ADM 2304
Assignment 4

Due Date: Wednesday, December 7, 2011 at 11:00 pm

Please remember that all hypothesis tests should include hypotheses, test statistic, p-value or critical value, decision and conclusion.

1. [4 marks] Problem 19.15 on MyStatLab.

2. [26 marks] An educational policy analyst in southwestern Ontario wanted to determine whether there was a relationship between income and the level of education at an aggregate level. Although there were potential problems with interpreting relationships based on aggregate data, she decided to begin with 2006 census data from the census tracts in Kitchener, Ontario.

She collected and used the following data for each of the census tracts:

CensusT:	identifying code for the census tract
MedInc:	Median Earned Income
AvgInc:	Average Earned Income
StdErr:	Standard Error of the Estimated Average Income
P_highsch:	the proportion of adults who graduated from high school
P_trades:	the proportion of adults with qualifications in a trade
P_collcert:	the proportion of adults with a college certificate
P_univdipl:	the proportion of adults with a pre-degree diploma
P_univdegr:	the proportion of adults with a university degree
No.Adults	the number of adults in the census tract

Each proportion represents the number reporting that level of education as their highest.

The data are in the files **Kitchener_IncomeEducation** in Minitab and Excel formats.

Please append only the most relevant computer text output in appendices; these should be minimized as much as possible. Manual calculations are only required if requested.

(a) Plot the average income against the median income. What does the graph suggest about the shape of income distributions? You may consider other plots as well but do not include these in your solutions. Why might it be better to use the median income as the response variable rather than the average income?
[2]

(b) Perform a multiple regression analysis using the five educational variables as predictor variables and the **median income** as the response variable. Summarize the regression equation along with a few summary statistics.
[3]

(c) Graph the standardized residuals against the fitted values and comment on whether the model assumptions are warranted.

[2]

(d) Calculate the fitted values for the regression model Now calculate the correlation coefficient between the fitted values and the median incomes. To what regression summary value is this coefficient most closely related?

[1]

(e) Are there any problems with multicollinearity? Explain briefly.

[1]

(f) Identify the two most extreme residual values from the regression in (b). For these observations, replace the values of the median incomes by an asterisk in the worksheet and re-estimate the model without these observations. Summarize your regression model. Explain whether one should drop these observations from the regression and what are the effects and consequences of dropping them.

[3]

(g) Perform an F-test for the overall usefulness of the model re-estimated in part (f), using the .01 level of significance. What do you conclude?

[3]

(h) Test the marginal usefulness or importance of the **p_univdegr** variable (proportion with a university degree), given the other variables in the model, using the .01 level of significance. Would you conclude that a university degree is beneficial in terms of increasing incomes?

[4]

(i) We think of the coefficient of the **P_univdegr** variable as estimating the average change in the median income when the P_univdegr variable increases by one unit. Is it meaningful to think of this variable as increasing by 1? Explain briefly.

[1]

(j) If you wanted to simplify the model, which variable(s) might you drop? Why?

[1]

(k) Using either stepwise regression or a best subsets approach, find what might be considered the best model to use instead of the original full model. Explain briefly why this model is superior to the original model.

[2]

(l) Use the best model in part (k) to calculate a 99% prediction interval for the **actual** median income in census tract number 106.01. (Click the Options button, and copy and paste the values of the predictor variables in the model for this census tract, making sure that there is a space and nothing else between each value). Now show how the standard error for the prediction interval is calculated (you can use the standard errors in the output). Does the prediction interval cover the actual median income for this census tract? What does your answer imply about how well the model predicts the median income for this census tract?

[3]