You should attempt all questions.
- There are 7 questions with marks as shown. The total number of marks available is 90.

Question 1 (10 marks)

(a) [**2 marks**] Give an example of a 3×3 idempotent matrix which is not 0 or I_3.

(b) [**2 marks**] Show that the matrix $A = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$ is positive definite.

(c) [**3 marks**] Show directly that $\frac{\partial}{\partial \mathbf{y}} \mathbf{y}^T A \mathbf{y} = A\mathbf{y} + A^T\mathbf{y}$.

(d) **[3 marks]** Show directly that for any $n \times k$ matrix A with $n \geq k$, the matrix $I - A(A^T A)^c A^T$ has a rank of $n - r(A)$. (You may assume that this matrix is idempotent.)

Question 2 (14 marks)

(a) **[5 marks]** Let $X_1 \sim \chi^2_{k_1,\lambda_1}$ and $X_2 \sim \chi^2_{k_2,\lambda_2}$ be independent. Show directly that $X_1 + X_2 \sim \chi^2_{k_1+k_2,\lambda_1+\lambda_2}$.

(b) **[3 marks]** Let

$$\mathbf{y} \sim MVN\left(\begin{bmatrix} -3 \\ 8 \end{bmatrix}, \begin{bmatrix} 5 & -3 \\ -3 & 2 \end{bmatrix}\right), \qquad A = \begin{bmatrix} 0 & -6 \\ 6 & 7 \end{bmatrix}.$$

Calculate $E[\mathbf{y}^T A\mathbf{y}]$.

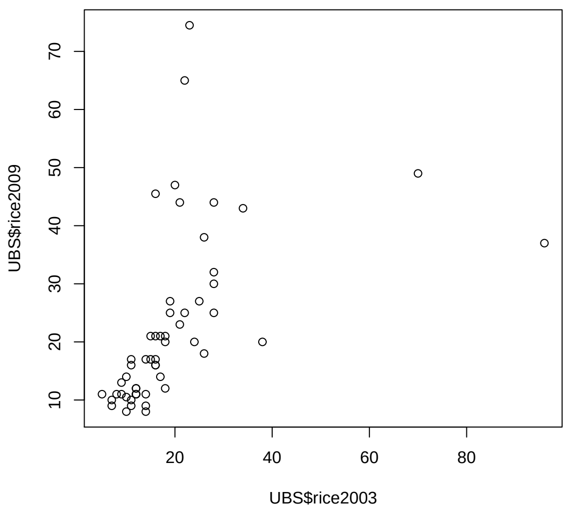
(c) **[3 marks]** Describe the distribution of $3y_1 - 2y_2$.

(d) **[3 marks]** Find all values of a and b for which $ay_1 + by_2$ is independent of $3y_1 - 2y_2$.

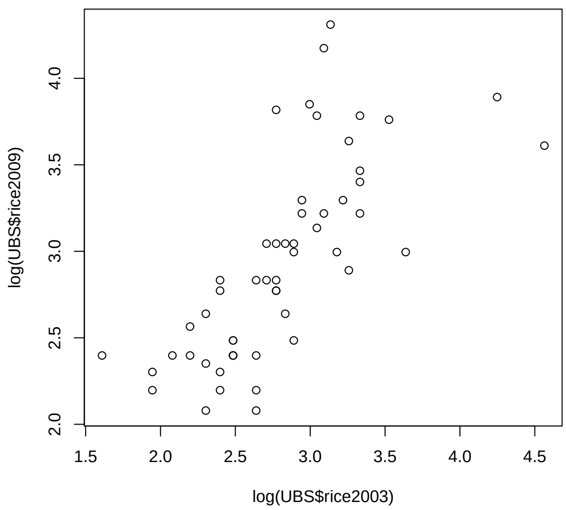
Question 3 (18 marks)

The international bank UBS produced a report on prices and earnings in major cities throughout the world. One of the variables that they measured was the price of 1kg of rice, measured in minutes of labour required for a "typical" worker to purchase the rice. This was measured in 2003 (`rice2003`) and again in 2009 (`rice2009`).
We wish to model the 2009 price in terms of the 2003 price, using the linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The following R calculations are performed:

```
> UBS <- read.csv('UBSprices.csv', header=T)
> plot(UBS$rice2003, UBS$rice2009)
```




```
> plot(log(UBS$rice2003), log(UBS$rice2009))
```



```
> (n <- length(UBS$rice2009))

[1] 54

> X <- cbind(1, log(UBS$rice2003))
> y <- log(UBS$rice2009)
> t(X)%*%X

         [,1]     [,2]
[1,]  54.0000 151.5818
[2,] 151.5818 440.4496

> t(X)%*%y

         [,1]
[1,] 158.3701
[2,] 456.1961

> t(y)%*%y

         [,1]
[1,] 481.9005

> sum(y)

[1] 158.3701

> qt(0.975,50:55)

[1] 2.008559 2.007584 2.006647 2.005746 2.004879 2.004045

> qf(0.95,1,50:55)

[1] 4.034310 4.030393 4.026631 4.023017 4.019541 4.016195

> qf(0.95,2,50:55)

[1] 3.182610 3.178799 3.175141 3.171626 3.168246 3.164993
```

(*Hint: To alleviate rounding error, keep as many digits in internal calculations as possible.*)

(a) **[2 marks]** A logarithmic transformation has been applied to both variables. Give two reasons to justify this transformation.

(b) **[3 marks]** Calculate the least squares estimates of $\boldsymbol{\beta}$.

(c) **[3 marks]** Calculate the sample variance s^2.

(d) **[4 marks]** In 2003, it cost 50 minutes of labour to buy 1kg of rice in the Republic of Linearmodelstan. Calculate (with 95% probability) an interval for the 2009 price of rice (in minutes of labour) in Linearmodelstan.

(e) **[3 marks]** Test for model relevance at the 5% significance level, using a corrected sum of squares.

(f) **[3 marks]** It is claimed that, on average, the price of rice in 2003 is the same as the price of rice in 2009, in terms of labour. This corresponds to a parameter estimate of $\boldsymbol{\beta} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Determine if this point lies within the joint 95% confidence region for the parameters.

Question 4 (12 marks)

Consider the full rank linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with p parameters. Now suppose that we transform the design variables x in a linear manner:

$$z_i = \sum_{j=1}^{p} a_{ji} x_j, \qquad i = 1, \ldots, p.$$

(Note that the x variables include the intercept term.) Now consider the linear model $\mathbf{y} = Z\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2$, which also has p parameters.

(a) **[2 marks]** Express the design matrix Z in terms of X, and state a condition under which the second linear model is also full rank.

(b) **[3 marks]** Calculate the least squares estimators for $\boldsymbol{\beta}_2$ from the second model, and express them in terms of $\mathbf{b}$, the least squares estimators for $\boldsymbol{\beta}$.

(c) **[2 marks]** Consider a subject with design variables $\mathbf{x}^*$ (for the first model). Calculate a point estimate for the average response for this subject, using the second model, and express it in terms of $\mathbf{b}$.

(d) **[3 marks]** Calculate the sample variance for the second model, and express it in terms of the sample variance for the first model.

(e) **[2 marks]** Briefly discuss the implications of the results you have derived above in the context of fitting a linear model, with particular reference to variable standardisation.

Question 5 (12 marks)

Consider the general linear model, $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. This model may be of full or less than full rank.

(a) **[2 marks]** Define the term BLUE (best linear unbiased estimator), and give an example of when one might choose not to use the BLUE.

(b) **[2 marks]** Describe how the parameters $\boldsymbol{\beta}$ of a linear model may be estimated by the method of maximum likelihood, and relate this to least squares estimation.

(c) **[2 marks]** Define the Cook's distance and explain its purpose.

(d) **[2 marks]** Define estimability, and explain its significance for a linear model.

(e) **[2 marks]** Define interaction between a categorical and a continuous predictor, and explain how to model it.

(f) **[2 marks]** Define single and double blinding, and describe their use in experimental design.

Question 6 (16 marks)

Data on 220 agricultural land sales in Minnesota over the period 2002–2011 were collected. The dataset contains the following variables:

- `id`: ID
- `acrePrice`: Sale price, in thousands of dollars per acre
- `region`: One of six major agricultural regions in Minnesota
- `improvements`: Percentage of property value in buildings
- `year`: Year of sale
- `acres`: Size of property
- `tillable`: Percentage of tillable area of the land
- `financing`: Type of financing (title transfer or seller financed)
- `crpPct`: Percentage of land in the US Conservation Reserve Program
- `productivity`: A score measuring the productivity of the land

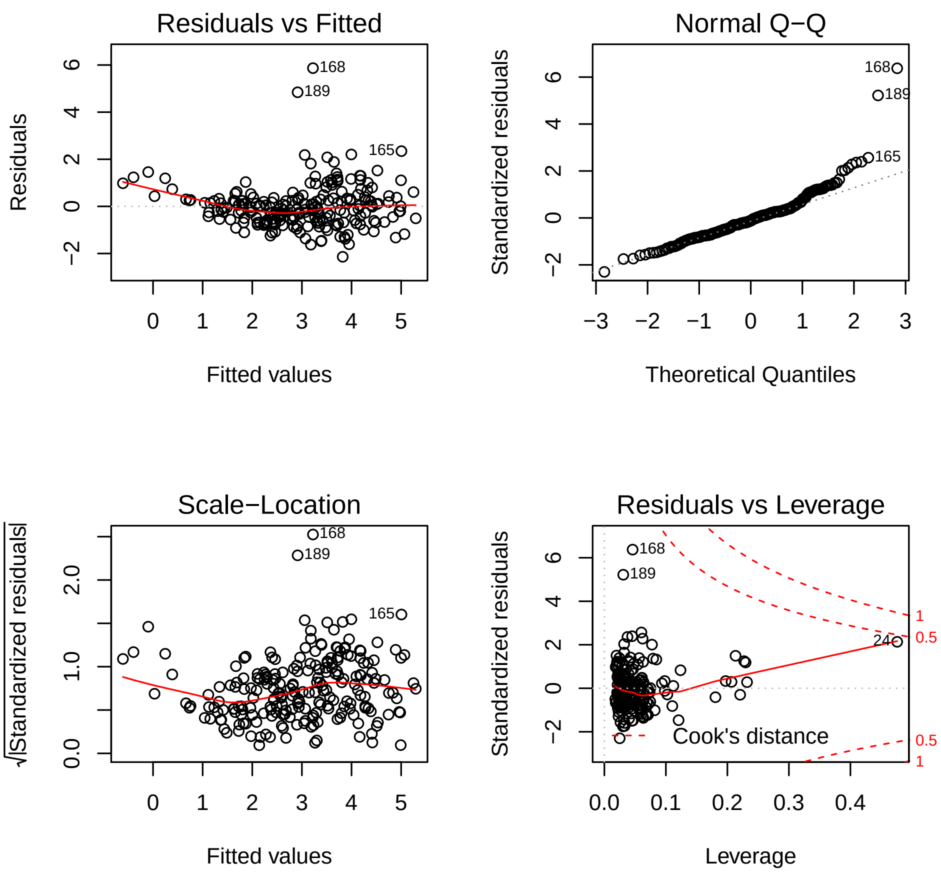
We wish to model the selling price (`acrePrice`) in terms of the other variables (except `id`). The following R calculations are produced:

```
> ML <- read.csv('ML2.csv', header=T)
> interaction_model <- lm(acrePrice ~ (. - id)^2, data=ML)
> additive_model <- lm(acrePrice ~ . - id, data=ML)
> anova(additive_model, interaction_model)

Analysis of Variance Table

Model 1: acrePrice ~ (id + region + improvements + year + acres + tillable +
    financing + crpPct + productivity) - id
Model 2: acrePrice ~ ((id + region + improvements + year + acres + tillable +
    financing + crpPct + productivity) - id)^2
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    207 182.99
2    153 125.15 54    57.845 1.3096 0.1034
```

```
> selected_model <- step(additive_model)

Start:  AIC=-14.52
acrePrice ~ (id + region + improvements + year + acres + tillable +
    financing + crpPct + productivity) - id

               Df Sum of Sq    RSS     AIC
- financing     1     1.135 184.13 -15.159
- improvements  1     1.431 184.42 -14.806
- acres         1     1.582 184.58 -14.626
<none>                      182.99 -14.519
- productivity  1     4.189 187.18 -11.540
- crpPct        1     5.001 187.99 -10.588
- tillable      1     6.770 189.76  -8.527
- region        5    64.123 247.12  41.571
- year          1   140.960 323.95 109.134

Step:  AIC=-15.16
acrePrice ~ region + improvements + year + acres + tillable +
    crpPct + productivity

               Df Sum of Sq    RSS     AIC
- improvements  1     1.509 185.64 -15.363
- acres         1     1.596 185.72 -15.260
<none>                      184.13 -15.159
- productivity  1     4.168 188.30 -12.235
- crpPct        1     5.079 189.21 -11.173
- tillable      1     6.439 190.57  -9.596
- region        5    64.875 249.00  41.245
- year          1   140.494 324.62 107.588

Step:  AIC=-15.36
acrePrice ~ region + year + acres + tillable + crpPct + productivity

               Df Sum of Sq    RSS     AIC
<none>                      185.64 -15.363
- acres         1     1.737 187.37 -15.314
- crpPct        1     4.353 189.99 -12.264
- productivity  1     4.666 190.30 -11.902
- tillable      1     5.163 190.80 -11.328
- region        5    63.368 249.01  39.247
- year          1   143.335 328.97 108.516
```

```
> summary(selected_model)

Call:
lm(formula = acrePrice ~ region + year + acres + tillable + crpPct +
    productivity, data = ML)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1397 -0.5763 -0.1042  0.3114  5.8682

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -6.763e+02  5.338e+01 -12.670  < 2e-16 ***
regionNorthwest     -1.915e+00  2.879e-01  -6.654 2.46e-10 ***
regionSouth Central  1.191e-03  2.376e-01   0.005   0.9960
regionSouth East     5.592e-02  2.887e-01   0.194   0.8466
regionSouth West    -5.216e-01  2.236e-01  -2.332   0.0206 *
regionWest Central  -1.064e+00  2.332e-01  -4.565 8.53e-06 ***
year                 3.379e-01  2.660e-02  12.703  < 2e-16 ***
acres               -7.921e-04  5.664e-04  -1.398   0.1635
tillable             1.109e-02  4.599e-03   2.411   0.0168 *
crpPct              -9.941e-03  4.490e-03  -2.214   0.0279 *
productivity         1.319e-02  5.753e-03   2.292   0.0229 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9425 on 209 degrees of freedom
Multiple R-squared:  0.6094,       Adjusted R-squared:  0.5907
F-statistic: 32.61 on 10 and 209 DF,  p-value: < 2.2e-16

> qt(0.975,209:214)

[1] 1.971379 1.971325 1.971271 1.971217 1.971164 1.971111

> qf(0.95,5,209:214)

[1] 2.257274 2.257066 2.256860 2.256657 2.256455 2.256255

> qf(0.95,6,209:214)

[1] 2.142153 2.141943 2.141736 2.141530 2.141327 2.141125
```

```
> par(mfrow=c(2,2))
> plot(selected_model)
```



(a) [**2 marks**] Interpret the output of the `anova` function.

(b) [**2 marks**] Identify the variable selection procedure that has been used here.

(c) [**3 marks**] From the model `selected_model`, test for the relevance of the `region` variable, at the 5% level. Clearly state your F-statistic and critical value, and interpret your results in the context of the study.

(d) [**2 marks**] Perform one step of backwards elimination on `selected_model`.

(e) [**4 marks**] From the diagnostic plots, comment on the suitability of the linear model for this data. If the model is not suitable, suggest how it can be improved.

(f) **[3 marks]** For a plot of agricultural land in the South West, calculate a 95% confidence interval for the effect of the year on the price per acre in the model `selected_model`.

Question 7 (8 marks)

(a) **[5 marks]** You wish to perform a study to determine if 3 treatments each produce no effect, using a completely randomised design. To do this, you will test the hypothesis $H_0 : \mu + \tau_1 = \tau_1 - \tau_2 = \tau_2 - \tau_3 = 0$. You are given resources to study 50 sample units. Determine the optimal allocation of the number of units to assign to each treatment. (*Hint: In a completely randomised design with treatment effects* τ_i, *we have var* $(\mu+\tau_i) = \sigma^2 \frac{1}{n_i}$ *and var* $(\tau_i - \tau_j) = \sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_j}\right)$. *To minimise a function* $f(\mathbf{x})$ *under the constraint* $g(\mathbf{x}) = c$, *minimise* $f(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda(g(\mathbf{x}) - c)$.)

(b) [**3 marks**] Compare and contrast blocking and randomisation as tools for eliminating confounding, and discuss their best use in experimental design.

End of Exam—Total Available Marks = 90