UNIVERSITY OF TORONTO SCARBOROUGH
Department of Computer and Mathematical Sciences
Sample Midterm Test

STAC67H3 Regression Analysis
Duration: One hour and fifty minutes

Last Name:________________________ First Name: ________________________

Student number: ________________________

Aids allowed:

- The textbook: Applied Regression Analysis by Ketner et al published by McGraw Hill
- Class notes
- A calculator (No phone calculators are allowed)

No other aids are allowed. For example you are not allowed to have any other textbook or past exams.

All your work must be presented clearly in order to get credit. Answer alone (even though correct) will only qualify for **ZERO** credit. Please show your work in the space provided; you may use the back of the pages, if necessary, but you MUST remain organized. Show your work and answer in the space provided, in ink. Pencil may be used, but then any re-grading will NOT be allowed.

There are 14 pages including this page. Please check to see you have all the pages.

Good Luck!

Question:	1	2	3	4	Total
Points:	19	20	10	11	60
Score:					

1. The following data (and the summary statistics) relate biomass production of soybeans to cumulative intercepted solar radiation over an eight-week period following emergence. Biomass production is the mean dry weight in grams of independent samples of four plants (Virginia Lesser and Mike Unsworth).

Row	SolarRadiation(x)	PlantBiomass(y)
1	29.7	16.6
2	68.4	49.1
3	120.7	121.7
4	217.2	219.6
5	313.5	375.5
6	419.1	570.8
7	535.9	648.2
8	641.5	755.6

$\sum x = 2346$, $\sum y = 2757$, $\sum x^2 = 1039943.1$, $\sum y^2 = 1523629$, $\sum xy = 1255267$
Assume that the data satisfied the necessary assumptions for the tests and confidence intervals below.

(a) (6 points) Calculate b_0 and b_1 for the linear regression of plant biomass(y) on intercepted solar radiation(x).

Solution: $b_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{1255267 - \frac{2346\times 2757}{8}}{1039943.1 - \frac{2346^2}{8}} = 1.27$.
$b_0 = \bar{y} - b_1\bar{x} = \frac{2757}{8} - 1.27 \times \frac{2346}{8} = -27.8$

(b) (6 points) Calculate a 95% confidence interval for β_1.

Solution: $SSTot = 1523629 - \frac{2757^2}{8} = 573497.875$, $SSReg = b_1^2 SS_{XX} = 1.27^2 \times (1039943 - \frac{2346^2}{8}) \approx 567706$, $SSE = SSTot - SSReg \approx 573497.875 - 567706 \approx 5791.875$, $s = \sqrt{MSE} = \sqrt{5791.9/6} \approx 31$, $s_{b_1} = \frac{s}{\sqrt{SS_{XX}}} = \frac{31}{\sqrt{1039943.1 - \frac{2346^2}{8}}} \approx 0.05$ and the CI is $b_1 \pm t_{0.975,6} s_{b_1} = 1.27 \pm 2.4469 \times 0.05$.

(c) (3 points) Test the null hypothesis $H_0 : \beta_1 = 1$ against the alternative $H_1 : \beta_1 \neq 1$.

Solution: $t = \frac{b_1 - 1}{s_{b_1}} = \frac{1.27 - 1}{0.05} = 5.4$ and the p-value < 0.05 and so reject $H_0 : \beta_1 = 1$ in favour of the alternative $H_1 : \beta_1 \neq 1$. (Or just see that $1 \notin CI$ in part b above).

(d) (4 points) Calculate the 95% confidence interval estimates of the mean biomass production for X = 300.

Question 1 continues on the next page. . .

Solution: $SS_{XX} = 1039943.1 - \frac{2346^2}{8} = 351978$. The predicted value $\hat{y} = -27.8 + 1.27 \times 300 = 353.4$ and the CI for the mean at $x = 300$ is $= \hat{y} \pm t_{n-2,\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{SS_{XX}}} = 353.4 \pm 2.4469 \times 31 \times \sqrt{\frac{1}{8} + \frac{\left(300 - \frac{2346}{8}\right)^2}{351978}}$.

Question 1 continues on the next page. . .

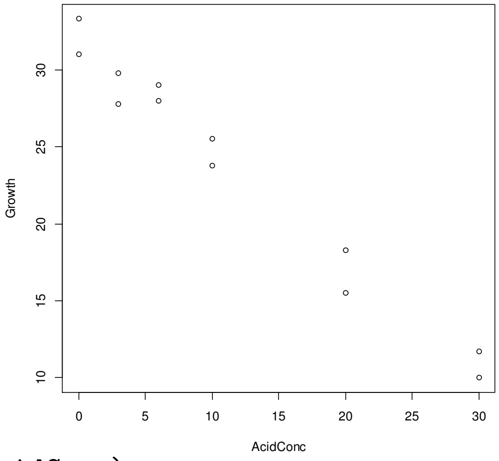
You may continue your to answer to question 1 on this page

2. (Samuel, M. et. al.) Laetisaric acid is a compound that holds promise for control of fungus diseases in crop plants. The accompanying data show the results of growing the fungus Pythium ultimum in various concentrations(in μG/ml) of laetisaric acid. Each growth value is the average of four radial measurements(mm) of a colony grown in a petri dish for 24 hours. Some R outputs (with codes) from the analysis of the data from this study based on the Normal error regression model are given below:

```
> fungus=read.table("C:/Users/Mahinda/Desktop/fungus.txt", header=T)
> fungus
   Row AcidConc Growth
1    1        0   33.3
```

⋮ only the first and the last observations are printed here

```
12  12       30   10.0
> Growth <- fungus[,3]
> AcidConc <- fungus[,2]
> plot(AcidConc, Growth)
```



```
> fit = lm(Growth ~AcidConc)
> summary(fit)

Call: lm(formula = Growth ~ AcidConc)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0896 -0.8498  0.2743  0.9004  1.4702

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.82979    0.55693   57.15 6.53e-14 ***
 AcidConc   -0.71201    0.03589  -19.84 2.32e-09 ***
```

Question 2 continues on the next page…

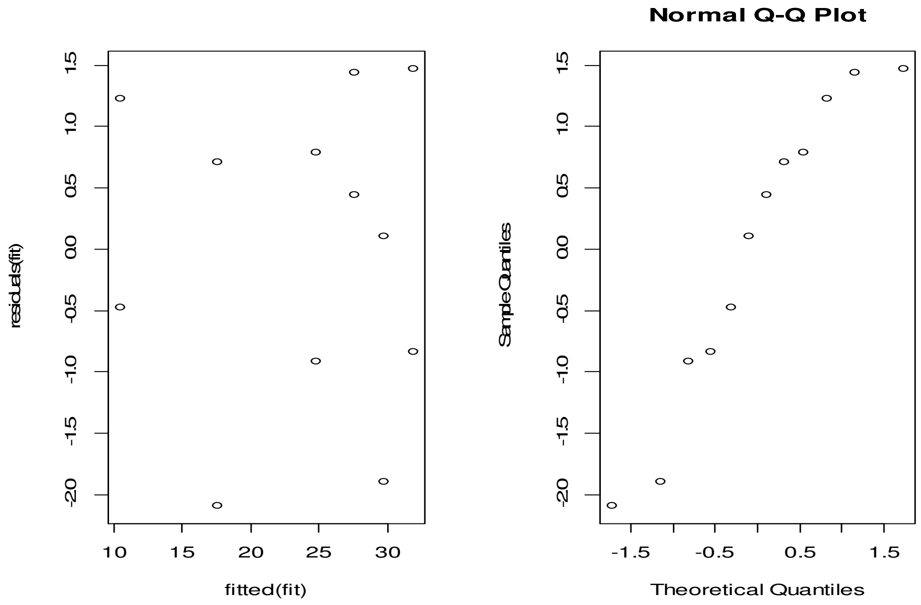
```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.295 on 10 degrees of freedom Multiple
R-squared: 0.9752,     Adjusted R-squared: 0.9727 F-statistic: 393.6
on 1 and 10 DF,  p-value: 2.321e-09

> anova(fit)
Analysis of Variance Table

Response: Growth
          Df Sum Sq Mean Sq F value    Pr(>F)
AcidConc   1 660.57  660.57  393.64 2.321e-09 ***

Residuals 10 16.78    1.68
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> par(mfrow=c(1,2))
> plot(fitted(fit),residuals(fit))
> qqnorm(residuals(fit))
```



```
> x0 <- data.frame(AcidConc=15)
> predict(fit, x0, interval="confidence", level=0.95)
       fit      lwr     upr
1 21.14963 20.27066 22.0286
> predict(fit, x0, interval="prediction", level=0.95)
       fit      lwr      upr
1 21.14963 18.13238 24.16688
```

Question 2 continues on the next page. . .

(a) (3 points) Describe what you learn from the scatterplot of growth on Acid concentration.

Solution: A strong, linear, negative relationship.

(b) (2 points) What proportion of the variability in fungal growth is explained by this regression linear regression with acid concentration?

Solution: 97.52% (The proportion of the variability in fungal growth is explained by this regression linear regression with acid concentration means R-sq)

(c) (4 points) Comment on the residual plots (i.e the plot of residuals vs fitted values and the Normal QQ plot of residuals)

Solution: The plot of residuals vs fitted values looks random. This indicates that there is no serious violation of the assumption of independence of errors. The Normal QQ plot is close to s straight line indicating no serious violation of the assumption of Normality of errors.

Question 2 continues on the next page. . .

Note: For the remaining parts of this question, assume that the data satisfied all the necessary assumptions, regardless of your answers to parts a, b and c above.

(d) (3 points) Is there significant evidence (at $\alpha = 0.05$) for a linear relationship between fungus growth and acid concentration? Give reasons to your answer.

Solution: Yes, the p-value for testing the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 \neq 0$ is less than 0.05 (p-value = 2.32e-09)

(e) (3 points) Construct a 95% confidence interval for the change in mean fungal growth when the acid concentration increases by 1 μG/ml.

Solution: $-0.71201 \pm t_{0.975,10} \times 0.03589 = -0.71201 \pm 2.228 \times 0.03589$ (df can be obtained from the ANOVA table)

(f) (5 points) Calculate a **90%** confidence interval for the mean fungal growth when the acid concentration is 15 μG/ml.
Note This question requires a numerical answer. Just stating whether this interval is wider or narrower is not an acceptable answer.

Solution: Note that the R output gives the 95% CI for the mean of Y at x = 15, and the question wants the 90% CI.
$\hat{y} = 21.14963$ and $t_{0.975,10} \times s_{\hat{y}} = (22.0286 - 20.27066)/2 = 0.87897$ and so $s_{\hat{y}} = 0.87897/2.228 = 0.394510772$ and the 90% CI for the mean of Y is $\hat{y} \pm$

Question 2 continues on the next page. . .

$t_{0.95,10} \times s_{\hat{y}} = 21.14963 \pm 1.812 \times 0.394510772 = (20.43, 21.86)$.
Here is the R output (just to compare the above calculations)

```
> predict(fit, x0, interval="confidence", level=0.90)
       fit      lwr      upr
1 21.14963 20.43464 21.86462
```

3. A simple linear regression was run on a data set with $n = 6$ observations . You are given only the following information:

- The regression equation is y = - 5.36 + 0.0405 x
- The value of the t-test statistic for testing null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 \neq 0$ is 9.25.
- $MSE = 2.16$

(a) (4 points) Calculate a 95% confidence interval for β_1.

Solution: $s_{b_1} = 0.0405/9.25 = 0.004378$, $t(0.975, 6 - 2) = 2.776$ and so the CI is $(0.0405 \pm 2.776 \times 0.004378) = (0.028, 0.053)$

(b) (6 points) Calculate and interpret the coefficient of determination (R^2).

Solution: $F = \frac{MSReg}{MSE} = t^2 = 9.25^2 \implies MSReg = 9.25^2 \times 2.16 = 184.815$. $SSE = MSE \times (n-2) = 2.16 \times 4 = 8.64$ and $SSReg = MSReg \times 1 = 184.815$, $SSTot = SSReg + SSE = 184.815 + 8.64 = 193.455$, $R^2 = \frac{SSReg}{SSTot} = \frac{184.815}{193.455} = 0.955$, i.e. 95.5% of the variability in y is explained by this linear regression with x.

Here is a software(MINITAB) output (just to compare the calculated values):

```
The regression equation is y = - 5.36 + 0.0405 x


Predictor      Coef   SE Coef     T      P

Constant     -5.360 1.558     -3.44  0.026

x          0.040488 0.004375   9.25  0.001


S = 1.46824   R-Sq = 95.5%   R-Sq(adj) = 94.4%

Analysis of Variance

Source          DF      SS      MS      F      P

Regression       1  184.63  184.63  85.65  0.001

Residual Error   4    8.62    2.16

Total            5  193.26
```

Question 3 continues on the next page. . .

You may continue your to answer to question 3 on this page

4. Consider the regression model with no intercept given by $Y_i = \beta x_i + \varepsilon_i, i = 1 \ldots n$ with the following assumptions:

    1. ε_i's are independent.
    2. $E(\varepsilon_i) = 0$ for $i = 1 \ldots n$
    3. $Var(\varepsilon_i) = \sigma^2$ for $i = 1 \ldots n$
    4. x_i's are known constants.

    Note that ε_i's are NOT necessarily Normally distributed and recall that in class we showed (in an exercise) that $B = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$ is the least squares estimator of β and it is an unbiased estimator of β.

    (a) (3 points) Prove that $Var(B) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$

    Solution: $Var(B) = \frac{\sum_{i=1}^n x_i^2 Var(Y_i)}{\left(\sum_{i=1}^n x_i^2\right)^2} = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{\left(\sum_{i=1}^n x_i^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}.$

    (b) (5 points) Prove that $E(Y_i - Bx_i)^2 = \sigma^2 \left(1 - \frac{x_i^2}{\sum_{i=1}^n x_i^2}\right)$

    Solution: $E(Y_i - Bx_i)^2 = E(Y_i - \beta x_i + \beta x_i - Bx_i)^2 = Var(Y_i) + x_i^2 Var(B) - 2x_i Cov(Y_i, B) = \sigma^2 + \frac{\sigma^2 x_i^2}{\sum_{i=1}^n x_i^2} - \frac{2x_i^2}{\sum_{i=1}^n x_i^2} Var(Y_i) = \sigma^2 - \frac{\sigma^2 x_i^2}{\sum_{i=1}^n x_i^2} = \sigma^2 \left(1 - \frac{x_i^2}{\sum_{i=1}^n x_i^2}\right).$

    (c) (3 points) Prove that $S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - Bx_i)^2$ is an unbiased estimator of σ^2.

    Solution: $E(S^2) = \frac{1}{n-1} \sum_{i=1}^n E(Y_i - Bx_i)^2) = \frac{1}{n-1} \sum_{i=1}^n \sigma^2 \left(1 - \frac{x_i^2}{\sum_{i=1}^n x_i^2}\right) = \sigma^2.$

Question 4 continues on the next page…

You may continue your to answer to question 4 on this page

END OF TEST

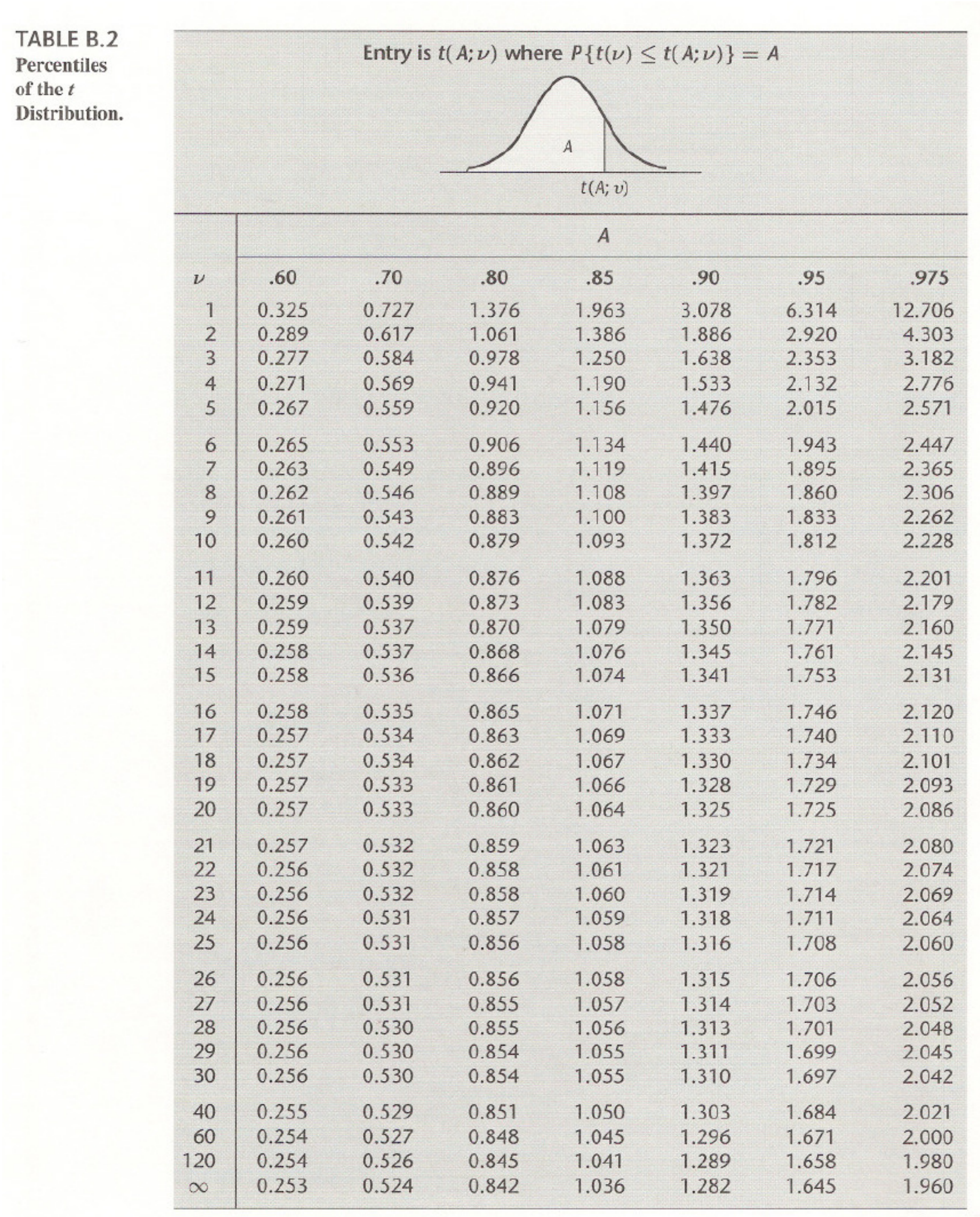


TABLE B.2
Percentiles of the *t* Distribution.

Entry is $t(A;\nu)$ where $P\{t(\nu) \leq t(A;\nu)\} = A$

ν	.60	.70	.80	.85	.90	.95	.975
				A			
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.537	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960