Question 1. [6 marks]

A shortage of long term care capacity has long hindered our health care system from providing appropriate care to Canadian citizens. One issue is whether the age of entry into long term care is related to the living arrangements of residents before entry. The following data were compiled:

	<60	60-69	70-79	80-89	90-99	>100	Totals
Alone	14	48	140	524	278	12	1016
Non-Private Residence	16	28	98	285	192	8	627
Spouse/Life Partner	22	99	374	783	232	4	1514
With Family	37	32	127	311	209	10	726
Totals							

The data outlines four potential pre-admission living arrangements – alone, non-private residence, living with spouse and living with family. The columns represents the age of residents at time of admission. Using the partially completed output in the Appendix A, perform the appropriate hypothesis test to determine if there is indeed a relationship between the age of entry and the living arrangement pre-admission. Use a 5% significance level.

Ho: age of entry and pre-admission living arrangements are independent
Ha: the two factors are related

Chi-Sq = 178.840, DF = 15, P-Value = 0.000
Critical value: 24.996

Since Chi-Sq > 24.996, we reject the null hypothesis, and conclude that there is a relationship between the age of entry and the pre-admission living arrangement

1 mark for hypotheses
2 marks for test statistic
1 mark for degrees of freedom
1 mark for critical value
1 mark for decision/conclusion

Question 2. [8 marks]

a) Using the same data, determine if the proportion of residents entering long term care before the age of 70 is different for residents living alone prior to admission versus residents living with family. Use a 5% significance level and the critical value approach.

[4]

```
Sample   X     N   Sample p

1       62  1016  0.061024
2       69   726  0.095041

Difference = p (1) - p (2)
Estimate for difference:  -0.0340177
95% CI for difference:  (-0.0599357, -0.00809972)
Test for difference = 0 (vs not = 0):  Z = -2.65  P-Value = 0.008
```

Ho: P1=P2; Ha: P1≠P2, where P1 is the proportion entering long term care before 70, among those living alone, and P2 is the same proportion among those living with family.

Pooled proportion is (62+69)/(1016+726)=0.0752,
SE is sqrt[(0.0752*0.9248)(1/1016 + 1/726)] = 0.0128 and z = .034/.0128 = 2.66

0.5 mark for correct hypotheses,
1 mark for correct sample proportions and pooled proportion calculation
1 mark for correct test statistic,
1 mark for correct p-value or rejection region shown as |Z| > 1.96,
0.5 mark for the conclusion

Many misinterpreted these as the proportions of those under age 70 who lived alone or with family (p1-hat = 62/296 = 0.209 and p2-hat = 69/296= 0.233, which are correlated sample proportions). This leads to a t = 0.61, based on a SE of 0.038643.

b) Calculate a 95% confidence interval for the difference between the proportion of residents entering long term care before the age of 70 among those living alone prior to admission versus the proportion of residents entering before the age of 70 among those living with family. Does this confirm your conclusion in part (a)?

[3]The CI is (0.095 – 0.061) ± 1.96 * 0.0132 = 0.034 ± 0.026 = (0.008, 0.06) or (-.06, -0.008)

1 mark for correct z-value, 1 mark for correct interval based on SE of 0.0132, 1 mark for correct conclusion SE would be 0.0128, if pooled proportion were used.

c) What key assumption justifies the use of the Z distribution in this question?

[1]

Assume that the sample proportions are normally distributed, or that the samples are large enough for the sampling distribution of the proportions to be normally distributed.

Question 3. [4 marks]

Another issue of whether appropriate care can be provided is whether there has been a shift in the age distribution of incoming long term care residents. Some changes (such as increased severe obesity and increased dementia) might lead to earlier entry into long term care while increased life expectancy may mean that people require long term care later. In 2008, the **population** of residents entering long term care had the following distribution:

	<60	60-69	70-79	80-89	90-99	>100
2008	2.96%	5.45%	18.21%	48.71%	23.57%	1.11%

In 2012, a random sample of residents entering long term care were distributed as follows:

	<60	60-69	70-79	80-89	90-99	>100	Total
2012	33	59	234	649	383	17	1375

Perform the appropriate hypothesis test to determine whether there has been an overall shift in the distribution of the age of residents at time of admission into long term care over the 4 year period. Use a 1% significance level.

Ho: p1=0.0296, …, p6=0.0111, Ha: at least one proportion not equal to that specified

(note that the hypothesized probabilities are exact values (not sample estimates) from the 2008 population)

Chi-square statistic = 17.48

Rejection region is $\chi^2 > 15.086$

Reject Ho, Conclude shift in age distribution of residents admitted into long term care

	<60	60-69	70-79	80-89	90-99	>100
Observed	33	59	234	649	383	17
Expected	40.66543	74.97689	250.3466	669.7089	324.0526802	15.24954
Contribution	1.444934	3.404531	1.067363	0.640364	10.72290625	0.200932
					Chi-square	17.48103

1 mark for hypotheses, 1 mark for chi-square statistic, 1 mark for critical value, 1 mark for decision/conclusion

Question 4. [10 marks]

Appendix B shows the average salaries for male and female assistant professors in the Faculty of Arts at 22 different universities.

(a) Explain whether the two samples are independent or paired.

[1] Paired -- each pair of salaries is from the same university *(note that correlation between the two samples is 0.972)*

(b) With specific reference to the appropriate boxplot(s), explain whether a parametric or non-parametric test is more appropriate to test for a difference in average salaries (mean or median).

[2] For a paired test, only the distribution of the differences is relevant. The boxplot of differences is skewed, suggesting the population of differences are not normally distributed. Therefore the non-parametric test may be more appropriate. Accept answer that the boxplot is mildly skewed, but it is reasonable to assume normally distributed errors (the upper whisker is longer due to one difference of 2.5). If answer to (a) is independent, look for comments on separate boxplots that the normality assumptions are valid.

(c) Without regard to your answer in (b), perform the appropriate non-parametric test.

[3] Ho: theta 1 = theta 2; Ha: theta 1 $\neq$ theta 2 *(theta = population median)*

p-value for Wilcoxon test is .339 not < 0.05 (p-value for MW test is 0.73)

Do not reject Ho, cannot conclude median salaries are different

(d) Without regard to your answer in (b), perform the appropriate t-test.

[3] Ho: $\mu 1 = \mu 2$; Ha: $\mu 1 \neq \mu 2$

T = .232 / (.846 / sqrt(22)) = .232 / 0.18 = 1.29

(or t = .232/(3.5738*sqrt(2/22)) = 0.215 for 2-sample test)

Rejection region is | t | > 2.08 (df = 21) For 2-sample test, |t| > 1.96 or 2.0.

Do not reject Ho, there is insufficient evidence of difference in average salaries

(e) Explain how the paired t-test results compare with the two-way analysis of variance results.

[1] The F-statistic for testing the effect of gender is 1.65, which is the square of t = 1.29.

The tests are equivalent (F* = 4.35, based on 1, 20 d.f. is the square of t* of 2.08, df=21)

Question 5. [15 marks]

A consumer organization wanted to examine the effect of age and gender on the amount (in hundreds of dollars) used car dealerships were willing to pay for a used car. Twelve "owners" (six male and six female) were recruited from each of three age groups (young, middle and elderly). Each "owner" was randomly assigned to one of thirty-six used car dealers and brought the same car to solicit a cash offer. The data and analyses can be found in Appendix C.

(a) Examine the residual plot and explain whether the model assumptions are warranted.

[2] Since the standard error is 1.55, only one of the residuals is beyond 2 standard errors, suggesting the errors follow a normal distribution.

The vertical spread is relatively constant, suggesting constant variance of errors.

(b) Describe the effects of age and gender on the average offer. Explain why you do not suspect any interaction between age and gender.

[3] -Middle age "owners" seem to get higher offers. -Gender does not seem to have an effect. -Even though the lines are not exactly parallel, they are close to being so, suggesting no interaction.

(c) Test whether the effect of age depends on gender.

[3] Ho: no interaction; Ha: some interaction between age and gender

F = 1.06 Rejection region if F > 3.32 (2,30 df)

Do not reject Ho, cannot conclude there is interaction

(d) Now test the effect of age on the average offer.

[3] Ho: no main effect due to age; Ha: some age effect

F = 66 Rejection region if F > 3.32 (2,30 df)

Reject Ho, conclude age affects the offer

(e) Using the Bonferroni method, can you conclude that the middle age group received higher cash offers than either of the other two age groups?

[4] K = 3 age groups, J = 3. $\alpha/(2J) = .05/6 = 0.00833$ $t^* = 2.55$ (based on prob'y=0.008)

Margin of error is 2.55 * 1.546 * sqrt(1/12 + 1/12) = 2.55 * 0.631 = 1.61

Average offer for Middle is 27.75, for Senior it is 21.42, for Young it is 21.5

Clearly Middle is more than 1.61 higher than the other two.

Question 6. [22 marks]

To model the Canadian/U.S. dollar exchange rate ('ExchangeRt'), stock indices for various sectors of the economy are used as predictors. Some of the variables used in the four models shown in Appendix D are:

- ExchangeRt: The Canadian dollar equivalent to the U.S. dollar.
- Energy: Index of energy stocks, including oil and gas.
- Financial: Index of Banks and other financial institutions
- Industrial: Industrial Index
- Materials: Index of Lumber, Potash and other materials
- M&M: Index for Metals and Mining sector
- Utilities: Index of for power production sector
- IT: Information technology sector index

Simple Linear Regression [4 marks]

(a) What problem(s) can you identify in the residual plots?

[1] The Standardized Residuals versus Order graph shows a pattern of a train of positive and negative residuals indicating:

i. Lack of randomness in the residuals. {1/2 mark}

ii. Positive 'autocorrelation' among the residuals. {1/2 mark}

(b) What is the simple linear regression equation which relates 'ExchangeRt' to the predictor variable 'Energy'?

[1] The regression equation is ExchangeRt = 1.43 - 0.00109 Energy {1 mark}

(c) Explain why the negative coefficient of the 'Energy' variable makes economic sense.

A large part of the Canadian economy is resource based. When the oil and gas industry is thriving with large exports, its stock index is high. Large exports of oil and gas bring in large amounts of US dollars into the country, thereby relatively weakening the American dollar and consequently strengthening the Canadian dollar. This causes the exchange rate (of Canadian dollar versus the American dollar) to become lower; that is why the negative coefficient makes sense. {Must mention **exports, influx of US dollars** etc}

[1]

(d) Calculate the correlation coefficient between 'ExchangeRt' and 'Energy'.

[1] R^2 = SSR/SSTotal = 0.14275/0.19187 = 0.7440 {1/2 mark}

Correlation coefficient, r = SQRT(0.7440) x sgn(b1) = 0.8626 x -1 = -0.8626{1/2 mark}

{'r' has the same sign as 'b1' in the SLR}

Comparison of models [3 marks]

(e) Between Model 2 and Model 3, which is the better model? Explain briefly.

[2]

Model 2 is better because:
-the standard error is lower (or, equivalently, the R-square (adj) is higher)
- it is a simpler model
- the Cp statistic is lower (meaning the total squared prediction errors are lower);

1 mark each for any two of the three points above

(f) The high Variance Inflation Factors suggest that there may be a problem that affects the estimated model coefficients. Describe this effect.

[1]

-the coefficients tend to have high standard errors (meaning the coefficients are highly variable from sample to sample) or, equivalently, the t-statistics are lower than expected.

The following questions relate only to Model 4 [15 marks]

(g) Explain if the linear model assumptions are warranted or justified.

[2] The assumptions are that the residuals are: i. reasonably normally distributed: the residuals fall on the normal probability plot line and are reasonably normal. ii. the residuals must have constant variance: residuals versus the fitted values plot shows no pattern suggesting constant variance. iii. Residuals should be random without any autocorrelation: residuals versus order plot shows randomness. Thus the assumptions are justified. {1 each for at least 2 of the above facts}

(h) How would you improve the regression equation in Model 4

[1] Unusual Observation #22 with a standardized residual value of -2.59 can be dropped. Then this outlier will not affect the quality of the regression model.

(i) Test whether this model is useful. Use the 5% level of significance.

S1: Ho: $\beta 1 = 0, \beta 2 = 0$; Ha: at least one of $\beta 1$ or $\beta 2$ is nonzero

[3] S2: $F_{Calc} = MSR/MSE = 241.99$

S3: $F_{Crit} = F_{0.05}(df_R = 2, df_E = 21) = 3.49$ (based on 2, 20 df)

S4: Since $\{F_{Calc} = 241.99\} > \{F_{Crit} = 3.49\}$ ==> Reject Ho.

There is sufficient evidence to suggest that the model is useful.

{S1, S3, S4 and the managerial statement 1/2 mark each; S2, 1 mark}

(j) Test at the 1% level of significance whether the 'Financials' variable is important to the model. State your conclusion carefully.

S1: Ho: $\beta_1 = 0$, Ha: $\beta_1 \neq 0$

S2: $t_{Calc} = (b_1 - 0)/SE(b_1) = (-0.002257/0.0002247) = -10.05$

[3] S3: $t_{Crit} = t_{\alpha/2}(dfE) = t_{0.005}(21) = 2.83$

S4: Since {|tCalc| = 10.05} > {tCrit = 2.83}==> Reject H0.

There is sufficient evidence to suggest that the 'Financials' variable is useful, given the the Materials index is already in the model.

{S1, S3, S4 and the managerial statement 1/2 mark each; S2, 1 mark}

(k) How do you interpret the value of the estimated coefficient of the 'Financials' variable? Be as precise as possible.

[2] [1] If the 'Financials' index increases by 1 unit, then the 'ExchangeRt' is estimated to go down by 0.0022571 and the Canadian dollar will strengthen against the American dollar, assuming the 'Materials' index is unchanged.

{1 mark for the 'If' clauses and 1 mark for the 'then' clause.}

(l) Using Model 4, calculate a 99% interval prediction for 'ExchangeRt' when the 'Financials' index is 190 and the 'Materials' index is 400.

[2] Prediction Interval: $\text{Y-Hat}_v \pm t_{\alpha/2}(df_E)\ SE(Y_v)$

$\{SE[Y_v]\}^2 = \{s_e^2 + SE[\mu\text{-hat}_v]^2\} = (0.01949)^2 + (0.00768)^2 = 0.000439$ {1 mark}

$SE[Y_v] = SQRT(0.000439) = 0.02095$, and $t_{\alpha/2}(df_E) = t_{0.025}(21) = 2.83$ and $\text{Y-Hat}_v = 0.99361$

PI: $0.99361 \pm 2.83 \times 0.02095 = 0.99361 \pm 0.0566$

PI: (0.94, 1.05) {1 mark}

{If the 'Financials' index is 190 and the 'Materials' index 400, the specific exchange rate at that time could be anywhere from US $ 0.94 to US $ 1.05}.

(m) If 'ExchangeRt' is regressed on the residuals from Model 4, the R-square is 0.0416. How is this related to the R-square of Model 4? Explain the relationship.

[2]

$R^2_{Model4} = 1 - R^2_{Resids4}$

$= 1 - 0.0416$

$= 0.9584$ {1 mark}

This relationship holds because R^2_{Model4} of 0.9584 means that 95.84% of the variation in the 'ExchangeRt' or 'Y' data is explained by the regression with the two predictor variables, leaving 4.16% of the variation unexplained (or "explained" by randomness). Obviously, this 4.16% variation in the 'Y' data is precisely what the regression with residuals explains with $R^2_{Resids4}$. {1 mark}