STAT3008: Applied Regression Analysis
2019/20 Term 2
Final Examination

Date: 8th May 2020 (Friday)

Time: 9:30am – 12:30pm (180 minutes, including the time to upload your work to Blackboard)

Total Score: 100 points

- Please reserve at least 15 minutes of the exam time to save, convert and upload the file. Late submission will NOT be graded.
- Please present your answers in 4 significant figures.
- Submission Requirement: (1) **Name and SID on the 1st page** of your work,
  (2) Only a **single file in .pdf or .doc* format (size < 15MB)** will be accepted
  (3) **Filename** in the format of "**LAST NAME First Name – SID.pdf/doc***"
- **How to submit your exam work?** A dropbox button is now available on Blackboard.

Problem 1 [1 point]: Let n_o **= Number of characters in your last name (in English)**, and s_9 **= The 9th digit (i.e. second last digit) of your SID**. What are your values of (n_o, s_9)?

*(E.g $(n_o, s_9) = (4, 9)$ for Mr. **CHAN** with SID 1234567890, $(n_o, s_9) = (5, 0)$ for Miss **ZHANG** with SID 1234567801)*

***Note**: Both n_o and s_9 will be used in later problems.*

Problem 2 [18 points]: Consider a 2×n matrix $\mathbf{W} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ w_1 & w_2 & \cdots & w_n \end{pmatrix}$, where n>2.

For i, j =1, 2, …, n, let q_{ij} be the $(i, j)^{\text{th}}$ element of the matrix $\mathbf{Q} = \mathbf{W}'(\mathbf{W}\mathbf{W}')^{-1}\mathbf{W}$.

(a) [10 points] Show that $q_{ij} \geq \frac{1}{n} - \frac{1}{2}$ for $i < j$.

(b) [8 points] Suppose n = 10. Based on the result in part (a), and the values of n_o and s_9 from Problem 1, provide the possible values of $\{w_1, w_2, \ldots, w_{10}\}$ such that

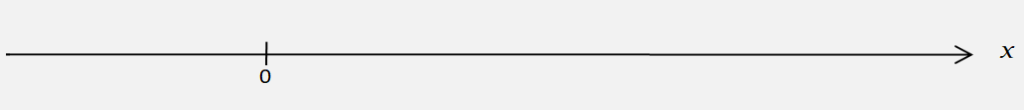
$$q_{23} = -0.40 \quad \text{and} \quad \sum_{i=1}^{10} w_i = (n_0 + 2)(s_9 + 3) + 7n_0 .$$

Problem 3 [14 points]: Consider simple linear regression $y_i = \beta_0 + \beta_1 x_i + e_i$, i = 1, 2, 3 and 4.

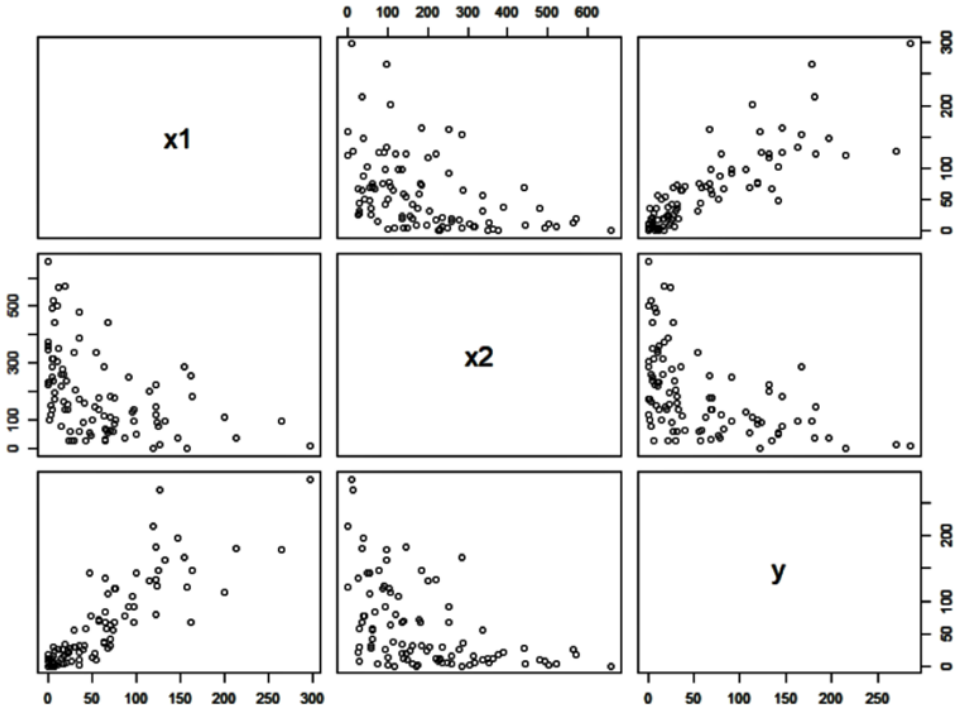
Suppose the hat matrix is given by $\mathbf{H} = \begin{pmatrix} ?? & 0.2500 & ?? & ?? \\ ?? & 0.2500 & ?? & 0.2500 \\ 0.4167 & ?? & ?? & ?? \\ ?? & ?? & ?? & 0.9166 \end{pmatrix}$.

(a) [8 points] Fill in ALL the missing values from the hat matrix above.

(b) [6 points] Given that $x_4 = 9.50 + (n_0/100)$ and $\bar{x} = (x_1 + x_2 + x_3 + x_4)/4 = 2.50(1 + (s_9 + 3)n_0/100)$, where n_o and s_9 are the values you have in Problem 1. On the real line (like the one below with 0 included), put down the possible locations of x_1, x_2, x_3, x_4 and $\bar{x}$.



Problem 4 [22 points]: Suppose we would like to explain the response variable Y by two explanatory variables x_1 and x_2 . Below is the scatterplot matrix for the 3 variables, together with their selected summary statistics:



(Sample) Quartiles of the 3 Variables

	Q_0(Min)	Q_1	Q_2	Q_3	Q_4(Max)
x1	0.001330	12.30	41.85	82.32	298.2
x2	1.443	77.40	150.4	259.2	656.4
y	0.009014	10.49	29.88	90.62	286.4

Sample Variance-Covariance Matrix

	x1	x2	y
x1	3586.7	-3759.2	3293.9
x2	-3759	21222	-4599.5
y	3293.9	-4599.5	4130.5

The data was fitted by the following multiple linear regression, and below are the R codes and output of the OLS estimates: **(Model #1)**: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$

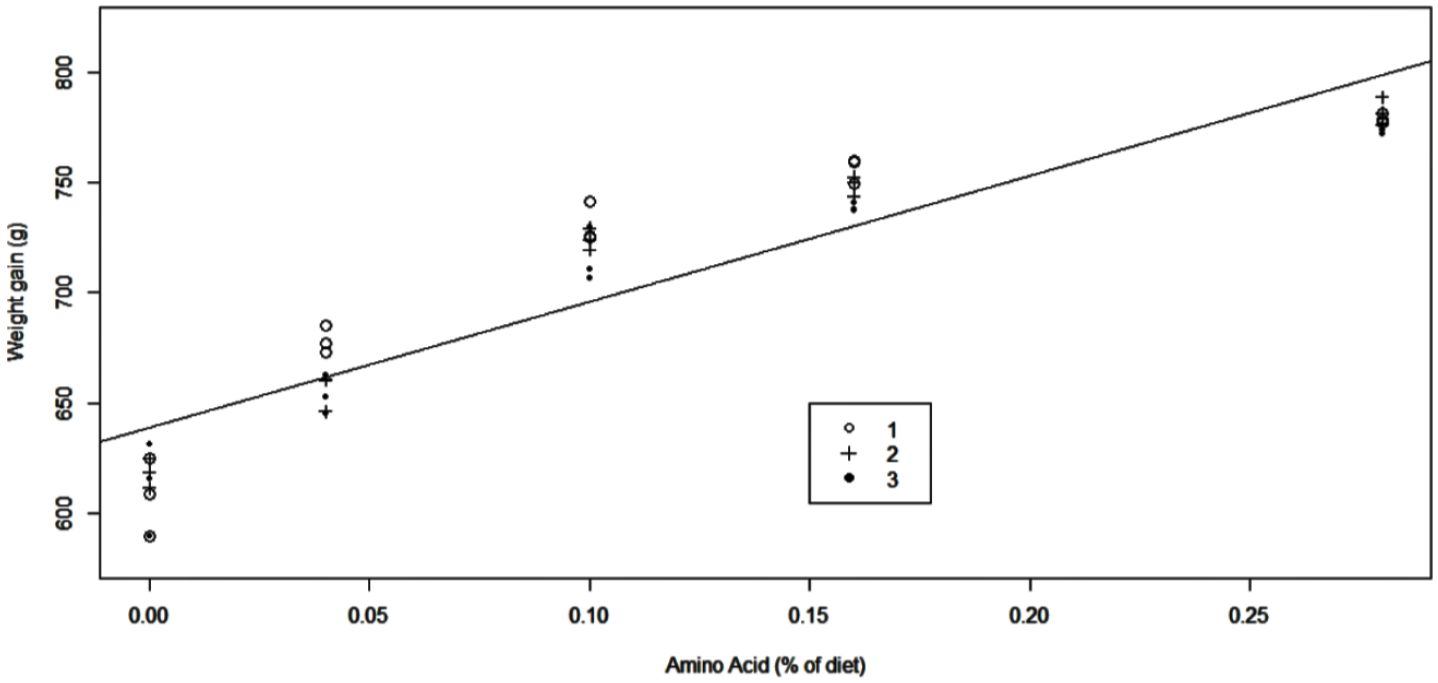
```
> fit0<-lm(y~x1+x2); summary(fit0)
Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept) 20.16407    7.66704      2.63    0.00993 **
x1           0.86133    0.05940     14.50    < 2e-16 ***
x2          -0.06404    0.02482     -2.58    0.01137 *
Residual standard error: 32.07 on 97 degrees of freedom
Multiple R-squared:  0.7614,    Adjusted R-squared:  0.7565
F-statistic: 154.8 on 2 and 97 DF,  p-value: < 2.2e-16
```

(a) [8 points] Describe in details on whether Model #1 is appropriate in fitting the data.

(b) [14 points] Based on the material from Section 5.1 to 8.3, propose TWO possible models you think is best to fit the data. You should specify your models as much as you could, and explain on why the models you proposed are reasonable.

Problem 5 [14 points]: The owner of a turkey (火雞) farm would like to investigate how the weight gain of a turkey is affected by the (i) type and (ii) amount of amino acids in their diet. n = 45 turkeys, which were selected at random from the farm, were divided into 15 dens of 3 turkeys each. Turkeys from different dens were then fed by different amount (0%, 4%, 10%, 16% and 28% of their diet) and types of amino acids (Type 1, 2, 3 labeled as ○, + and • in the scatterplot) as in the table below:

Den	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Turkey Number (i)	1 to 3	4 to 6	7 to 9	10 to 12	13 to 15	16 to 18	19 to 21	22 to 24	25 to 27	28 to 30	31 to 33	34 to 36	37 to 39	40 to 42	43 to 45
Amount of Amino Acid (% of diet)	0%	4%	10%	16%	28%	0%	4%	10%	16%	28%	0%	4%	10%	16%	28%
Type of Amino Acid	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3



Let x = Amount of amino acid in the diet, and
y = Weight gain (in gram) of turkey during the experiment.
The following model was fitted into the data, and the fitted line is shown in the scatterplot.

(Model #1) $y_i = \beta_0 + \beta_1 x_i + e_i, \quad E(e_i) = 0, \mathrm{Var}(e_i) = \sigma^2, \quad i = 1,2,\cdots,45$

(a) [3 points] Do you think the null plot assumptions hold for the residual plot of Model #1? Explain.

[part (b) to (c)] Suppose we would like to construct a more complicated model to explain the behavior of the data, but limiting the model to have a maximum of 12 parameters only. Based on the material covered in Section 5.1 to 8.3,

(b) [10 points] Provide the specification of a model you think is best to explain the data, and outline your idea on how to come up with the model. The specification should be a single formula applicable for i = 1, 2, …, 45 observations similar to Model #1 above.

(c) [1 point] How many parameter are there in your model in part (b)?

Problem 6 [13 points]: Consider a multiple linear regression with 4 terms:

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

The table below shows the AIC for models with different subsets of terms.

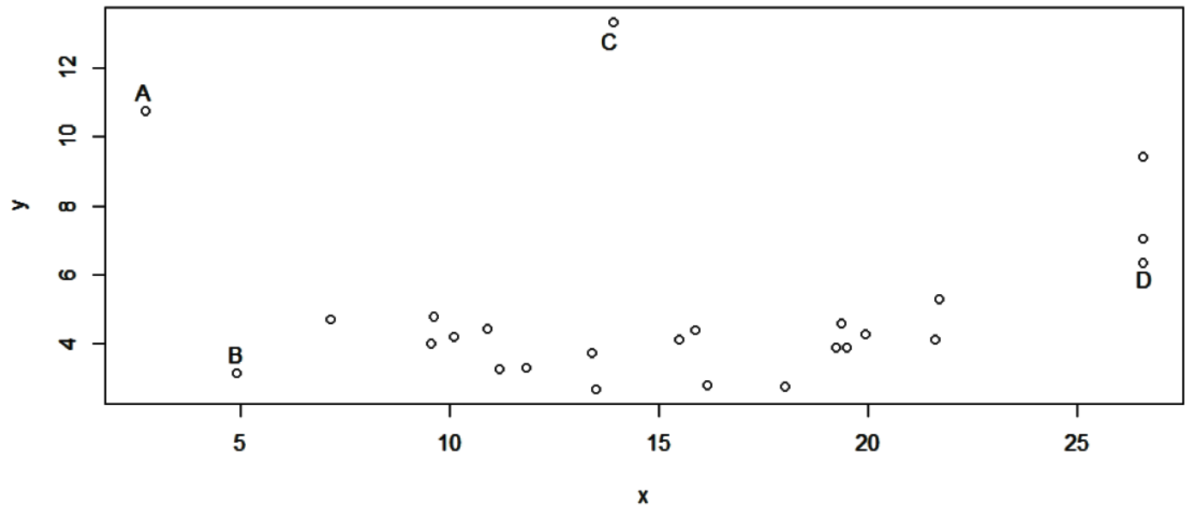
Model	x1	x2	x3	x4	AIC
1					23.723
2	x				-22.456
3		x			24.359
4	x	x			-51.994
5			x		23.970
6	x		x		-20.822
7		x	x		24.242
8	x	x	x		-50.745
9				x	-17.395
10	x			x	-58.310
11		x		x	-15.719
12	x	x		x	-79.601
13			x	x	-26.388
14	x		x	x	-58.273
15		x	x	x	-25.111
16	x	x	x	x	-78.503

(For instance, AIC = 24.242 for Model #7: $y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + e$*)*

(a) [9 points] Note that AIC increases slightly when x_2 is added to the intercept model (i.e. from Model #1 to Model #3), but decreases significantly when x_2 is added from Model #14 to Model #16. Explain in details on why the changes in AIC are in opposite direction.

(b) [4 points] Implement the forward selection method based on AIC. Show your steps in details on how you come up with the parsimonious model.

Problem 7 [18 points]: The scatterplot below shows n=25 data points $\{(x_i, y_i), i = 1, 2, \ldots, 25\}$:



A quadratic regression was fitted to the data (R-codes: x2<-x^2; fit0<-lm(y~x+x2)), and the following are the output from the command influence.measures(fit0):

Observation	dfb.1_	dfb.x	dfb.x2	dffit	cov.r	cook.d	hat
1	0.0823	-0.106	0.0969	-0.17	1.1682	9.95E-03	0.0689
2	0.0281	-0.034	0.028	-0.062	1.2247	1.34E-03	0.0682
3	-0.108	0.073	-0.051	-0.138	1.2874	6.60E-03	0.1238
4	0.0007	0.0008	-0.001	0.0047	1.2313	7.64E-06	0.0662
5	-0.034	0.0074	0.0057	-0.083	1.226	2.42E-03	0.0738
6	-0.014	-9E-04	0.0076	-0.045	1.231	7.17E-04	0.0697
7	-0.006	0.0051	-0.001	0.0234	1.247	1.91E-04	0.0788
8	-0.641	1.2363	-1.363	2.2047	0.0172	4.08E-01	0.0683
9	-0.008	0.0119	-0.012	0.0186	1.2357	1.21E-04	0.0702
10	-0.012	-0.027	0.0419	-0.119	1.196	4.88E-03	0.0657
11	0.0301	-0.036	0.0287	-0.068	1.223	1.62E-03	0.0683
12	-0.007	0.0085	-0.007	0.0159	1.2334	8.87E-05	0.0683
13	0.0626	-0.091	0.091	-0.14	1.1922	6.74E-03	0.0703
14	0.0302	-0.028	0.0093	-0.118	1.218	4.79E-03	0.0778
15	0.014	-0.016	0.0122	-0.034	1.2322	4.01E-04	0.0689
16	-0.034	0.0533	-0.072	-0.113	1.5778	4.45E-03	0.2748
17	-0.098	0.1515	-0.206	-0.322	1.5252	3.57E-02	0.2748
18	1.6979	-1.509	1.3181	1.7191	1.4378	8.95E-01	0.4787
19	0.0074	-0.016	0.0186	-0.032	1.2305	3.47E-04	0.0674
20	0.0376	-0.081	0.0911	-0.153	1.178	8.04E-03	0.0676
21	0.1881	-0.29	0.3954	0.6165	1.3784	1.27E-01	0.2748
22	0.0044	-9E-04	-8E-04	0.0109	1.2406	4.11E-05	0.0734
23	0.0015	-0.035	0.0462	-0.104	1.2033	3.76E-03	0.0654
24	-0.022	0.0325	-0.033	0.0501	1.2309	8.74E-04	0.0703
25	-0.93	0.7606	-0.626	-0.993	1.0173	3.01E-01	0.2455

Of the 25 observations in the table,

(a) which observation corresponds to data point A in the scatterplot? Explain.

(b) which observation corresponds to data point B in the scatterplot? Explain.

(c) which observation corresponds to data point C in the scatterplot? Explain.

(d) which observation corresponds to data point D in the scatterplot? Explain.

- End of the Exam -