Stat-340 – Assignment 1 – 2015 Spring Term

Part 0 - Getting started - nothing to hand-in

1. Install *SAS* on your machine. Visit
   `http://www.stat.sfu.ca/~cschwarz/CourseNotes/HowGetSoftware.html`
   for details.

2. Create a new (test) directory and copy the sample *SAS* program to the new directory as described in
   `http://www.stat.sfu.ca/~cschwarz/CourseNotes/SAS-install/FirstSASProgram.html`.

3. Right-click on the *sample-program.sas* and open with SAS 9.4. Careful that you don't open it with the Enterprise product. This will open *SAS* in the control window environment.

4. Read the section on running programs in the *SAS* Windowing Environment at
   `http://support.sas.com/documentation/cdl/en/basess/58133/HTML/default/viewer.htm#a001310736.htm`

5. Run the sample *SAS* program.

6. Look at the *Log* to see if there are any errors. If you want to rerun a *SAS* program that creates pdf output, you must close the pdf file before re-running *SAS*.

7. Look at the *Results* window. Note that if you are running *SAS* 9.4, the pdf file created may not be displayed in the *Results* window and should open in a separate window. If not, visit
   `http://support.sas.com/kb/51/125.html` for the fix to this problem.

8. Congratulations – you've just run your first *SAS* program.

Part 1 - Breakfast cereals - Easy

Preamble:. All of the assignments in this course are similar. There will be typically three questions ranging from easy to hard. In each part, there are numerous questions posed, asking, for example, what is the purpose of a particular option on a statement. These questions often form the basis for the term tests and exams, so it is a good idea to "answer them", using, for example, the commenting feature of Adobe Reader. It is NOT necessary to hand in your answers to these types of questions.

Many of the answers to the questions can be found in the course text book. Additionally, *Google is your friend*.

I also give hints at the end of the assignment about typical problems that the marker found in previous years. Read these and correct your assignments before handing-in your work - learn from the mistakes of others!

Remember, you will use the Online Assignment Submission system to hand in your code, output from your program, and the write ups. The link is found on the Assignment web page in the course web site.

Introduction: Many people in North America start their morning with breakfast cereals. Every two years the Section on Statistical Graphics of the American Statistical Association sponsors a special exposition where one or more data sets are made available, analyzed by anyone interested and presented in a special poster session at the Annual Meeting. One of the datasets in past years was nutritional information about breakfast cereals.

1. Read the information about the dataset and breakfast cereals from
   `http://www.stat.sfu.ca/~cschwarz/Stat-340/Assignments/Cereal/Cereal-description.pdf`

2. Download the cereal dataset from
   `http://www.stat.sfu.ca/~cschwarz/Stat-340/Assignments/Cereal/cereal.csv`
   and save it to your computer in an appropriate directory.

3. Read the data into *SAS* from the separate file (rather than including it inline as in the sample *SAS* program). Your code to read in the data will look something like:

```
title justify=left   'Schwarz, Carl James 123456'
      justify=center 'Assignment 1 - Part 1;
title2 'Cereal Dataset';

ods pdf file='cereal.pdf'  style=styles.printer;

data cereal;
   infile 'cereal.csv' missover dlm=',' dsd firstobs=2;
   length name $30 manufacturer $30.;
   input name $ manufacturer $ calories ... ;
   attrib calories label='Calories' format=7.0;
   ...
run;
....

ods pdf close;
```

If you are using the *SAS* University Edition, don't forget that you need to set a *&path* variable and your code will look something like:

```
%let path=/folders/myfolder/...;  /* replace the ... by the directory names with the ce

ods pdf file="&path/cereal.pdf"  style=styles.printer;  /* use double quotes */

data cereal;
   infile "&path/cereal.csv" missover dlm=',' dsd firstobs=2; /* use double quotes */
   length name $30 manufacturer $30.;
   input name $ manufacturer $ calories ... ;
   attrib calories label='Calories' format=7.0;
   ...
run;
....

ods pdf close;
```

You need double quotes (") around the filenames so that *SAS* can resolve the *&path* macro variable (i.e. replace the *&path* by the value in the *%let* statement.

Refer to `http://people.stat.sfu.ca/~cschwarz/CourseNotes/SAS-install/FirstSASProgram.html` for instructions on running the program using the *SAS* University Edition.

How do you read data "inline" rather than from an external file?
Why should you (almost) always use the MISSOVER option on the infile statement?
What does the *dlm* option on the *infile* statement do?

What does the *dsd* option on the *infile* statement do? This options should almost always be used when the the *dlm* option is specified.
What does the *firstobs=2* option on the *infile* statement do? Why is this needed?
Why should you always specify the (maximum) length of the character variables?
What does the *ODS PDF ... ODS PDF CLOSE;* statements do?
What will *ODS RTF ... ODS RTF CLOSE;* do? This will be helpful when doing your writeup.

4. Print out the first 10 records to verify that you have read the data properly. Do you notice some unusual values for Potassium or Serving Size? What do these represent? The same convention might be used for the other variables as well.

   Why should you always print out the first few (typically 10 or so) records from a dataset?

5. Recode some data values as appropriate to deal with missing values. Do this in a second datastep after the input data step above. Your code will look something like:

```
data new_cereal;
   set cereal;  /* access the original dataset */
   if calories = **** then calories = ****;
   if potass   = **** then potters = ****;
   ...
run;
```

   Of course, you need to change the **** above to appropriate values. It would be good practice to do the recoding for ALL variables, and not just the ones in this particular dataset that appear to have problematic values.

6. Again list the first 10 records to verify that your fixups previously have worked.

7. Use the *SGscatter* procedure to create a scatter plot matrix (also known as a casement plot) involving calories, protein, fat, complex carbohydrates, sugars, sodium and vitamins. Interpret the plot. What do you conclude from these plots. Why do plots involving vitamins look "odd"? What is "odd" about plots involving protein and fat (and other variables)?

8. Use the *SGplot* procedure to create a scatter plot of calories per serving vs. the number of grams of fat.

   - Create a base scatter plot using the *scatter* statement.
   - How many points appear to be plotted on the plot? How many observations are in the dataset? Why is there a discrepancy?

- In many cases, it is useful to *jitter* data values during the plotting process to avoid having data points with the same x- and y-coordinates overplotting. This is easily done in the *SGplot* procedures using the *jitter* option on the *scatter* statement. Redo the plot. Note that the points are jittered ONLY for plotting purposes, the actual data values are left unchanged.
- Add a regression line to the base scatter plot using the *reg* statement. Note that the regression line is computed using the original data points and NOT the jittered values. What is the general form of the relationship between the two variables. The actual regression is not shown on the plot – we will estimate the slope and intercept next.
- Don't forget to properly label the axes using the *xaxis* and *yaxis* statement. If is also good plotting practice to avoid having data points plotted on the margins of the plot. The *offsetmin* and *offsetmax* option can help you add some white space to the margins of the plots.

9. Do a regression of calories against grams of fat. If you find that not all of the diagnostic plots show in the pdf file, change the page orientation to *portrait* rather than *landscape*.
   - What is the equation of the fitted line. Interpret the coefficients.
   - Interpret the standard error of the estimate for the coefficient associated with grams of fat.
   - Use an appropriate option on the model statement to get confidence limits for the slope (and intercept). Use the *SAS* online help feature to find the name of the option. Interpret this interval.
   - What is the value of R^2. Interpret this value?
   - Have a look at the distribution of the residuals. Are you concerned? Why? You may wish to read `http://stats.stackexchange.com/questions/76226/interpreting-the-residuals-vs-fitted-values-plot-for-verifying-the` for information on interpretting residual plots.
   - Look at the plot of observed vs. predicted. What should this plot look like for a well fitting model? Are you concerned? Anything unusual about this plot?
   - Look at the QQ plot for the RESIDUALS. What should this look like if the assumptions of regression are satisfied? Are you concerned? What explains the unusual feature of the plot? You may wish to read `http://en.wikipedia.org/wiki/Q¿Q_plot` on interpretting QQ plots.
   - Examine the plot that shows the confidence intervals for the mean response and the prediction intervals for individual responses. Do you understand the difference? When would want to use either interval?

Estimate the number of calories per serving and the confidence and prediction intervals when there are 4 grams of fat from the plot.[1]

- Review the assumptions for linear regression available in the appropriate chapter at `http://www.stat.sfu.ca/~cschwarz/CourseNotes`. Are the assumptions satisfied for this analysis?

Hand in the following using the electronic assignment submission system:

- Your SAS code that did the above analysis.
- A PDF file containing all of your SAS output.
- A one page (maximum) PDF file that is spaces at least 1.5 lines apart (preferably double spaced) with 2.5 cm margins all around, containing a short write up on this analysis explaining the results of this analysis suitable for a manager who has had one course in statistics. You should include the following:
  - A (very) brief description of the dataset.
  - The fitted regression line. Interpret the slope, se, and confidence interval for the slope.
  - An estimate of the calories per serving for a cereal with 4 grams of fat along with measure(s) of uncertainty. You must carefully explain how to interpret these intervals and carefully explain the difference between the two measures of uncertainty.
  - A plot of the fitted regression line on the scatter plot of calories vs grams of fat.

  You will likely find it easiest to do the write up in a word processor and then print the result to a PDF file for submission. Pay careful attention to things like number of decimal places reported and don't just dump computer output into the report.

A sample write up showing the format is available on the Assignment section of the course web site.

[1] More advanced users of *SAS* can get the predictions computed directly from *Proc Reg* but that is NOT necessary for this assignments. Consult the assignments solutions for details.

Part 2 - Titanic Data - Intermediate

Many people are familiar with the *Titanic* disaster. Data on the passenger aboard the liner are available at
`http://www.statsci.org/data/general/titanic.html`.

The data have already been cleaned up, so you don't have to do any data screening.

1. Read in the dataset directly from the web using the URL option on the infile statement. Hint: Your INFILE statement will look something like:

   ```
   infile 'http://.......' url .......
   ```

   If you are using the *SAS* University Edition, don't forget to set the *&path* variable and use it in the *ODS PDF file="..."* statement. Because the *infile* statement is pointing to a URL, you don't need a *&path* variable here.

   Because the dataset is tab delimited, use a DLM='09'x on the INFILE statement (read `http://www.ciser.cornell.edu/FAQ/SAS/other_delimiters.shtml`).

   Define the name variable to have a length of at least 100 characters. Notice that some ages are given as NA to indicate a missing value. What does SAS do when it encounters such values (hint: read the log).

2. Print out the first 10 records to ensure that you've read the data properly.

3. Use *Proc Tabulate* to create a summary table for the number of passengers that lived and died as a function of passenger class and sex.

   Hint: Your *Proc Tabulate* will look something like;

   ```
   proc tabulate data=blah missing;
      title2 give a nice title (do not forget the semicolon);
      class var1 var2 var3;  /* classification variables */
      table var1*var2, var3;  /* table of var1*var2 along the left and var3 along the top
   run;
   ```

4. Because survival is coded using 0 = died and 1 = survived, you can compute the proportion that survived by taking the mean of this variable (why?).[2] Use *Proc Tabulate* to compute the proportion that survived as a

[2] Hint: Consider the set of values 1 0 0 0. What proportion of the values have the value of 1. What is the mean of the set of 4 numbers?

function of passenger class and sex. Hint: Put the survival variable as an analysis variable rather than a classificiation variable in the *Proc Tabulate*.

Hint: Your *Proc Tabulate* will look something like;

```
proc tabulate data=blah missing;
   title2 give a nice title (do not forget the semicolon);
   class var1 var2;     /* classification variables */
   var    var3;        /* analysis variable */
   table var1*var2, var3*mean*f=7.3;  /* table of var1*var2 along the left and var3 alo
run;
```

What does the '*f=7.3' do in the *Proc Tabulate*. Try different values, e.g. 10.4, 10.1, etc to see the effect.

5. There are many ways to compare proportions between groups, e.g. is the proportion that survived the same for male and females. You could compute the individual proportions and look at the difference in proportions between the two groups, in an analogous fashion to estimating means in each groups and looking at the difference in the means.

   However, for reasons beyond the scope of this course, it is more common to compare proportions by using odds and odds-ratios. Read the chapter on Logisitic Regresion in `http://www.stat.sfu.ca/~cschwarz/CourseNotes`ğ for more details.. Sort the data by passenger class. Hint: Use the *Sort* statement.

   Look at the output from *Proc Tabulate* above. Compute the following (by hand):

   - The probability of survival for males and female for each passenger class.
   - The odds of survival for males and females for each passenger class.
   - The odds ratio of survival (females:males) for each passenger class. If the survival probability was equal for males and females, what should the odds ratio be? If females had a higher survival probability than males, what values would the odds ratio take? If females had a lower survival probability than males, what values would the odds ratio take?

   We will now get *SAS* to do these computations. Use *Proc Freq* to do a comparison of the proportion survival between males and females by passenger class. Hint: Use a BY statement to get the separate analyses and use the RELRISK option on the Tables statement to get the odds ratio of survival for the two sexes. You need to figure out if the odds ratio is referring to survival or death and if the odds ratio is males:females or females:males. [3]

[3] There is a certain symmetry to odds ratios. For example compare the odds ratio of survival (male:female) to the odds ratio of death (females:males).

Look at your contingency tables and turn off non-relevant percentages (e.g. depending on your tables structure, either the rows or column percents are of most interest). You may have to try different arrangements of the row and column variables to get the odds ratio that you want.

Use the ODS system to extract the odds ratios (and confidence limits on same) to a separate dataset. The ODS table name is *RelativeRisks* and you need a code fragment similar to:

```
proc freq data=****;
   by pclass;
   table sex*survived / relrisk;
   ods output RelativeRisks=myoddsratio;
run;
proc print data=myoddsratio;
   title2 'Odds ratio of survival for males vs females by passenger class';
run;
```

6. Use *Proc SGplot* to plot the odds ratio of male vs. female survival for the three classess along with the confidence limits for the odds ratio for each class. Use the HIGHLOW statement in SGplot to get the plots. Be sure to use the XAXIS and YAXIS statements to improve the appearance of the plots and to apply proper labels.

   Hint: Your code will look something like:

```
proc sgplot data=RelRisks noautolegend;
    title3 'Comparison of ODDS ratio (and 95% confidence intervals) among passenger cla
   where studytype=....';
   scatter y=....  x=.....;
   highlow x=.... low=.... high=....;
   yaxis label="Odds ratio of SURVIVAL (male:female)";
   xaxis offsetmin=0.05 offsetmax=0.05 label='Passenger Class';
```

   Note that you will need to use the *Where* clause within *Proc SGplot* to select the observations from *myoddsratio* created earlier as more information than is required is present in the dataset.

   What does the 'offsetmin' and 'offsetmax' do in the *x-axis* statement?

   What does this plot tell you?

Hand in the following using the online submission system:

- Your SAS code.
- A PDF file containing the the output from your SAS program.

- A one page (maximum) 1.5 spaced (but preferably double spaced) PDF file containing a short write up on this analysis suitable for a manager who has had one course in statistics. You should include:
  - A (very) brief description of the dataset.
  - The plot you created of the odds-ratio of male vs. female survival for the three passenger classes. Use ODS RTF file=****; and ODS RTF close; around the SGplot above to also send the plot to an RTF file that you an then easily open in MSWord (and other word processors).
  - An explanation of what is measured by the odds ratio and what the plot shows you about the odds-ratio across the three passenger classes.
  - There is an old saying "Women and children first" (see `http://en.wikipedia.org/wiki/Women_and_children_first`). Does this seem to apply in the case of males and females? How do you tell?

Part 3 - American Time of Use Study - Hard

The American Time Use Survey (ATUS, `http://www.bls.gov/tus/`) is a large scale survey conducted in the United States to measure how people spend their time. This data is useful to measure how much time is spent, for example, on child or elder care, or how people use their leisure time.

The ATUS uses a complex sampling method to select participants to include in the study. All ATUS respondents are assigned a day on which they are to record their activities and the respondent is then called on the following day. A computer-assisted telephone interviewing is used to collect the information on their activities for the sampled day.

The core time diary of the ATUS is very similar to other time-budget surveys. The respondent is asked to take the interviewerthrough his or her day from 4 AM through 4 AM of the following day (the interview day). The respondent describes eachactivity, which the interviewer either records verbatim or, for a limited set of commonly performed activities (such as sleepingor watching television), hits a precode button.

Only the respondent's primary activity is recorded and coded; if the respondent mentions secondary activities performedsimultaneously, these are recorded but are not included in the total time inputs and are not classified using the three-tierscheme. For example, if a person was both watching television and knitting, then only the time for watching television is recorded.

Person-level Variables Included in the ATUS include

- Labour Force Status
- Income
- Earnings
- Gender
- Race
- Marital Status
- Age
- Region
- State
- Household Demographics

- Height
- Weight
- Body Mass Index
- Education
- And, of course – lots of time use variables!

Some research questions of interest are:

- What effect does the economy have on the amount of time spent watching TV and playing video games?
- Does this vary by gender?
- Does this vary according to your labour force participation?
- Does this vary across income?
- What are the strongest sociodemographic predictors of time spent watching TV?
- What activities have been replaced by increased time spent on TV and on video games?

ATS microdata (i.e. a subset of the respondents with suitable anonymization to preserve confidential information) from 2003 to 2012 is available from the U.S. Bureau of Labor Statistics at`http://www.bls.gov/tus/`. I've downloaded a subset of the (large) amount of information (see below for details).

For many students, the only exposure they get to real data analysis is through toy-examples taken from text books. However, most real jobs don't deal with simple toy-examples – real life data is challenging, large, and messy. Research questions are ambiguous and answers not simple. Your mission, should you decide to accept it[4] is to use the ATUS study to see how people spent their time.

FYI, there is a similar study done in Canada (and many other countries), but micro data is not readily available. Findings from the US are expected to be similar to those from Canada.

1. Visit `http://www.stat.sfu.ca/~cschwarz/Stat-340/Assignments/TUS`. Download and store on your computer the ATUS summary data file (*atussum_0313.csv*), a *SAS* file that will serve as a template to read the

[4] `http://en.wikiquote.org/wiki/Mission:_Impossible`

summary data file (*atussum_0313.sas*), and the user's guide (*atususers-guide.pdf*.

2. Read the following sections from the User's Guide:

   - 1.1 What do the ATUS data measure?
   - 1.2 How can the survey results be used?
   - 3.1, 3.2 and 3.3 An overview of the survey design. Unlike many toy-examples in textbooks, this data is collected using a complex sampling scheme – you will learn the theory of these more complex designs in Stat-410 (Survey Sampling). This has consequences in the analysis of the data set.

     What is the population of interest? Who is excluded? There are the obvious exclusions (e.g. young children), but who else is not considered to belong to the population of interest? Why is it important to understand the exclusions?[5]

     The frame is the set of people from the population from which samples are taken. We hope that the frame covers the population of interest – does it?

     The sampling scheme is NOT a simple random sample as often found in toy-examples. For example, certain ethnicities are oversampled relative to their proportion in the population; households with children are oversampled relative to their proportion in the population; (and as you will see later, weekend days are over sampled) relative to their proportion of the week. This will have consequences in the analysis.

     A key task of a statistician (and how you differ from a trained monkey that runs statistical packages) is that you understand the interplay between the sampling design, analysis and conclusions.
   - 3.6 Response rate. Not every contacted person responded. There are three types of missing data (missing completely at random (MCAR), missing at random (MAR), and informative missing (IM) - see my course notes on the web). The type of missingness affects how the data need to be analyzed. For simplicity, we will assume MCAR for the remainder of the course.
   - 6.2 Data imputation. The data has missing values, but these are "filled-in" using various imputation methods. This is beyond the scope of this course, and we will ignore any problems induced by imputation for this course.
   - 7.1 Why weights are necessary. Because of the complex sampling scheme, simple statistics (such as simple averages) will not give unbiased estimates of the population quantities of interest. Consequently, sampling weights must be used in weighted estimates to give unbiased

[5] Hint: Suppose the survey was about poverty rates? Would the ATUS population be suitable?

estimates of the population quantities. Researchers working with the multi-year data files created by BLS can use thestatistical weighting variable *TUFNWGTP* to account for the sampling scheme.[6]

3. Read in the ATUS data. Make a COPY of the *atussum_0313.sas* file for each assignment. Modify the COPY to include the title and to point to the *atussum_0313.csv* file. If you are using the *SAS* University Edition, don't forget to set and use the *&path* macro variable.

   Run this code to read the (large) summary file – there are over 100,000 records (which is fairly small compared to many data sets encountered in the real world). Check the log-file to ensure that it has been read properly. Print out the first 5 records to ensure that you've read the data properly. Notice that some variables are coded using -1 to indicate a missing value.

4. We will start with total time spent watching television. From the *atussum_0313.sas* file, you will see that two variables describe television viewing, *t120304* and *t120303*. A -1 may indicate a missing value. Recode any -1 to missing. Compute a derived variable *totaltv* (minutes) for the sum of these two variables and add this to the summary data. This should be done in a NEW data step. Similarly, create a derived variable *watchtv* that is 0 if $totaltv = 0$ and is 1 if $totaltv > 0$. This variable represents if this person watched tv during the selected day.

5. Use *Proc Univariate* to compute some summary statistics and draw a histogram of the total time spent tv watching in a day. The histogram should have breakpoints for the bars at 30 minute intervals.[7] There are two odd things you should spot from the histogram:

   - heaping. How do you recognize this? Why do you think this occurred? This is a very common problem when measuring many types

[6] **A few words about sampling weights**. Consider first a population of 100 people. Suppose that a simple random sample of size 20 is selected. In a simple random sample, each person has the same probability of selection and the selection of one person is independent of the selection of another person. So in this case, each person has a probability of $20/100 = 0.2$ of being selected. Conceptually, each of the 20 selected people "represents" 5 people in the population and this is known as the sampling weight. Notice in this example that every person has the same sampling weight. Sampling weights are highly related to the inverse of the probability of selection, but also account for other features of the design such as adjustments for missingness etc. The estimate of the mean of the population is found as $\widehat{\widehat{Y}} = \frac{\sum w_i y_y}{\sum w_i}$ where w_i is the sampling weight. Because all of the weights are equal, this reduces to the simple sample mean.

Now suppose that you stratified in advance by sex and that the 100 people in the population have 60 males and 40 females. From the 60 males, you take a sample of size 10. From the 40 females you also take a sample of size 10. Now the probability of selection for a male (10/60) is smaller than the probability of selection for a female (10/40). The sampling weights are 6 and 4 respectively. The estimate of the overall mean (over both males and females) is again the weighted average above.

In the ATUS, the sampling weights have to account for the complex stratification, over/under sampling by strata, and over/under sampling by day of the week.

[7] This can be done using one of the options from the *histogram* statement in *Proc Univariate*

of variables using recall.

- outliers. How many hours is a 1200 minutes of TV watching? Is this realistic? What is the largest value in the dataset? How many observations over the multiple years of the study have 1200+ minutes of TV watching? We will, for now, retain these outlying values because removing them complicates the computation of the sampling weights (see later) and they actually have little impact on the final result.

6. Use *Proc Tabulate* to compute the average amount of TV watched by year and the proportion who watch TV by year.[8] Because of the complex sampling design, you need to compute WEIGHTED statistics - see Section 7.4 of the ATUS User Guide. This is done in many of the *SAS* procedures using the *Weight* statement. The variable *TUFNWGTP* is the correct weighting variable to use.

7. Use *Proc Means* to compute the average amount of TV watched by year and the proportion who watch TV by year again using the *Weight* statement. Use the *Output* statement to send the results to a new data file, i.e. the new data file should have 1 line per year with the (weighted) average time watched and the proportion of the population of interest who watched tv at some point during the day. Don't forget that you need to sort a dataset first before using the *By* variables to get computations by individual groups.

8. Plot the average time watched per way over the years of the survey using *Proc SGplot*. Join the individual points over time using the *Series* statement. Add the regression line in much the same way as you did for the cereal example.

   Add to the plot the proportion of viewers over time (hint: you can plot a second variable on an axis using the right side of the plot, using the *yaxis2* options on many of the statements in *Proc SGplot*. You can adjust the color and type of lines (e.g. use a dashed line for the proportion of tv viewers).

9. What do you conclude based on these plots. What is the estimated increase in the average amount watched/year (hint: estimated it from the graph). Why was there a sudden increase in the average amount of TV watched in 2008?

Hand in the following using the online submission system:

- Your SAS code.

[8] Hint: Recall in the Titanic example, you computed a proportion by averaging a 0/1 variable.

- A PDF file containing the the output from your SAS program.
- A one page (maximum) double spaced PDF file containing a short write up on this analysis suitable for a manager who has had one course in statistics. You should include:
  - A (very) brief description of the dataset.
  - The histogram of the hours of TV watched per day over all years in the study.
  - The plot(s) you created of the changing average amount of TV watched and proportion of population who viewed TV in a day. Use ODS RTF file=****; and ODS RTF close; around the SGplot above to also send the plot to an RTF file that you an then easily open in MSWord (and other word processors).
  - What is your explanation for the sudden jump in average amount of TV viewing in 2008?

Hints for this assignment

DO NOT PANIC.

Yes, this assignment will take a fair amount of time, but later assignments will take less time. In this assignment, you are learning how to import data, learning how to run procedures, learning how to get output from SAS into various forms etc.

Future assignments will build on the material from Assignment 1. For example, in Assignment 2, I will have you do a multiple regression based on the cereal dataset.

Some general things to look out for:

- Some students did not put their name and student numbers on the write-ups.
- Some students did not put their name and student numbers in the header of the *SAS* code files.
- Students included raw code in the write up. This is seldom a good thing to do as you manager does NOT want to see raw code.
- Students used A4 rather than US letter-sized paper.
- Too many decimal places reported. It is quite rare to report more than 2 significant digits for estimates.

Part 1 Hints

***** How do I turn on ODS graphics for Proc Reg?
If you are running SAS 9.3+, the default is ON so you don't have to do anything.
If you are running SAS 9.2 or earlier, your code will look something like:

```
ods graphics on;
proc reg data=****;
    statements
run;
ods graphcs off;
```

***** How do I get predictions for 4 grams of fat.
For this assignment, simply read off the prediction, the confidence intervals for the mean, and the prediction interval for an individual response directly off the appropriate plot. No additional programming is needed.

If you want the computer package to do the computations, the general procedures is to create "dummy" data which is appended to the real dataset with the Y value set to missing and the remaining covariates (X values) set to the values of interest. In this case, you would have to add a "fake" observation where grams of fat $= 4$ and $calories = .$ This is not needed for this assignment; in later assignments you will see how to do this. You can look at `http://people.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms/SASTricks/SampleSASPrograms/` for details on how to do this.

***** What do I need to include in my report?
All of questions that I ask in the bulleted parts of the assignment are for your own personal use. These are examples of the types of questions that I will ask on exams/term tests.

***** How big should the report be?
The report is typically 1 or 2 pages at most. Start with two or three sentences describing the data set. Then give the fitted regression line along with the *se* and confidence interval for the slope. Use the Equation Editor of MSWord to create the equation of the line. Interpret the slope, the se and the ci for the manager. This is an additional 5 or 6 sentences.

Finally, give the estimate of the number of calories for 4 grams of fat along with the two intervals (the confidence interval for the mean response and the prediction interval for a future response). Explain the difference between the intervals. This is 5 or 6 sentences again.

No computer output should be included in your write up for part 1.

Part 2 Hints

Some common error made:

- Interchanging the terms "odds" and "probability". While there is a 1-1 relationship between the two terms, i.e. every probability has an associated

odds value and every odds value is associated with a probability, they are NOT the same thing. Don't mix them up.

- Saying that 'odds ratios measure association between two events', but not explaining further.
- Flipping back and forth from odds ratio of death and odds ratio of survival in the same explanation, often in the same sentence! Pick one response and make sure that your write up is always about the response variable that you chose.
- Many students were vague about the odds ratio being odds of death or survival for women compared to the odds of death or survival of men (most of these students were also vague about what an odds ratio actually is as well). Every odds ratio must be declared BOTH in terms of the variable of interest and the direction of the ratio. For example, you would say something like "the odds ratio of survival (male:female)" to indicate that the statistic is

$$OR = \frac{ODDS_{Survival}(males)}{ODDS_{Survival}(females)}$$

.

***** How do I use Proc Tabulate?
Proc Tabulate is your friend. I find it one of the more useful procedures. Your code will look something like

```
Proc tabulate data=****;
   class passengerclass sex survived;   /* defines your classification variables */
   table passengerclass*sex, survived;  /* get a table */
run;
```

***** How do I use Proc Tabulate to get the proportion that survived?
You code will look something like. Notice how the survival variable is now used.

```
Proc tabulate data=****;
   class passengerclass sex;   /* defines your classification variables */
   var survived;
   table passengerclass*sex, survived*mean;  /* get a table with the mean of the survived va
run;
```

**** How do I use a BY statement?
The BY statement is one of the most powerful features of SAS enabling you to get repeated analyses on subsets of your data.

Your code will look something like:

```
Proc freq data=****;
    by passengerclass;  /* separate analysis by passenger class */
    table .....
    ods output TableName=YourNewSASdataset;
run;
```

Don't forget to sort the data by the BY variables prior to the procedure.

**** My odds ratios are in the wrong order, i.e. dead:alive rather than alive:dead, or females:males rather than males:females. There are several options:

- Try a different order on the Tables statement, i.e. Tables x*y; vs Tables y*x;

- Create a new derived variable that has the levels with a different sort order SAS typically sorts levels of factors alphabetically or numerically, i.e. females come before males, 0 comes before 1.

  So create a new variable for alive status such as:

  ```
  if survival = 0 then  survival_new = 'dead';
  if survival = 1 then  survival_new = 'alive';
  ```

  and do the tables using the variable *survival_new* as it is sorted in a different order. Don't forget to specify the length for the *survival_new* variable. (Why?).

- After you get the odds, take the inverse. For example, if the odds are computed as male:female, then the female:male odds is 1/male:female odds;

  ```
  data relativerisks;
     set relativerisks;  /* this accesses the SAS dataset that was created earlier */
     odd2 = 1/ odds;
  ```

**** How do I select observations for plotting?
The ODS statement in *Proc Freq* will give you MORE data than you need for the plots. To use a subset of a data in any procedure, use the *Where* statement.

Your code fragment will look something like:

```
Proc SGplot data=xxx;
   where studytype='Case-Control (Odds Ratio)';
   ....
run;
```

**** How do I plot the confidence bounds?
The ODS statement will create a dataset that has the estimated odds ratio by passenger class along with the lower and upper confidence limits. Your code will look something like:

```
proc sgplot data=RelRisks;
   title3 'Comparison of ODDS ratio (and 95% confidence intervals) among passenger classes';
   where studytype='Case-Control (Odds Ratio)';
   scatter y=value x=pclass;
   highlow x=pclass low=lowercl high=uppercl; /* draws the ci lines */
   ...
run;
```

**** How do I get the plot into a file for my word processor?
Easiest is to create an RTF file. Your code will look something like

```
ods rtf file='myrtffile.rtf';
proc sgplot data=****;
   ...
run;
ods rtf close;
```

Part 3 Hints

TBA

Comments from the marker from previous years.

The marks are now up, however some of you may see 0's for your writeups. If you haven't written your name at the top of the writeup, that is the probable reason.

DON'T PANIC– I'll bring a stack of nameless assignments to the tutorials, they're all marked and waiting for you! Just pick yours out of the pile and give it to me and I'll update your mark on the spot.

A few notes on the target audience for your assignments

Much of your assignment grades depends mostly not on how well you program in SAS, but how well you communicate your results. And that's a pretty big deal. Out there in the "real world", your boss is not likely to look at your brilliantly clever code or your fancy charts, but give just a brief passing glance at the first page of your written report to determine how good you are. Which is.... you guessed, precisely what I would be doing every week this semester!

Therefore, as you compose your write-ups, imagine that I am your boss, rather than your Statistics TA. Imagine, in fact, that I never took statistics in college. I don't know what a confidence interval is, or what an odds ratio is, or what wine to order with regression. Instead, I'm extremely busy making millions, flying private jets to summits in Costa Rica and making keynote addresses at ribbon-cutting ceremonies. I need to be told, directly and concisely, what your data set is about, and what the important features of it are.

So, if something doesn't make sense to me, I won't read it twice to try and understand it – I'll just fire you and hire someone who makes sense. Yikes, that's harsh! But, of course, it's just pretend right now. :) Instead of firing you, your grade on the assignment is reduced which gives you another chance to impress me. So, no pressure there:)

In other words... impress me! Make me want to give you a promotion! Your write-ups are a direct reflection on your competence and professionalism. Be technical, but not text-bookey; be simple but not pandering; be concise but thorough. It's a delicate art, and it's worth learning.

A few notes on including plots with your write-ups.

Sometimes the directions ask you directly to include a specific plot. Other times, it's simply a good idea. A graphical plot within a written report can be a powerful, and dangerous thing. It can make everything beautifully clear, but can also confuse the reader entirely, and it's up to you which one it's going to be. Some things I consider important:

- Every plot, without exception should have a concise and informative Title. What is depicted on the plot? Is its purpose to predict something? To compare something? To illustrate a relationship? To highlight a feature in your data?

  Example of a good title: "Comparative Awesomeness of several Nifty Things vs. Time, 1984 - present"

- Every plot should have a reference to it somewhere in the text of the writeup, which should explain (a) why the plot is relevant and (b) which features of the plot I should be looking at. I never want to see plots that are purely "ornamental"!

  Example of a good reference: "Some Nifty Things have been found exhibit widely varying Awesomeness throughout the years. For instance, note in Figure 1, the Awesomeness of Transformers drops sharply in the early 90's, but rises considerably in 2007, and continues to rise every year afterwards. This could be due to...". example of a bad reference: "According to the plot, Awesomeness is a feature of Nifty Things."

- If you include a reference to a plot in your write-up, DO include the plot as well! It's just mean not to:)

- Not all plots require these, but check all that apply: aptly named axes, units, a legend, labels for specific points, etc. Ideally, your boss should be able to glean all the important information from the plot alone, before she even reads the write-up.

- Legends for Figures go at the bottom of the figure. Legends for Tables go at the top of a table. Legends need to explain any codes used in the plot. Every plot/figure should be able to stand on its own.

A few notes on explaining statistics in your write-ups.

Oftentimes you'll be asked to report statistick-ey information in your write-ups. Stuff like prediction intervals and confidence bands and slopes and regression

coefficients, and whatnot. Remember, I'm not a statistics grad student, I'm your boss, entirely uneducated in the subject. I am unlikely to read or follow your formulas should you decide to include them. I will not be impressed how well you quote Wikipedia or Websters or your textbook when giving me a definition of a technical term. In fact, I've probably been given the definitions of those things before, and forgot them, because I didn't really care, and I'm not about to start caring.

What interests me is one thing: What does this statistics mumbo-jumbo mean in relation to what the data is really about? What does that mean in terms of cereal calories, or deaths on the Titanic, stuff I actually care about?

So, explain those things, but in a way that is relevant to the specific data at hand. Feel free to include examples – they can be very useful! But don't get too long-winded, or I may lose interest.

A few notes on the "irregardless" effect.

You are in University now, so things like punctuation and capitalization and spelling and formatting are usually not included in the marking scheme. This doesn't mean they're not important – it means that we expect students to automatically include those features in their work.

So what happens if you don't? Will you be penalized for it? Yes and no. You shouldn't see marks taken off because you didn't write in complete sentences or didn't spell "forthwith" correctly. But when reading a paper with obvious (and not necessarily content-related) errors, I am naturally more inclined to expect content-related errors, and to look for them. Why would you want me to actively look for errors in your paper?:)

A really good technique is to have another student read your report before you hand it in – and it doesn't even have to be a Stat 340 student! In fact, it's better if they're not.

How to write good code

There is a sample piece of code on the assignment page. Follow its format. In particular:

- Start the code with a header with your name and student number
- The header should also state the purpose of the code and have a change log. The change log is VERY important in the real world because code gets constantly modified over the years and it is important to know who modified the code and why.
- Include plenty of comments. But is not necessary to document every line because much of the code in *SAS* is relatively self documenting. For example, *Proc Print* presumably prints out a data set; *Proc Reg* does a regression and the model is right in the code, etc.
- Use white space and blank lines to make the code easier. For example, I suggest you write procedures by indenting the statements within procedures such as seen in the sample program. Separate procedures with blank lines. Add blocks of comments if a section of code analyzes variable $Y1$ and then another block of code analyzes variable $Y2$ etc.