Question 1. [17 marks]

a. Identify the appropriate parametric test and state whether you think the assumptions required are warranted based on the boxplots in Appendix A.

[2]

The appropriate test would be a two sample t-test. The boxplot for high school graduates looks quite symmetrical. Though the boxplot for undergraduate+ is a bit skewed, the sample size is 44 so the skew is not as big an issue. Thus, I would suggest that the assumptions are valid.

1 mark for appropriate test, 1 mark for correct explanation.

b. Identify the appropriate non-parametric test and explain why it would be a valid test even if the assumptions in (a) were not warranted.

[2]

The appropriate non-parametric test would be the Mann-Whitney U test. It is still valid even if the normality assumption is unwarranted since it compares the medians rather than the means and thus is not as affected by skewed data.

1 mark for correct test, 1 mark for correct explanation.

c. Perform the parametric test you chose in (a) to answer the hypothesis posed at the beginning of the question. Use a significance level of 0.05. State any reasonable assumption you need to make.

[5]

Assume equal variances between the two populations.

$H_0: \mu_1 - \mu_2 = 0$ $H_a: \mu_1 - \mu_2 < 0$ or $\mu_2 - \mu_1 > 0$

```
Two-sample T for High-school vs Undergrad+

              N  Mean  StDev  SE Mean
High-school  15  49.1   13.2      3.4
Undergrad+   44  75.5   15.1      2.3

Difference = mu (High-school) - mu (Undergrad+)
Estimate for difference:  -26.4788
95% upper bound for difference:  -19.1609
T-Test of difference = 0 (vs <): T-Value = -6.05  P-Value = 0.000  DF = 57
Both use Pooled StDev = 14.6382
```

Assuming equal variances, the critical value is -1.67 or -1.645. Since the t-stat is much smaller than our critical value, we reject the null and conclude that those with a university degree do in fact earn more post-graduation.

1 mark for hypotheses, 1 mark for test statistic, 1 mark for pooled variance = 214.28 and rejection region < critical value of -1.67 based on 57 d.f. , 1 mark for assumption, and 1 mark for decision and conclusion.

Assuming unequal variances, we have:

```
Difference = mu (High-school) - mu (Undergrad+)
Estimate for difference:  -26.4788
95% upper bound for difference:  -19.5156
T-Test of difference = 0 (vs <): T-Value = -6.48  P-Value = 0.000  DF = 27
```

If they assume unequal variance, then have to calculate d.f. = 27 for critical value of -1.7. This d.f. calculation gets the 1 mark instead of the 1 mark for the pooled variance.

d. The researchers belatedly realized that they would like to analyse the data to determine whether having a graduate degree further enhances your chances of earning a higher salary. They thus divided their sample of undergraduate degree executives into those that had only an undergraduate degree and those that had a graduate degree as well. The sample mean for the 24 executives who only had an undergraduate degree (the rest had a graduate degree) was 72.71 and the sample standard *deviation* was 14.95 while for those who also had a graduate degree, the sample mean was 78.95 and the sample standard *deviation* was 14.91. They performed the appropriate analysis *on all three sub-samples*. Your task is to fill in the rest of the ANOVA table below, but you should show how the MSE = 211 is calculated for full marks.

e.

Start by summarizing the data:

Highest level	*High school only*	*Undergraduate degree*	*Graduate degree*
Mean	49.07	72.71	78.95
Standard deviation	13.16	14.95	14.91
Sample size	15	24	20

MSE = 210.52 is calculated by (14*13.16^2 +23*14..95^2 + 19*14.91^2)/56
Degrees of freedom for the error term is 56. Total degrees of freedom is 58.
SSE = 210.52 * 56 = 11789
MST = 4134, SSE=MSE*56=11789,SSTotal=SST+SSE=20057

	Degrees of freedom	Sum of Squares	Mean Square	F
(Education) Factor	2	8268	4134	19.64
Error	56	11789	211	
Total	58	20057		

1.5 mark for showing calculation of MSE, 1 mark for the degrees of freedom, 1 mark for the SS values (lose the mark if any of the SS are wrong) and .5 mark for MST and F.

f. Use a 95% Bonferroni confidence interval to determine whether there is a statistically significant difference between those with graduate degrees and those with only an undergraduate degree. To help you out, the necessary t critical value is 2.47, but you must describe how this would be derived.

[4]

78.95 – 72.71 +/- 2.47 * sqrt[211*(1/20+1/24)] = 6.24 +/- 10.85 = (-17.1, 4.6)

Thus, since zero is in the interval, there is no evidence of a difference in income between those with or without a graduate degree (provided they have an undergraduate degree).
1 mark for the correct difference in means, 1 mark for the correct upper and lower bounds and 1 mark for the correct conclusion.
*T = 2.47 is based on a tail probability of .05/(2*3)=.0083 and 56 d.f. (1 mark total here) since there are J=3 possible pairwise comparisons.*

Question 2. [13 marks]

A small independent stock broker has created four sector portfolios for her clients. Each portfolio always has five stocks that may change from year to year. The volatility of each stock is recorded for each year. Please refer to Appendix B.

a. Looking at the relevant graph, comment on the validity of the model assumptions. [2]

While there are a few outliers, they aren't sufficient in number to make the normality assumption unwarranted and though the spread is somewhat uneven, it too is reasonably consistent to allow the equal variance assumption to stand as well.
Note: I would accept arguments that the equal variance assumption is suspect.
1 mark for stating the reasons for the normality assumption and 1 mark for stating the reasons for the equal variance assumption.

b. Based on the partial analysis in Appendix B, perform the appropriate hypothesis test to determine if the impact of the type of portfolio on volatility varies depending on the year. Use a significance level of 0.05. [4]

H_0: There is no interaction H_a: There is interaction

To get the test statistic, we first calculate $SSAB = SSTotal - (SSA + SSB + SSE) = 265.2$

$$MSAB = \frac{265.2}{6} = 44.2$$

$$F = \frac{MSAB}{MSE} = 4.96$$

The critical value is 2.29 so we reject the null and conclude that there is interaction between the two factors.
0.5 mark for hypothesis, 0.5 mark for SSAB, 0.5 mark for MSAB, 0.5 mark for F, 1 mark for critical value, 1 mark for decision and conclusion.

c. Are the main effects test in this case relevant? Why or why not? [1] Two answers are accepted:

-Not relevant as the presence of significant interaction makes them difficult to interpret since the effect of one factor depends on the level of the other.
- while the interactions are significantly different from zero, they are not that serious as to say that the main effects are not meaningful.
1 mark for either answer.

d. Look at the interaction plot given in Appendix B. Does it confirm or contradict your conclusion in (a)? Why is the hypothesis test performed in (a) still necessary even if one has the interaction plot? [2]

The interaction plot lines are not parallel to each other suggesting the presence of interaction which is what we showed in part (a). We still need the test however as the interaction plot is based solely on sample data.
1 mark for correctly pointing out that the interaction plot suggests that there is interaction, 1 mark for correctly recognizing that an interaction plot is based on sample data.

e. Use a Bonferroni confidence interval to determine whether there is a difference between the mean volatility of Leisure portfolio in 2004 and the Retail portfolio in 2006. The required t-value is 3.597. How many pairwise comparisons are possible using the Bonferroni margin of error?

[4] Based on 12 cell means, there are 12*11/2 = 66 possible pairwise comparisons
The two means are 18.52 (for 2004 Leisure portfolio) and 19.9 (for 2006 retail portfolio). We therefore get:

19.9-18.52 +/- 3.597 sqrt(8.92*2/5) = 1.38 +/- 6.79 = (-5.41, 8.17)

Therefore, since zero is in the interval, we conclude that there is insufficient evidence of a difference between the 2004 leisure portfolio and the 2006 retail portfolio.
1 mark for correct formulation of the interval, 1 mark for correct calculations, 1 mark for conclusion. 1 mark for number 66.

Question 3 [15 marks]

As of April 18, 2011, the three most recent political polls taken around April 14 give the following Ontario breakdown in advance of the May 2, 2011 federal election:

	Cons	Green	Lib	NDP	undecided	Total
Nanos	121	8	104	48	17	298
Forum	289	52	237	155	48	781
Ekos	195	45	185	70	23	518

A chi-square analysis using Minitab yields the following output:

Chi-Square Test: Cons, Green, Lib, NDP, undecided

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

          Cons  Green     Lib     NDP  undecided  Total
 Nanos     121      8     104      48         17    298
       112.89  19.59   98.15   50.94      16.42
        0.582  6.859   0.348   0.170      0.020

 Forum     289     52     237     155         48    781
       295.87  51.35  257.24  133.51      43.04
        0.160  0.008   1.592   3.460      0.573

 Ekos      195     45     185      70         23    518
       196.24  34.06  170.61   88.55      28.54
        0.008  3.516   1.213   3.886      1.077

Total      605    105     526     273         88   1597
```

a. Test at the .01 significance level whether the data reveal any overall differences in how the electorate is currently leaning (or not leaning) among the Ontario populations addressed by the three polling firms. Use the critical value approach. [4]

Ho: Proportions supporting the various parties or undecided are the same from poll to poll, or Electoral choices (or lack thereof) and Poll are independent;
Ha: Proportions supporting the various parties or undecided differ from poll to poll, or Electoral choice and Poll are associated or related.

```
Chi-Sq = 23.472, DF = 8, P-Value = 0.003
```

1 for hypotheses, 1 for the chi-square statistic, 1 for the critical value of 20.09 based on 8 d.f. and 1 for decision and conclusion

b. What is the p-value of the result in part (a)? If an exact value is not feasible, gives the best bounds possible. Which observed count stands out as being the most responsible for the result? Explain briefly. [2]

Based on the chi-square table, the p-value is approx. 0 .003.
Based on the chi-square contribution, the result that stands out are the Greens identified by Nanos.

c. The two largest polls (Forum and Ekos) show 32.3% and 37.4% Liberal support among decided voters (n = 781-48=733 and n = 518-23=495, respectively). Do the data show any statistically significant differences (at the .05 level)? [4]

```
Sample    X    N  Sample p
1       237  733  0.323329
2       185  495  0.373737

Difference = p (1) - p (2)
Estimate for difference:  -0.0504086
95% CI for difference:  (-0.104842, 0.00402501)
Test for difference = 0 (vs not = 0):  Z = -1.82  P-Value = 0.070
```

-Ho: p1=p2; Ha: p1-p2 ≠ 0
Pooled proportion is (237+185)/(733+495) = .347 (I would not insist on this)
-Z = (.374 - .323) / sqrt[.347*.653*(1/733+1/495)] = .0504/.0277 = 1.82
-Since |1.82| is not > 1.96, we do not reject the null H. There is insufficient evidence to say there is a difference.
1 for hypotheses, 1 for standard error of .0277, .5 for z-statistic, .5 for critical value, 1 for decision and conclusion.

Some students tried to do a chi-square test—this is possible as a test of homogeneity of proportions in a 2x2 table:

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

            C1      C2  Total
    1      237     496    733
        251.89  481.11
         0.881   0.461

    2      185     310    495
        170.11  324.89
         1.304   0.683

Total      422     806   1228

Chi-Sq = 3.329, DF = 1, P-Value = 0.068
```

Note that 3.329 is the square of z = ±1.82 and the critical value of 3.84 (df = 1) is the square of 1.96.

d. Seeing that the estimated Liberal support (32.3%) by the Forum is lower than that estimated by Ekos (37.4%), is it fair to do the test above as a 1-sided test? Explain briefly.

[1] *No, the hypotheses must be set before seeing the data. Otherwise, we can set the hypotheses to obtain the conclusion we favour. In fact, a 1-sided test would end up rejecting the null hypothesis.*

e. During the 2008 federal election, 18.2% voted for the NDP in Ontario. During the past few weeks, NDP support has grown nationally. During the Easter weekend, a new poll of 1028 voters in Ontario showed 236 favouring the NDP. Test at the .05 significance level whether this constitutes evidence of a shift toward the NDP in Ontario.

[4]

```
Test of p = 0.182 vs p > 0.182
                                  95%
                                Lower
Sample    X     N  Sample p     Bound  Z-Value  P-Value
1       236  1028  0.229572  0.207997     3.95    0.000
```

Ho: p = .182; Ha: p > .182
Z = (236/1028 - .182)/sqrt(.182*.818/1028) = 3.95
Z is in the rejection region Z > 1.645
Reject Ho, conclude there is a shift toward the NDP in Ontario

1 for hypotheses, 1.5 for calculating Z, .5 for critical value and region, 1 for decision/conclusion.

This can be done also as a goodness of fit test; however, the test becomes 2-sided and there are 2 categories: those supporting the NDP and those not supporting the NDP, and not simply 1 category:

	observed	expected	(O-E)^2	chi-sq cont.
	236	187.096	2391.601	12.78275
	792	840.904	2391.601	2.844084
total	**1028**	1028		**15.62683**

The expected values are .182*1028 and .818*1028, respectively.
The chi-square statistic is 15.62 based on 1 d.f. ; this is the square of z = 3.95
Here the critical value for the test is 3.84 which is the square of the critical value of 1.96 for the 2-sided test.

Question 4. [25 marks]

A medical researcher wanted to establish a relationship between % body fat of male subjects, the response variable, and eight predictor variables:

Fat%: % body fat
Weight: body weight in pounds (lb.)
Height: height in inches
Neck: circumference of neck in cm
Chest: circumference of chest in cm
Abdomen: circumference of abdomen in cm
Hip: circumference around hip in cm
Thigh: circumference of thigh in cm
RegExer: 1 if regular physical exercise is done and '0' otherwise.

a. Carefully look at the three models and specify which model is the best one. Justify your choice on the basis of *three* numerical criteria. You will need to calculate some of these values.

[3]

Model	s or s_e	R^2	R^2_{Adj}	Comment
2	3.6802	0.853	0.830	Best Model
3	3.7781	0.859	0.827	
1	5.2028	0.678	0.672	

Model 2 is the "Best" model because it has the lowest value of 's/s_e' and/or the highest value of R^2_{Adj} . Note that these two criteria are equivalent.
While Model 3 has the highest R-square, it is only marginally higher than the R-sq for Model 2; however, Model 2 is much simpler as it is based on only 5 predictors while Model 3 has 9 predictors.
1 mark for calculating the missing s and R^2 *and 1 mark for arguing Model 2 is best based on the s and/or* R^2_{Adj} *, while the 3rd mark is for the argument based on* R^2 .

b. Based on the output given, explain briefly if this best model satisfies the basic assumptions of the regression model.

[2] If you observe the four plots (I would insist only on the residual plot) you see from:
1. Normal Probability plot and the Histogram that the residuals are reasonably normal.
2. From the Residuals vs Fits, you see that the residuals show no pattern and are within ± 2 Std. Dev (6/3.6802 = 1.63). They are reasonably normal and seem to have a constant variance.
3. From the Residuals vs Order, it can be seen that the residuals are quite random and do not have any discernible 'runs'.
Thus the residuals do seem to be random and normally distributed with constant variance and thus satisfy the basic assumptions underlying this model.

Notwithstanding your answers above, please use **Model 2** for the remaining sub-questions.

c. Test whether the model is useful for predicting % body fat, using a 5% significance level.

[3] S1:

H0: *beta1=0, beta2=0, … betaK=0 vs Ha: at least one beta is nonzero.*

Whether the model is useful or not is an interpretation, not a statement of the hypotheses.

S2: F_{Calc} = MSR/MSE = 689.02/13.54 = 50.87

S3: With LS = α = 0.05, $F_{Crit} = F_{\alpha}(df_R = 5, df_E = 44) \approx 2.43$

S4: Since $\{F_{Calc} = 50.87\} > \{F_{Crit} = 2.43\}$ ➔ Reject H0.

Based on the statistical evidence, the regression model is useful.

1 for hypotheses as stated (.5 for not useful/useful?), 1 for F-stat and critical value, 1 for decision and conclusion.

d. What steps would you take to improve this model? Explain your rationale.

[2] 1. Drop observation# 39. This observation has large leverage suggesting that it may be an outlier unduly affecting the coefficients of the regression model.

2. Drop the predictor variable "Neck" which has a very high p-Val of 0.197, suggesting that it does not explain the %body fat too well. It could/may improve the regression.

3. If possible, get even a larger sample, say of size 100 or more. ***(A larger sample reduces the standard errors of the coefficients and of the CI for the mean of Y, but not the standard error of the regression or MSE)***

One mark for each of the above, to a maximum of 2.

e. Explain if there are any specific problems with multicollinearity (refer to specific numerical evidence).

[2] "Weight" variable has a VIF (Variance Inflation Factor) of 17.658. This is more than 10, which is considered as a nominal threshold for considering the presence of multi-collinearity.

f. Based on this value, calculate the coefficient of determination (R^2 value) of this specific predictor variable when regressed against the other predictor variables. What does the R^2 value suggest?

[2] For "Weight", $VIF_1 = 17.658$

Bur $VIF_1 = 1/(1 - R_1^2)$

$17.658 = 1/(1 - R_1^2)$ ➔ $17.658 - 17.658\ R_1^2 = 1$ ➔ $16.658 = 17.658\ R_1^2$

Solving for R_1^2, gives $R_1^2 = 16.658/17.658$ or $R_1^2 = 0.9434$

This indicates that the (variation in) predictor variable "Weight" is almost completely (94.34%) explained by the other 4 predictor variables. In other words, "Weight" and the other 4 predictor variables have a high degree of multi-collinearity. The "Weight" variables is, in simple vernacular, is a predictor variable which is not "independent enough"!

g. Explain the meaning of the estimated coefficient of the 'RegExer' variable. Try to be as careful and as specific as possible.

[2] If all values of the other 4 predictor variables remain unchanged, then if a large group of male subjects is engaged in regular physical activity, then the *estimated expected value or long term average % body fat* declines by 3.214 units. In other words, if the expected value of %body fat without exercise were to be 18.214%, with regular exercise it would be 15%.

h. Test whether the variable 'RegExer' is helpful to predict % body fat. State your conclusion as carefully as possible.

[3] S1: H0: $\beta_5 = (\beta_{50} = 0)$ Ha: $\beta_5 \neq (\beta_{50} = 0)$
S2: $t_{Calc} = (b_5 - (\beta_{50} = 0))/SE(b_5) = -3.214/1.228 = -2.6173$
S3: With LS = $\alpha = 0.05$, $t_{Crit} = t_{\alpha/2}(df_E = 44)$ ➔ $t_{Crit} = t_{0.025}(44) = 2.015$
S4: Since $\{|t_{Calc}| = 2.6173\} > \{t_{Crit} = 2.015\}$ ➔ Reject H0.
Conclude that RegExer is a useful predictor, even given the other variables.
1 mark for hypotheses, 1 for t-stat and critical value, 1 for decision and conclusion (the dependence on the other variables must be mentioned).

If the 4 predictor variables are 170 lbs, 38 cm, 90 cm, 58 cm respectively and if regular physical exercise is done, then

i. Calculate the point estimate/'fitted value' of 'Fat%'.

[1]
$Y_{Hat} = -39.89 - 0.2097 \times 170 - 0.4618 \times 38 + 0.8685 \times 90 + 0.5647 \times 58 - 3.214 \times 1$
$Y_{Hat} \approx 14.6162$ (= 14.604 from the output)

j. Now calculate a 95% confidence interval for % body fat, given the predictor variable values above. Explain what this interval estimates.

[3] With CC = $1 - \alpha = 0.95$, $\alpha = 0.05$
CI: $Y_{Hat} \pm t_{\alpha/2}(df_E) \times SE(Y_{Hat}) = 14.604 \pm 2.015 \times 1.189$
CI : 14.6162 ± 2.3958 ➔ (12.2082, 16.9998)
With the calculated value of 14.6162, CI : (12.2204, 17.0120)
The expected value of % body fat for a large group of males with these characteristics is covered by the interval given by: (12.2082, 16.9998) with 95% confidence.

k. Modify the interval above to obtain a 95% prediction interval for the % body fat for a specific individual with the given characteristics.

[2] $[SE(Y_v)]^2 = [SE(Y_{Hat\,v})]^2 + s_e^2 = 1.189^2 + 3.6802^2 = 14.9576$
Hence SE(Yv) = 3.8675, and
PI : $Y_{Hat} \pm t_{\alpha/2}(df_E) \times SE(Y_v) = 14.604 \pm 2.015 \times 3.8675$
PI: 14.604 ± 7.7930 ➔ (6.8110, 22.3970)
95% of the time, the ***predicted*** value of % body fat for one specific male with the given characteristics would be covered by (6.8110, 22.3970)