

--	--	--

**Semester Two 2017
Examination Period****Faculty of Business and Economics**

EXAM CODES: ETC1000 / ETW1000

TITLE OF PAPER: BUSINESS AND ECONOMIC STATISTICS - PAPER 1

EXAM DURATION: 2 hours writing time

READING TIME: 10 minutes

THIS PAPER IS FOR STUDENTS STUDYING AT: (tick where applicable)

- | | | | | |
|------------------------------------|---|--|--|--|
| ☐ Berwick | ☒ Clayton | ☒ Malaysia | ☐ Off Campus Learning | ☐ Open Learning |
| ☐ Caulfield | ☐ Gippsland | ☐ Peninsula | ☐ Monash Extension | ☐ Sth Africa |
| ☐ Parkville | ☐ Other (specify) | | | |

During an exam, you must not have in your possession any item/material that has not been authorised for your exam. This includes books, notes, paper, electronic device/s, mobile phone, smart watch/device, calculator, pencil case, or writing on any part of your body. Any authorised items are listed below. Items/materials on your desk, chair, in your clothing or otherwise on your person will be deemed to be in your possession.

No examination materials are to be removed from the room. This includes retaining, copying, memorising or noting down content of exam material for personal use or to share with any other person by any means following your exam.

Failure to comply with the above instructions, or attempting to cheat or cheating in an exam is a discipline offence under Part 7 of the Monash University (Council) Regulations.

AUTHORISED MATERIALS

OPEN BOOK	☐ YES	☒ NO
CALCULATORS	☐ YES	☒ NO
SPECIFICALLY PERMITTED ITEMS	☐ YES	☒ NO

if yes, items permitted are:

Candidates must complete this section if required to write answers within this paper

STUDENT ID: _____

DESK NUMBER: _____

INSTRUCTIONS TO CANDIDATES:

Answer ALL questions in this examination paper.

Paper is out of 100 marks

Where you are asked to perform calculations, you should write out the solution as an equation containing the appropriate numerical values from within the question. You do not need to calculate exact values in order to receive full marks for that part of the question.

Question 1 (22 marks)

- (a) The pivot table below tabulates a sample of data on whether a woman is working or not working against the number of children they have.

Count of children		Column Labels					
Row Labels		0	1	2	3	4	5 Grand Total
Not Working		34.49%	48.35%	44.81%	53.41%	67.04%	76.72%
Working		65.51%	51.65%	55.19%	46.59%	32.96%	23.28%
Grand Total		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

- What is the probability of a woman not working? (2 marks)
- Interpret the value 65.51% in the pivot table. (2 marks)
- Discuss what independence means and whether you think the number of children a woman has is independent of their work status. (4 marks)

- (b) The table below provides some descriptive statistics on hours worked by women. For those who are not working the hours worked is zero.

<i>hours</i>	
Mean	17.148
Standard Error	0.189263
Median	10
Mode	0
Standard Deviation	18.92631
Sample Variance	358.2053
Kurtosis	-0.61262
Range	99
Minimum	0
Maximum	99
Sum	171480
Count	10000

- i. Comment on the skewness of the distribution. (2 marks)
- ii. With data that is Normally distributed then around ± 2 standard deviations covers 95% of the data. Construct this range and comment on why it might not be reasonable in this case. (4 marks)
- iii. Suppose that instead of measuring hours worked we measure *minutes* worked. What does this do to the mean, median, variance and standard deviation? (4 marks)
- iv. Suppose we discover two mistakes in the data. First, one individual was recorded as having worked zero hours but actually worked 10 hours. Second, an individual was recorded as working 99 hours but actually worked 89 hours. What would this do to the mean and the median and why? (4 marks)

Question 2 (32 marks)

- (a) We have data on whether particular students passed a course for two schools; A and B. School A has a student population of 100 students, school B has 200 students. We ran a regression with a dummy variable for School B (=1 if student attends School B, 0 otherwise).

The Dependent variable = 1 if the student passed the course, and =0 if they did not pass.

The results are shown below.

Regression Statistics						
Multiple R	0.083690525					
R Square	0.007004104					
Adjusted R Square	0.003671903					
Standard Error	0.450540317					
Observations	300					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	0.426666667	0.426666667	2.101945225	0.14816348	
Residual	298	60.49	0.202986577			
Total	299	60.91666667				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.77	0.045054032	17.09059037	4.04852E-46	0.681335625	0.858664375
School_B	-0.08	0.055179694	-1.449808686	0.14816348	-0.188591238	0.028591238

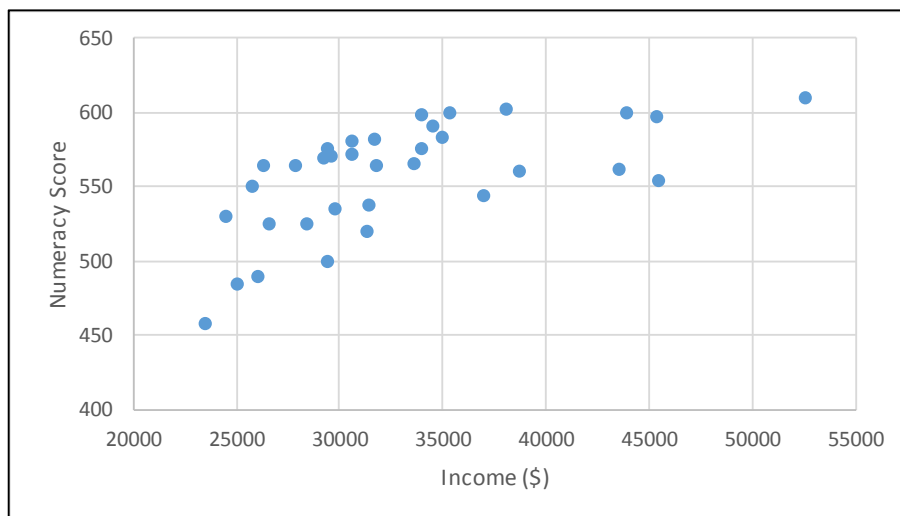
- Interpret the coefficient on the Intercept. (2 marks)
 - Explain in simple language what the values 0.6813 and 0.8586 (Lower 95% and Upper 95%) tell us. (2 marks)
 - Show how you would calculate the proportion of students that passed for the full sample of 300 students. (2 marks)
 - Some parents claim that School B is better than School A in helping people pass the course. Is there any evidence in the analysis above to support this claim? Explain your conclusion without reference to the *t Stat* or *P-value*. (3 marks)
- (b) We have a score which represents the level of numeracy skills of Year 9 students for each state (Victoria, NSW, ACT, SA, WA, NT, Tasmania, Queensland) by the geolocation of the school. The values of geolocation are; (1) Major City, (2) Inner Regional, (3) Outer Regional, (4) Remote, and (5) Very Remote. In addition to students' scores across all these locations, we also have each child's family income. We constructed two variables Y and X and estimated a regression with an intercept with results shown below.

$$Y = (\text{score}) / (\text{standard deviation of score})$$

$$X = (\text{income}) / (\text{standard deviation of income})$$

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	12.25148188	0.684960632	17.88640297	3.1962E-18	10.85626273	13.64670103
X	0.613089383	0.139655985	4.389997214	0.00011576	0.328619452	0.897559315

- What does the coefficient on X measure? (2 marks)
- How do you judge the magnitude of the coefficient on X? (3 marks)
- What does the intercept measure and does it provide useful information? (2 marks)
- The chart below plots the Numeracy Score against Income. Do you think there might be any problems with the previous regression and if so why? (3 marks)



- (c) Let us explore the results by region. We run a regression with dummy variables for each geolocation except “Major City” and an intercept.

The Dependent variable is Y = the child’s test score, a value between 0 and 800. The results are shown below.

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	594.9166667	11.61114785	51.23667998	5.33296E-30	571.1692029	618.6641304
d_InnerRegional	-16.2452381	15.8233273	-1.026663848	0.313066549	-48.60757613	16.11709994
d_OuterRegional	-48.65416667	15.36010483	-3.167567358	0.00360455	-80.06910836	-17.23922497
d_Remote	-51.03095238	15.8233273	-3.225045619	0.003112517	-83.39329041	-18.66861435
d_VeryRemote	-71.65	16.42064277	-4.363410191	0.000148003	-105.2339853	-38.06601467

- What do the coefficients of the Intercept and the VeryRemote variable tell us? (4 marks)
- Based on the results in (b), do you think we should have included income in this regression? Explain why or why not? (2 marks)
- If we include income in the regression this gives the following results. What do these results, and the regression results without income included, tell us about the likely relationship between income and the geolocation dummy variables. (3 marks)

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	489.8959	21.74029	22.53401	1.74E-19	445.3629	534.4288
income	0.002878	0.00055	5.237283	1.45E-05	0.001752	0.004004
d_InnerRegional	-5.07223	11.64244	-0.43567	0.666418	-28.9207	18.77623
d_OuterRegional	-30.4533	11.6411	-2.61601	0.014175	-54.299	-6.60756
d_Remote	-43.7948	11.52841	-3.79886	0.000719	-67.4097	-20.1799
d_VeryRemote	-59.3266	12.1082	-4.89971	3.65E-05	-84.1291	-34.5241

- (d) The previous data was for 2016. We also have data for 2008 and want to explore whether there is any improvement in numeracy score over this 8-year period. Below are the results of a regression of scores from 2016 and 2008 with a dummy variable for 2008.

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	562.7266667	6.089832696	92.40429003	1.15819E-64	550.5365421	574.9167912
d_2008	-9.663333333	8.612323991	-1.122035509	0.266471047	-26.90277284	7.576106178

Conduct a one-sided hypothesis test at the 5% level to investigate whether numeracy scores improved between 2008 and 2016. (4 marks)

Question 3 (27 marks)

- (a) We have data on weekly rental prices (\$) and housing characteristics for a particular suburb. Below we provide results from a regression of rent on the number of bedrooms (beds) and the number of bathrooms (baths).

<i>Regression Statistics</i>						
Multiple R	0.840366018					
R Square	0.706215044					
Adjusted R Square	0.702758751					
Standard Error	143.0367696					
Observations	173					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	2	8360874.797	4180437.399	204.3272726	6.06055E-46	
Residual	170	3478117.966	20459.51745			
Total	172	11838992.76				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	195.6819601	35.77533798	5.469744554	1.5883E-07	125.0608452	266.303075
beds	190.3711999	11.90915933	15.9852761	1.05763E-35	166.8623204	213.8800794
baths	113.7225809	23.86377274	4.765490443	4.02285E-06	66.61509482	160.830067

- Explain what the intercept measures in this regression and discuss whether it is meaningful. (3 marks)
- Interpret the coefficient on beds. (2 marks)
- Suppose you were looking for a place to stay. You have two options: (1) rent a one bedroom, one bathroom property, (2) rent a two bedroom, one bathroom property and rent out the second room for \$150 per week. Show how you would use the model above to calculate the predicted cost of each and determine which is likely to be the cheaper option. (4 marks)

- (b) Instead of including beds as a linear variable we now include a dummy variable for every value of number of bedrooms. For example, $\text{beds}_2=1$ when the home has 2 bedrooms, and equals zero if it does not have 2 bedrooms.

The estimated model below uses this set of dummy variables, excluding the dummy variable for beds equals 1.

Regression Statistics						
Multiple R	0.848158488					
R Square	0.71937282					
Adjusted R Square	0.709229669					
Standard Error	141.4712527					
Observations	173					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	6	8516649.615	1419441.602	70.92202565	2.91996E-43	
Residual	166	3322343.148	20014.11535			
Total	172	11838992.76				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	478.7795044	45.81743958	10.44972196	5.84371E-20	388.3194869	569.239522
baths	110.0666494	23.65777746	4.652450957	6.67098E-06	63.3577328	156.775566
beds_2	84.26958929	42.75883133	1.970811331	0.050408186	-0.151641309	168.6908199
beds_3	279.4661973	44.33564149	6.303420632	2.52828E-09	191.9317793	367.0006152
beds_4	509.5938565	50.4142887	10.1081235	5.09368E-19	410.0580142	609.1296988
beds_5	694.8538591	66.00104259	10.52792247	3.55271E-20	564.5441882	825.16353
beds_6	901.0871967	148.7055921	6.059538071	8.87464E-09	607.4891565	1194.685237

- Do the coefficients on each of the dummy variables follow the pattern you would expect? Explain. (3 marks)
- With reference to the coefficient on beds_2 , discuss what the p-value and confidence interval are telling us. Are they giving a consistent message? (3 marks)
- Consider again the two options: (1) rent a one bedroom one bathroom property, (2) rent a two bedroom one bathroom property and rent out the second room for \$150 per week. Show how you would calculate the predicted cost of each and determine which is likely to be cheaper using the new model. (3 marks)

- (c) We tried another modification of our regression and made the dependent variable the rent divided by the number of bedrooms. This gave the following results.

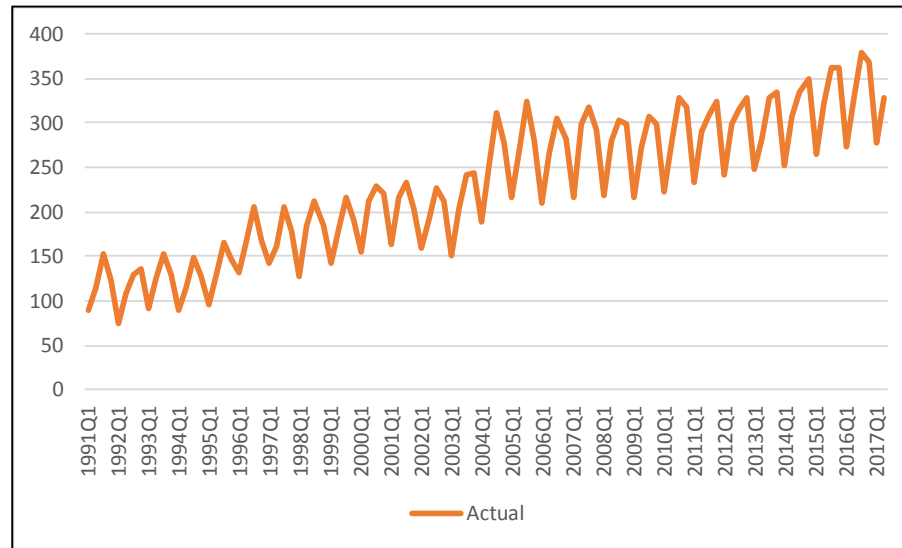
Regression Statistics						
Multiple R	0.586881605					
R Square	0.344430018					
Adjusted R Square	0.33671743					
Standard Error	79.88483928					
Observations	173					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	2	569979.9628	284989.9814	44.65816371	2.58595E-16	
Residual	170	1084869.883	6381.587547			
Total	172	1654849.846				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	453.9520057	19.980227	22.72006248	2.17193E-53	414.5107038	493.3933076
beds	-62.58390327	6.651165861	-9.40946363	3.44856E-17	-75.7134158	-49.45439073
baths	40.49242359	13.32771745	3.038211438	0.002755689	14.18328667	66.80156052

- Is the R-Squared of this model comparable with that in part (a) or (b) of the question? (2 marks)
- What does the coefficient on beds tell us in this model and why do you think the coefficient is negative? (4 marks)
- Once again consider the two options; (1) rent a one bedroom one bathroom property, (2) rent a two bedroom one bathroom property and rent out the second room for \$150 per week. Show how you would calculate the predicted cost of each and determine which is likely to be cheaper using our new model. (3 marks)

Question 4 (19 marks)

(a) We have data which records the number of people coming from New Zealand to Australia for a holiday. The data is recorded quarterly. We are interested in modelling this data. Below is a chart of the data and the results from a regression model on this variable (where arrivals are measured in thousands). The independent variables are:

- time (1 in 1991Q1, 2 in 1991Q2, up to 106 in 2017Q2), and
- dummy variables for each quarter. For example, d_Q1=1 when the observation is for quarter 1, and zero for other quarters.



Regression Statistics						
Multiple R	0.976223819					
R Square	0.953012945					
Adjusted R Square	0.951152072					
Standard Error	17.13672368					
Observations	106					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	4	601585.7441	150396.436	512.1320515	4.22885E-66	
Residual	101	29660.39714	293.6672985			
Total	105	631246.1412				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	123.1619739	4.463843815	27.59101326	7.827E-49	114.3069085	132.0170392
time	2.265162877	0.054404765	41.63537671	2.09549E-65	2.157238458	2.373087296
d_Q1	-61.98227302	4.708972561	-13.16258955	1.22331E-23	-71.32360788	-52.64093816
d_Q2	-16.62891738	4.70865827	-3.531561737	0.000624188	-25.96962877	-7.288205986
d_Q3	16.10362442	4.753183361	3.387966167	0.001004947	6.674587219	25.53266161

- i. Discuss how well you think the model fits the data, considering both Rsquare and Standard Error? (3 marks)
- ii. What is the purpose of the quarterly dummy variables and what are they reflecting? (2 marks)
- iii. What does the coefficient on the intercept mean? (2 marks)
- iv. Interpret the coefficient on the Q2 Dummy variable. (2 marks)
- v. On average, how much higher are arrivals in Q3 than Q2? (2 marks)
- vi. Use the model to show the calculation you would do to forecast the number of arrivals in 2018Q1. (2 marks)
- vii. Illustrate how you would calculate how many people are expected to arrive in the whole of 2018. (3 marks)
- viii. The graph below plots the model errors - the difference between the actual values and those predicted by the model. Discuss the features of this graph. Comment on what they suggest about how the model and forecasts may be improved. (3 marks)

