School of Mathematics and Statistics

Te Kura Mātai Tatauranga

STAT 292 | **Test 2** | **Due by 8:00am, Saturday 20 June 2020**

Instructions: There are 10 questions given on pages 2–14 worth a total of 100 marks.

Answer **ALL** questions.

Solutions must be either typed or written neatly, and questions must be answered in order.

Be sure to submit your Test 2 answers as a PDF file and follow the instructions specified in the submission system.

By proceeding with this Test you are in agreement and consent to comply with the following.

Recognising the trust that the University and the Academic Staff teaching this course have placed in me in this current situation, I affirm that:

- I have logged on to Blackboard using my own credentials;
- I will complete all parts of this Test on my own;
- I will not give anyone else access to this Test; and
- I understand that breaking any of the above will likely constitute academic misconduct at level 2 or 3, to be investigated according to the Student Conduct Statute, with consequences for my studies as per the statute.

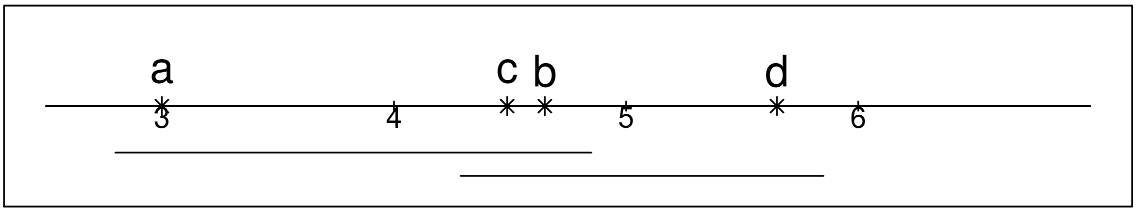
Section A: Multi-Choice [40 marks]

For Section A questions there is a single correct answer in each case. Only record the letter corresponding to your answer. Do not present working to support your choice of answer.

Note that the possible answers for each question are ordered alphabetically (or by 'not significant' then 'significant', etc.), or they are listed in ascending numerical order.

Use the following information for Questions 1 to 3 [5 marks each]

A one-way ANOVA was estimated to see if a single factor with four levels (denoted **a**, **b**, **c** and **d**) had any effect on a certain response variable, Y. A balanced design was used, and the conventional null hypothesis was rejected at a 5% significance level. A post-hoc Tukey test was carried out with a 5% experiment-wise error rate and the results from the Tukey test are presented in the following underlining diagram.



1. The advantage of a balanced design is that it:
   (a) allows the use of Q-Q plots to check for normality of the error random variables.
   (b) ensures constant variance of the errors.
   (c) ensures there are no outliers in the data.
   (d) gives the highest power for the ANOVA.
   (e) helps to reduce bias.

2. The sample mean of the response variable, $\overline{y}$, is approximately equal to:
   (a) 3.14
   (b) 4.00
   (c) 4.45
   (d) 5.65
   (e) 6.00

3. The Tukey test and underlining diagram indicate that the population mean of Y:
   (a) has no significant differences between any levels of the factor.
   (b) has no significant differences between levels **b** and **c** but has significant differences between (**b** and **c**) and all other levels of the factor.
   (c) is significantly lower at level **a** than at all other levels of the factor.
   (d) is significantly lower at level **a** than at level **d**.
   (e) has significant differences between all levels of the factor.

Use the following information for Questions 4 and 5 [5 marks each]

Consider the random effects model for a one-way ANOVA with $p = 5$ groups,

$$Y_{ij} = \mu + A_i + E_{ij}$$

where Y_{ij} is the j^{th} observation in the i^{th} group, $i = 1, 2, 3, 4, 5$.

4. The number of random variables in the random effects model for a one-way ANOVA with $p = 5$ groups is:

   (a) 1
   (b) 2
   (c) 3
   (d) 4
   (e) 5

5. The number of components of variation in the random effects model for a one-way ANOVA with $p = 5$ groups is:

   (a) 1
   (b) 2
   (c) 3
   (d) 4
   (e) 5

6. [5 marks] In a linear multiple regression model the:

   (a) errors are all assumed to be zero.
   (b) errors are assumed to be independent of all the explanatory variables.
   (c) residuals are spread evenly around the line of best fit in the Q-Q plot.
   (d) response variable is assumed to be independent of all the explanatory variables.
   (e) response variable is assumed to be independent of the error term.

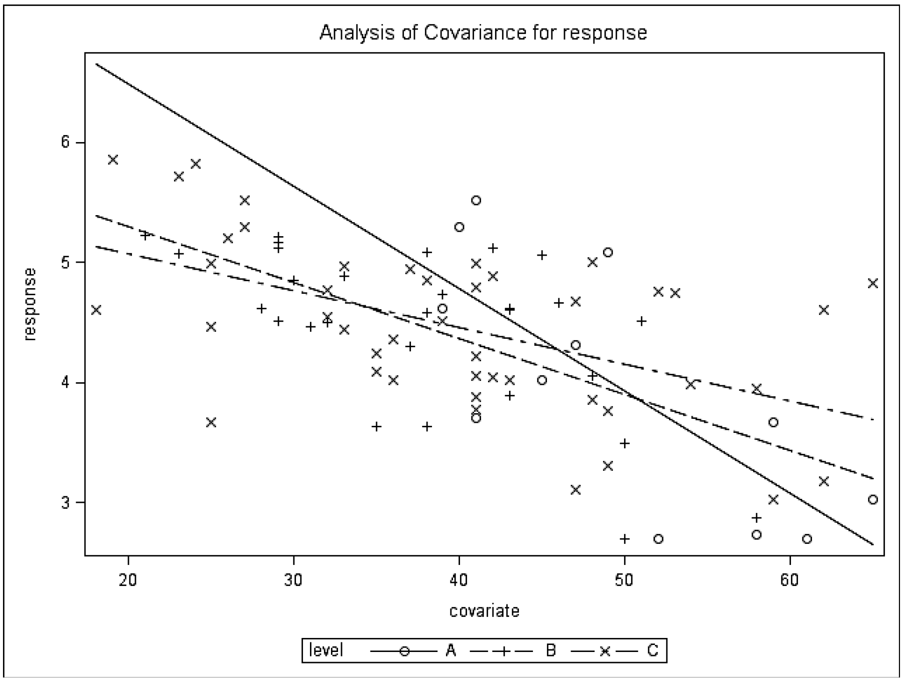
Use the following information for Questions 7 and 8 [5 marks each]

Consider the complete model for an ANCOVA with response variable Y, one qualitative factor with $p = 3$ levels, and one covariate, x:

$$Y = \alpha_i + \beta_i x + E,$$

for the factor at level i, $i = 1, 2, 3$.

A graph showing that complete model fitted to 85 observations from an unbalanced, observational study with data from 3 groups labelled A, B and C follows.



7. The relationship between the response variable and the covariate is:
   (a) different, depending on the level of the factor, but always negative.
   (b) not displayed in that graph.
   (c) not significant, since the lines have different slopes.
   (d) of no interest, since the covariate is always a nuisance variable in any ANCOVA.
   (e) positive in every ANCOVA model by assumption, but sometimes it is estimated to be negative due to sampling variability, as in this case.

8. Given there are $n = 85$ observations and $p = 3$ levels of the factor, the number of independent parameters fitted in the complete ANCOVA model, as displayed in the graph, is:
   (a) $3 = p$
   (b) $4 = p + 1$
   (c) $5 = 2p - 1$
   (d) $6 = 2p$
   (e) $82 = n - p$

Section B: Written Answers [60 marks]

For Section B questions, you must write out your answers. Ensure you label your work, to make it clear which part of each question you are answering.

9. **[20 marks]**

   Four educational tests for 10-year-olds have been developed. They are meant to be of the same standard, so that they may be used interchangeably. Forty 10-year-olds were randomly selected, and ten were allocated randomly to each test. Their scores out of 100 follow:

Test	Y = Score out of 100										Mean
Test 1	64	81	53	73	50	67	69	53	53	52	61.5
Test 2	53	60	64	56	71	60	57	56	54	71	60.2
Test 3	52	54	69	51	61	61	60	74	45	59	58.6
Test 4	76	69	63	61	68	48	77	74	72	68	67.6

   Relevant SAS output is on pages 6 and 7.

   (a) Give the model equation and hypotheses for a one-way analysis of variance. State the meaning of all the terms in the equation.

   (b) State the assumptions of a one-way ANOVA. Do you think they are satisfied? Justify your answer.

   (c) Using a 5% significance level, give the test statistic, the degrees of freedom and the p-value for the test, and your statistical conclusion plus interpretation of the result.

   (d) Failure to reject a null hypothesis may be caused by different groups having equal population means. However, if there really is a difference of population means, there may still be failure to reject H_0. Give a statistical reason why this may happen.

SAS Output for Q9

One-Way Analysis of Variance

Results

The ANOVA Procedure

Class Level Information		
Class	**Levels**	**Values**
test	4	T1 T2 T3 T4

Number of Observations Read	40
Number of Observations Used	40

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	464.075000	154.691667	2.01	0.1293
Error	36	2764.900000	76.802778		
Corrected Total	39	3228.975000			

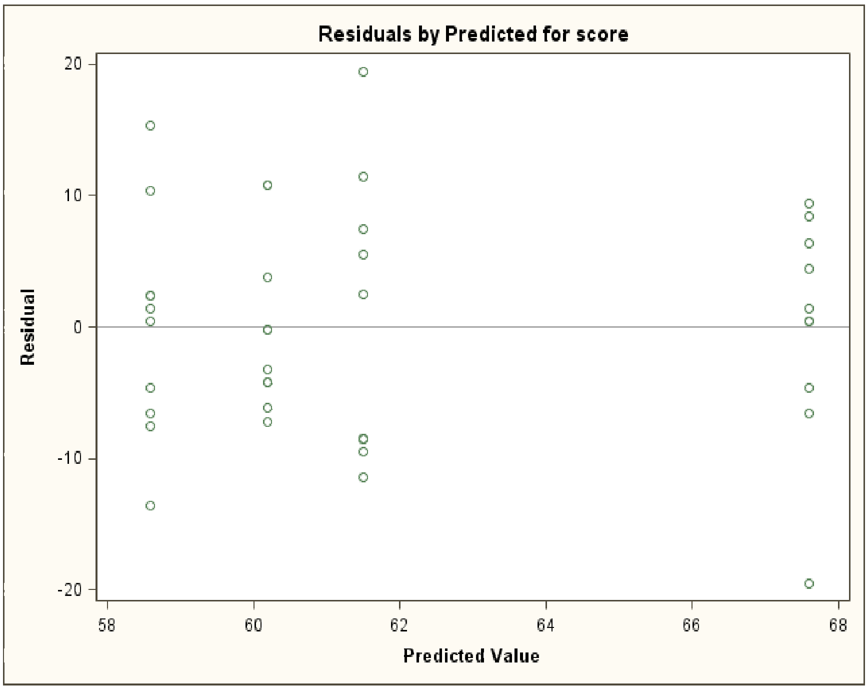
R-Square	Coeff Var	Root MSE	score Mean
0.143722	14.14073	8.763719	61.97500

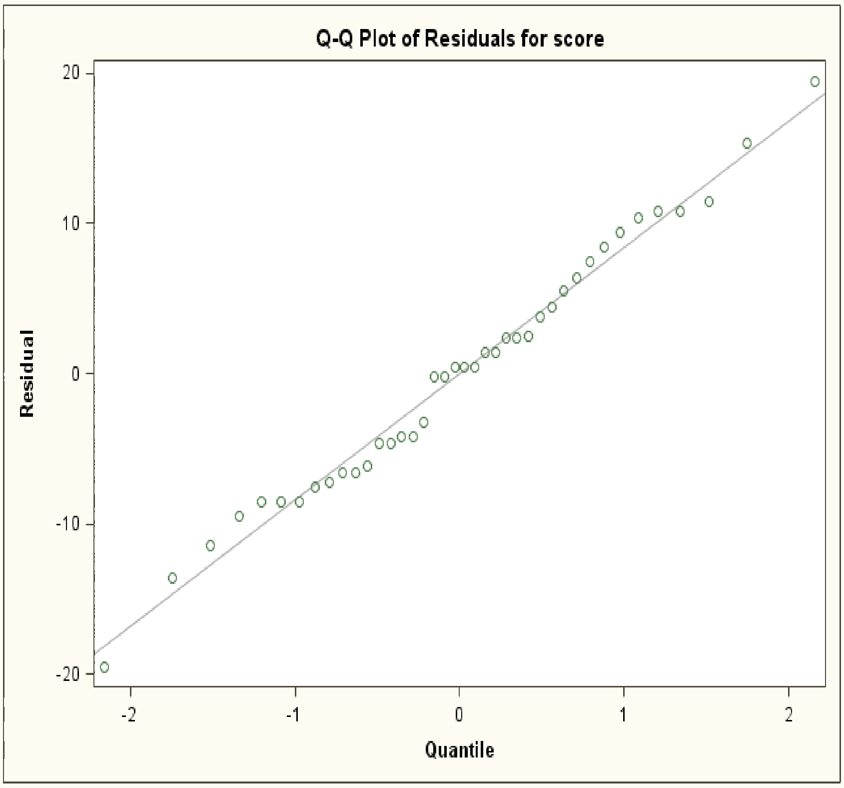
Source	DF	Anova SS	Mean Square	F Value	Pr > F
test	3	464.0750000	154.6916667	2.01	0.1293

Dependent Variable: score

Levene's Test for Homogeneity of score Variance ANOVA of Squared Deviations from Group Means					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
test	3	22048.6	7349.5	0.88	0.4586
Error	36	299311	8314.2		

SAS Output for Q9 continued





10. **[40 marks]**

Sixty children (who were not colour-blind) aged 7 to 12 years were given the online Farnsworth-Munsell 100 hue test, in which they sort blocks of colour into order by hue from purple to magenta. Their scores are TES = total error score, with a score of 0 indicating perfect sorting by colour. This is regarded as an exploratory experiment, with possible predictors of the TES scores being Age (in years) and Time (time taken to do the sorting, in seconds).

TES = total error score
Age = age in years
Time = time taken (seconds)

Age (yrs)	Variable	Observed value									
7	TES	180	197	173	171	130	208	180	224	173	148
	Time (sec)	228	183	263	212	259	223	311	256	217	260
8	TES	129	191	136	157	160	195	154	179	191	178
	Time (sec)	341	336	355	308	313	220	283	191	278	373
9	TES	166	104	127	179	120	139	120	122	165	94
	Time (sec)	351	207	375	227	347	295	225	372	343	392
10	TES	131	169	109	144	115	108	91	156	101	111
	Time (sec)	206	299	389	212	380	264	320	202	351	265
11	TES	113	102	99	98	94	102	106	103	83	90
	Time (sec)	212	186	384	347	217	342	189	209	345	383
12	TES	82	92	64	68	83	85	83	71	73	93
	Time (sec)	228	247	299	317	370	259	376	261	286	261

SAS output from fitting six different regression models is on pages 9 to 14. A new variable logTES = log(TES) was also defined in the data file.

(a) In general, for any data set, give the model equation for a multiple regression with two predictors, X_1 and X_2. Define all the terms in the model.

(b) State the assumptions for a multiple regression model with two predictors, X_1 and X_2.

(c) For this specific data set, we have X_1 = Age and X_2 = Time. Why is the analysis using Y = logTES preferable to the one using Y = TES? Support your answer by referring to the SAS output.

(d) Using the logTES analysis, compare all the models, with one and two predictors, using t-tests and a 5% significance level. For each comparison, give the hypotheses, the t statistic and the p-value. You may present these results either in words or using a lattice diagram of the models.

(e) Interpret your results for all three models using the logTES analysis from part (d). Include:

i. comments for each model on the meaning of the negative signs of the beta (slope) estimates,

ii. a brief explanation of why Time is significant in the multiple regression but not significant if it is the only predictor.

SAS Output for Q10

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: TES

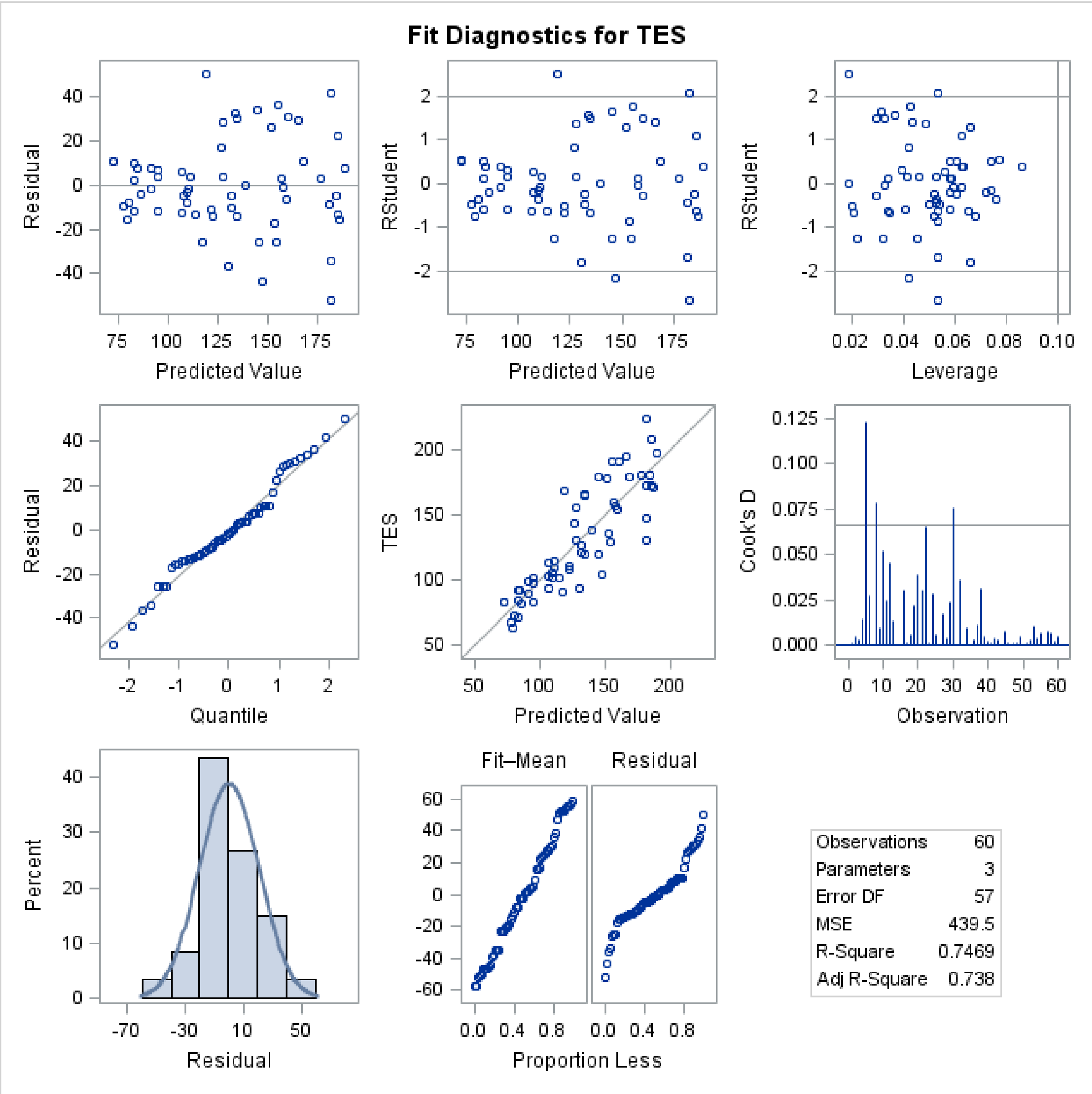
Number of Observations Read	60
Number of Observations Used	60

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	73930	36965	84.11	<.0001
Error	57	25051	439.49956		
Corrected Total	59	98982			

Root MSE	20.96424	R-Square	0.7469
Dependent Mean	130.15000	Adj R-Sq	0.7380
Coeff Var	16.10776		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	344.92450	18.38970	18.76	<.0001
Age	1	-19.82001	1.59749	-12.41	<.0001
Time	1	-0.09266	0.04241	-2.19	0.0330

SAS Output for Q10 continued



SAS Output for Q10 continued

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: TES

Number of Observations Read	60
Number of Observations Used	60

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	71832	71832	153.45	<.0001
Error	58	27150	468.10034		
Corrected Total	59	98982			

Root MSE	21.63563	R-Square	0.7257
Dependent Mean	130.15000	Adj R-Sq	0.7210
Coeff Var	16.62361		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	322.62000	15.78631	20.44	<.0001
Age	1	-20.26000	1.63550	-12.39	<.0001

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: TES

Number of Observations Read	60
Number of Observations Used	60

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6276.68814	6276.68814	3.93	0.0523
Error	58	92705	1598.36141		
Corrected Total	59	98982			

Root MSE	39.97951	R-Square	0.0634
Dependent Mean	130.15000	Adj R-Sq	0.0473
Coeff Var	30.71803		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	175.59005	23.50406	7.47	<.0001
Time	1	-0.15897	0.08022	-1.98	0.0523

SAS Output for Q10 continued

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: logTES

Number of Observations Read	60
Number of Observations Used	60

Analysis of Variance

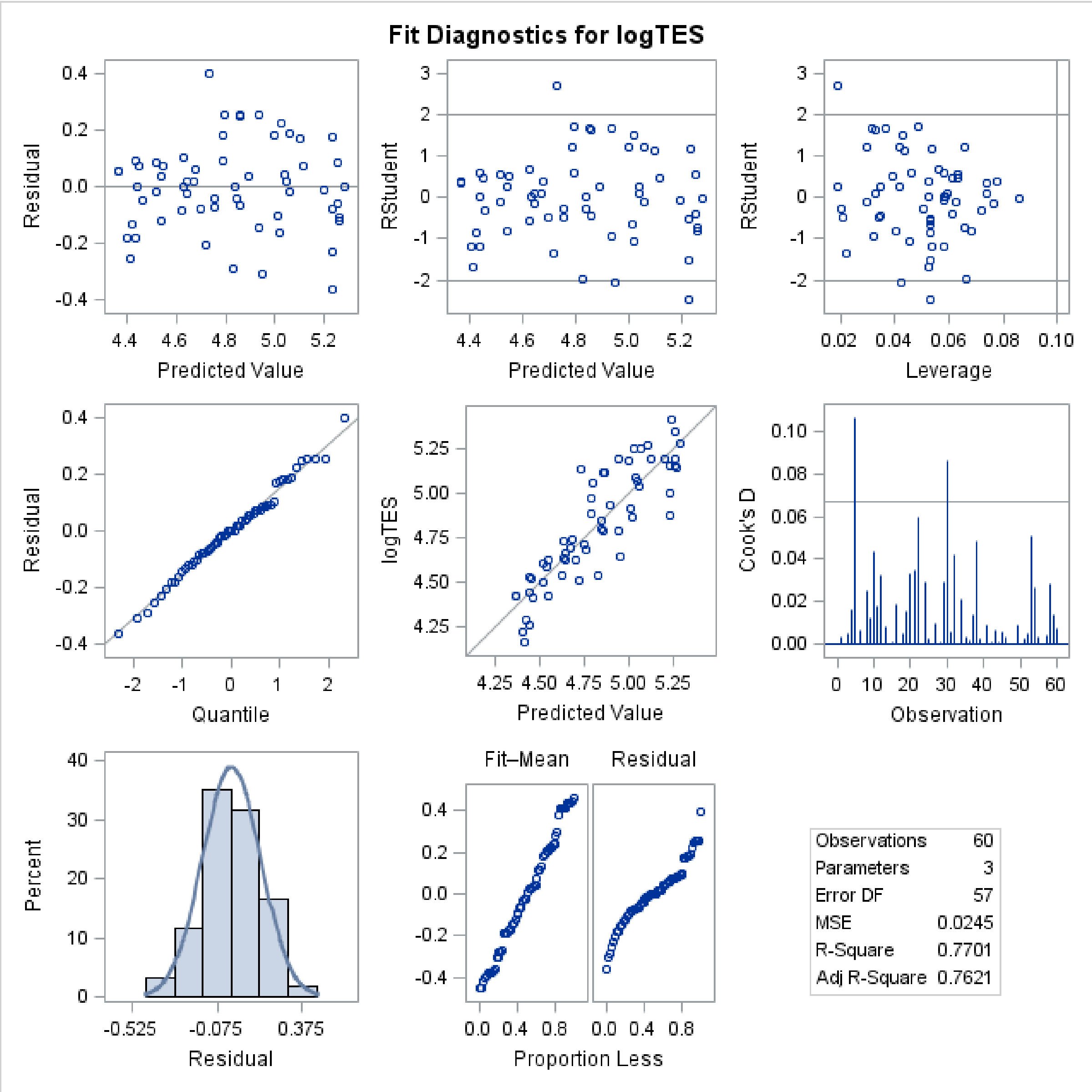
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4.67546	2.33773	95.48	<.0001
Error	57	1.39558	0.02448		
Corrected Total	59	6.07103			

Root MSE	0.15647	R-Square	0.7701
Dependent Mean	4.81833	Adj R-Sq	0.7621
Coeff Var	3.24745		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.50972	0.13726	47.43	<.0001
Age	1	-0.15859	0.01192	-13.30	<.0001
Time	1	-0.00064657	0.00031650	-2.04	0.0457

SAS Output for Q10 continued



SAS Output for Q10 continued

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: logTES

Number of Observations Read	60
Number of Observations Used	60

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.57328	4.57328	177.10	<.0001
Error	58	1.49775	0.02582		
Corrected Total	59	6.07103			

Root MSE	0.16070	R-Square	0.7533
Dependent Mean	4.81833	Adj R-Sq	0.7490
Coeff Var	3.33510		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.35408	0.11725	54.19	<.0001
Age	1	-0.16166	0.01215	-13.31	<.0001

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: logTES

Number of Observations Read	60
Number of Observations Used	60

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.34417	0.34417	3.49	0.0670
Error	58	5.72686	0.09874		
Corrected Total	59	6.07103			

Root MSE	0.31423	R-Square	0.0567
Dependent Mean	4.81833	Adj R-Sq	0.0404
Coeff Var	6.52150		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.15482	0.18474	27.90	<.0001
Time	1	-0.00118	0.00063053	-1.87	0.0670